Simulation Foundations, Methods and Applications

Claus Beisbart
Nicole J. Saam *Editors*

Computer Simulation Validation

Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives

Springer

Simulation Foundations, Methods and Applications

More information about this series at http://www.springer.com/series/10128

Claus Beisbart · Nicole J. Saam
Editors

Computer Simulation Validation

Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives

Springer

Editors
Claus Beisbart
University of Bern
Bern, Switzerland

Nicole J. Saam
Friedrich-Alexander-Universität
Erlangen-Nürnberg
Erlangen, Germany

ISSN 2195-2817 ISSN 2195-2825 (electronic)
Simulation Foundations, Methods and Applications
ISBN 978-3-319-70765-5 ISBN 978-3-319-70766-2 (eBook)
https://doi.org/10.1007/978-3-319-70766-2

Library of Congress Control Number: 2018966848

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

Preface

This volume is interdisciplinary all the way through. It all started, when a social scientist (NJS), who had done simulations for quite some time, consulted the philosophical literature to obtain a better grip on what she was doing. At some point, she raised a question to a philosopher (CB) who had written on computer simulations. The cross-disciplinary exchange that emerged in this way eventually led to our collaboration.

The first significant step was the organization of an interdisciplinary conference about the validation of computer simulations. It took place at Herrenhausen Castle in Hanover, Germany, in 2015 and was generously supported by the Volkswagen Foundation. We invited working scientists doing simulations in the natural and social sciences, e.g., in astrophysics, ecology, and economics, further, mathematicians, computer scientists, and philosophers to share their experiences of validating simulations, to propose frameworks to think about validation, and to offer philosophical reflections about it. The conference was different from most conferences at which we had been before and which were typically meetings with lots of old friends. This time, by contrast, even the organizers had not met most of the participants before. Despite this (or, rather, therefore?), the conference featured very lively discussions, and we learned a lot from the talks. But in our view, the conference also showed that, first, more cross-disciplinary dialogue is needed to facilitate exchange about validation across disciplinary boundaries. Second, our impression was that many valuable accounts of, and techniques for, validation have been developed for restricted areas of research, but that researchers from other disciplines might benefit from them. So it is important to find out whether, and, if so, to what extent, these accounts and techniques may be transferred to other areas. Third, we felt that validation raises a lot of questions of broader relevance to scientific methodology and a philosophical understanding of scientific practice. This was motivation enough to edit this volume, which is meant to serve as a handbook about the validation of computer simulations.

Many people have helped us to publish this volume, and we are grateful to all of them. Our first thanks go to our authors (some of which were at the conference, while we became only aware of others when putting this volume together). The authors have not only contributed their chapters, which was often significant work because there is a lot of uncharted territory when it comes to validation and also because they had to address a broad readership. Most authors have also reviewed at least one other chapter of the volume and often given invaluable advice to their fellow contributors. We are particularly grateful to David Murray-Smith, William Oberkampf, and Patrick J. Roache who have each kindly refereed more than one chapter. Finally, we also owe our authors a lot of useful comments and suggestions. To give just one example, Richard B. Rood worked through the whole introduction and sent us very useful comments.

We are also grateful to a number of "external" referees who have helped us to review the chapters, viz., Petra Ahrweiler, Stefan Gruner, Paul Humphreys, Stephan Poppe, Rush Stewart, and Eric Winsberg.

Our thanks go further to our (former) student assistants who have helped with proofreading and literature reviews, viz., Soham Astik, Sarah Gloor, Christoph Merdes, and Audrey Salamin.

We are grateful to the series editor, Louis Birta, for his encouragement and support.

Last but not least, we wish to thank our publisher and the Springer team with which we have collaborated, in particular, Simon Rees and Wayne Wheeler. The cooperation was extremely smooth, and we are grateful for their support, their professional advice, quick answers to our questions and for the patience that is needed for long-term projects such as editing this volume.

Bern, Switzerland Claus Beisbart
Erlangen, Germany Nicole J. Saam
October 2018

Contents

Part III Methodology—Preparatory Steps

Part IV Methodology—Points of Reference and Related Techniques

Part V Methodology—Mathematical Frameworks and Related Techniques

Part VI Methodology—The Organization and Management of Simulation Validation

Part VII Validation at Work—Best Practice-Examples

Part VIII Challenges in Simulation Model Validation

Part IX Reflecting on Simulation Validation: Philosophical Perspectives and Discussion Points

Contributors

Eckhart Arnold Bavarian Academy of Sciences and Humanities, Munich, Germany

Michael Baldauf Deutscher Wetterdienst, Offenbach am Main, Germany

Anouk Barberousse Sorbonne Université, Paris, France

Christoph Baumberger ETH Zurich, Institute for Environmental Decisions, Zürich, Switzerland

Claus Beisbart Institute of Philosophy, University of Bern, Bern, Switzerland

Keith Beven Lancaster Environment Centre, Lancaster University, Lancaster, UK

Seamus Bradley TiLPS, Tilburg University, Tilburg, Netherlands;
University of Leeds, Leeds, UK

Alan C. Calder Stony Brook University, Stony Brook, NY, USA

Xueyu Cheng Clayton State University, Morrow, GA, USA

Christine S. M. Currie Mathematical Sciences, University of Southampton, Southampton, UK

Lori A. Dalton The Ohio State University, Columbus, OH, USA

Roozbeh Dehghannasiri Texas A&M University, College Station, TX, USA

Edward R. Dougherty Texas A&M University, College Station, TX, USA

Giorgio Fagiolo Istituto di Economia, Scuola Superiore Sant’Anna, Pisa, Italy

Mathias Frisch Institute for Philosophy, Leibniz Universität Hannover, Hanover, Germany

Axel Gelfert Technische Universität Berlin, Berlin, Germany

Mattia Guerini Istituto di Economia, Scuola Superiore Sant’Anna, Pisa, Italy;
OFCE - Sciences Po, Paris, France

Gertrude Hirsch Hadorn ETH Zurich, Institute for Environmental Decisions, Zürich, Switzerland

Cyrille Imbert Archives Poincaré, CNRS, Université de Lorraine, NANCY Cedex, France

Julie Jebeile Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Xiaomo Jiang Tongji University, Shanghai, China

Reto Knutti ETH Zurich, Zürich, Switzerland

Josef Köstlbauer History Department, University of Bremen, Bremen, Germany

Francesco Lamperti Istituto di Economia, Scuola Superiore Sant'Anna, Pisa, Italy;
FEEM, Milano, Italy

Stuart Lane Institute of Earth Surface Dynamics, Université de Lausanne, Lausanne, Switzerland

Johannes Lenhard HLRS, University of Stuttgart, Stuttgart, Germany

Fei Liu School of Software Engineering, South China University of Technology, Guangzhou, China

Robert E. Marks School of Economics, University of New South Wales, Sydney, NSW, Australia

Michael Mäs Department of Sociology/ICS, University of Groningen, Groningen, The Netherlands

Peter Mättig Physikalisches Institut der Universität Bonn, Bonn, Germany

Matthias Meyer Hamburg University of Technology, Hamburg, Germany

Alessio Moneta Istituto di Economia, Scuola Superiore Sant'Anna, Pisa, Italy

David J. Murray-Smith School of Engineering, University of Glasgow, Glasgow, UK

William L. Oberkampf Sandia National Laboratories - retired, Albuquerque, NM, USA

Oliver Reinhardt University of Rostock, Institute of Computer Science, Rostock, Germany

William J. Rider Sandia National Laboratories, Center for Computing Research, Albuquerque, NM, USA

Patrick J. Roache Consultant, Socorro, NM, USA

Andrew P. Robinson CEBRA, School of BioSciences, The University of Melbourne, Parkville, Victoria, Australia

Richard B. Rood Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA

Andrea Roventini Istituto di Economia, Scuola Superiore Sant'Anna, Pisa, Italy; OFCE - Sciences Po, Paris, France

Christopher J. Roy Crofton Department of Aerospace and Ocean Engineering, Virginia Tech, Blacksburg, VA, USA

Andreas Ruscheinski University of Rostock, Institute of Computer Science, Rostock, Germany

Nicole J. Saam Institute for Sociology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

K. Heinke Schlünzen Meteorological Institute, Center for Earth System Research and Sustainability (CEN), Universität Hamburg, Hamburg, Germany

Jan Seibert Department of Geography, University of Zurich, Zurich, Switzerland

Maria Staudinger Department of Geography, University of Zurich, Zurich, Switzerland

Susanne Theis Deutscher Wetterdienst, Offenbach am Main, Germany

Dean M. Townsley University of Alabama, Tuscaloosa, AL, USA

Adelinde M. Uhrmacher University of Rostock, Institute of Computer Science, Rostock, Germany

H. J. (Ilja) van Meerveld Department of Geography, University of Zurich, Zurich, Switzerland

Tom Warnke University of Rostock, Institute of Computer Science, Rostock, Germany

Ming Yang Control and Simulation Center, Harbin Institute of Technology, Harbin, China

Yong Yuan Tongji University, Shanghai, China

Chapter 1
Introduction: Computer Simulation Validation

Claus Beisbart and Nicole J. Saam

Abstract To provide an introduction to this book, we explain the motivation to publish this volume, state its main goal, characterize its intended readership, and give an overview of its content. To this purpose, we briefly summarize each chapter and put it in the context of the whole volume. We also take the opportunity to stress connections between the chapters. We conclude with a brief outlook.

The main motivation to publish this volume was the diagnosis that the validation of computer simulation needs more attention in practice and in theory. The aim of this volume is to improve our understanding of validation. To this purpose, computer scientists, mathematicians, working scientists from various fields, as well as philosophers of science join efforts. They explain basic notions and principles of validation, embed validation in philosophical frameworks such as Bayesian epistemology, detail the steps needed during validation, provide best practice examples, reflect upon challenges to validation, and put validation in a broader perspective. As we suggest in our outlook, the validation of computer simulations will remain an important research topic that needs cross- and interdisciplinary efforts. A key issue is whether, and if so, how very rigorous approaches to validation that have proven useful in, e.g., engineering can be extended to other disciplines.

1.1 Introduction

During the second half of the twentieth century, computer simulation has established itself as a new method in most natural and social sciences. These days, computer simulations are used to investigate as diverse phenomena as the healing of wounds (e.g., Walker et al. 2004), the climate of our planet (see, e.g., IPCC 2014, in particular,

C. Beisbart (✉)
Institute of Philosophy, University of Bern, Bern, Switzerland
e-mail: claus.beisbart@philo.unibe.ch

N. J. Saam
Institute for Sociology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
e-mail: nicole.j.saam@fau.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_1

Chap. 9 for an overview), or the distribution of taxpayers in societies (e.g., Harding et al. 2010). In many areas, computer simulations have replaced experiments and led to insights that would otherwise have been utterly impossible. But to what extent can we trust the results from simulation research, e.g., the predictions from climate simulation models?

To address this question, scientists try to validate their simulations. Validation comprises the efforts to show that computer simulations provide adequate representations of their target systems. According to a very famous definition, validation is supposed to substantiate "that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model" (Schlesinger et al. 1979, p. 104, emphasis deleted).

This may seem straightforward and not particularly difficult. However, as it happens, validation of computer simulations is a challenging issue both from a practical and a more theoretical perspective.

As far as practice is concerned, some working scientists have expressed their unease about validation. As, e.g., Kleindorfer et al. (1998, p. 1087) report, "[t]here is still considerable doubt and even anxiety among simulation modelers as to what the methodologically correct guidelines or procedures for validating simulation models should be." In a more recent contribution, Ghetiu et al. (2010, p. 1) complain that "a cohesive understanding of what scientific validation requires, is not captured by the existing efforts that mainly try to solve pieces of the 'puzzle'" (emphasis deleted). Accordingly, in practice, activities of validation are often neglected or only done sloppily and superficially. As a consequence, some results from computer simulations have later turned out to rest on erroneous numerical artifacts. In the absence of clear guidelines of how to validate simulations, the method of computer simulation, successful as it might seem, is not yet fully developed.

It is remarkable in this regard that some working scientists from, e.g., business analysis have resorted to the philosophical literature. Authors such as Herskovitz (1991), Kleindorfer et al. (1998), Feinstein and Cannon (2003), Klein and Herskovitz (2005) have examined various philosophical positions to see whether they can make sense of validating simulations.

As far as the theoretical understanding of validation is concerned, several issues have emerged that haven't yet been treated satisfactorily. Already the very term "validation" is a matter of controversy. In an often-quoted paper, Oreskes et al. (1994) try to show that the term is misleading because a simulation cannot be shown to be true or valid except in trivial cases. However, as Rood in Chap. 30 of this volume argues, Oreskes et al. have gone too far with their skeptical outlook. We are not merely talking about verbal disputes here; there is a real issue about how we may properly characterize the epistemic situation that successful validation can optimally produce. A related key question is how one can determine the overall confidence of simulation results if a battery of tests have been carried out. What is also a matter of discussion is the question of how validation is related to what people call verification, i.e., the attempt to show that a simulation reliably traces the predictions of a model. In the philosophical literature, we find a discussion about the extent to which computer simulation and its validation are novel and require a new epistemology (see, e.g., Winsberg 2001).

This volume is meant to be a reaction to this dissatisfying understanding of validation. It aims at a methodological and philosophical discussion of the validation of computer simulation and its techniques. True, the last few years have seen impressive systematic work on validation, notably the monographs by Roache (2009), Oberkampf and Roy (2010) and Murray-Smith (2015). But the focus of these books is mostly on the physical sciences and engineering, so it is not clear whether and how their approaches may be generalized to other disciplines. To give just one example, Oberkampf and Roy stress the importance of model validation experiments, which cannot be done in many parts of the social sciences.

In our view, validation needs a much broader discussion. Validation is an issue not just for the physical sciences and engineering but is rather relevant to all disciplines in which computer simulations have currency. Thus, cross-disciplinary communication about the understanding and the practices of validation is needed to put validation in a broader perspective and to avoid misunderstandings that arise in interdisciplinary research, when scientists from various fields differ in their tacit assumptions about validation. Scientific exchange across disciplinary boundaries can also help to disseminate knowledge of new techniques and to share experiences about validation. It is thus time to approach the topic by taking seriously the perspectives of all parties that can contribute to a better understanding of validation. These parties comprise first and foremost practicing simulation scientists from the natural and social sciences as well as computer scientists. Mathematicians are needed first not only because many computer simulations are based on approximation schemes, the appropriateness of which needs close analysis, but also because validation involves the testing of statistical hypothesis and the quantification of uncertainties. Finally, the expertise of philosophers of science is called for, since some questions raised about validation touch upon more general issues about scientific method.

1.2 Goals and Readership of this Handbook

This handbook aims to significantly improve the understanding of validation of related methods and thus ultimately to promote more thorough and sustained validation practices. To this purpose,

- Experts in validation and philosophers of science explain and clarify basic concepts and principles used to frame validation, e.g., the notions of validation, verification, and error.
- Philosophers embed validation in general philosophical frameworks that try to explain how scientific method works, e.g., in Bayesian epistemology.
- Experts from some sciences and engineering as well as mathematicians explain the steps needed for validation, provide practical advice, and introduce related techniques.

- Working simulation scientists from various fields share their experience about what they take to be best practice of validation.
- Practicing simulation scientists reflect upon challenges of validation, e.g., the sparseness of data.
- Philosophers address philosophical questions that have been raised about validation.

The focus of this book is on the validation of *computer simulations in scientific research*. By computer simulation, we mean, very roughly, a method in which a computer program traces the dynamics of a target system by providing (typically approximate and partial) solutions to a model of the target. But a lot of results from the chapters collected in this volume are also relevant to other scientific computer-based methods in which a target system is represented using a model, for instance to representations of spatial structures (a building, a bone) at one time. Further, many results are transferable to computer simulations done outside science, e.g., in business companies or administration. These days, computer simulation programs are distributed on a commercial basis, and customers rightly expect that the programs be validated. Note, too, that computer simulations are also a valuable tool in education. Although the validation of simulations used in areas such as education raises issues outside the scope of this volume, many concepts and approaches from validation in scientific research can be applied, as far as the adequate representation of the target system is concerned (see Chap. 2 by Beisbart, Chap. 17 by Saam and Chap. 36 by Köstlbauer in this volume for contributions that touch upon the use of simulations outside science).

This volume is unique in at least three respects: To begin with, it is the first, and so far only, handbook-like collection on validation in which international experts join forces to explain the fundamental concepts, strategies, and techniques of validation to a broad readership. Second, it is first to draw on the expertise from various fields ranging from engineering and the physical sciences to the social sciences and history. Third, it is unique in providing new and original philosophical reflections about validation. Although the philosophy of science has featured a lively debate about computer simulations during the last few years, the thorny issue of validation was so far much neglected by philosophers. Part IX of the book provides the first collection of philosophical articles on validation of simulations.

The intended readership of this book is mainly comprised by working scientists from all natural and social sciences (see Table 1.1). We hope that this volume will be particularly helpful for young scholars who start research with computer simulations. On top of this, we wish to engage with philosophers of science as well as methodologists who want to increase their understanding of simulation validation.

Before we start with an overview of this book, two remarks are in order. First, since we have aimed to bring together the perspectives of various disciplines, we have decided not to base the chapters on a shared understanding of validation. Instead, the contributors were asked to specify their preferred view of validation. Note though that Chap. 2 by Beisbart systematically compares various understandings of validation. Second, depending on their principal aims, the chapters differ significantly in

Table 1.1 Readership of this volume

Disciplines	Readership	Features that will appeal to this audience
Empirical sciences, natural and social	working scientists as well as advanced students, methodologists	- exposition of the foundations of validation and related concepts - introduction to steps of validation and corresponding techniques - introduction to mathematical frameworks that are useful for validation - best practice examples - reflection upon challenges of validation
Philosophy of science	Philosophers of simulation, methodologists	- exposition of the foundations of validation and related concepts - reflections about whether validation can fit into existing philosophical frameworks - general philosophical reflection about validation and its significance

style. While many chapters are survey articles, others provide introductions or even tutorials. A couple of chapters, finally, are something like original research articles because they address questions that haven't been discussed yet. But we have always tried to make sure that the chapters are accessible to a broader readership and that they explain the needed background knowledge.

1.3 Structure and Topics

This handbook comes in nine parts. Foundations are explained in Parts I (about basic concepts) and II (about philosophical frameworks for thinking about validation). Parts III through VI detail the methodology of validation by considering the crucial steps and building blocks: preparatory steps (Part III), points of reference and related techniques (Part IV), mathematical frameworks (Part V), and the organization and management of simulation validation (Part VI). Part VII provides best practice examples from various fields. Challenges that arise from peculiarities of certain types of simulation models or some disciplines are reflected upon in Part VIII. Part IX, finally, collects chapters that address philosophical issues surrounding validation.

1.3.1 Foundations (Parts I–II)

Part I of the book begins with a series of chapters that explain fundamental concepts and principles related to validation.

Chapter 2 by Claus Beisbart addresses the question: “What Is Validation of Computer Simulation?” (thus, the title). This question has so far been answered using various definitions. The objective of the chapter is not to argue for one specific definition, but instead to understand how and why existing characterizations of the validation of simulation differ. Beisbart considers a representative sample of attempts at defining validation, e.g., the famous definition by Schlesinger et al. (1979) and definitions that authors from this volume use in their chapters. The definitions agree that validation is an evaluation, but differ on what exactly the proper object of the evaluation is, e.g., results from a simulation, a simulation program or a model. Beisbart thus clarifies how these entities hang together. The definitions, too, mention different standards of evaluation, e.g., truth, accuracy, credibility, and adequacy for a purpose. Beisbart explains these standards and argues that the validation of a whole computer simulation can at best establish a high credibility for claims on the accuracy of the simulation outputs. This view is to some extent compatible with the idea that simulations should be adequate for a purpose, as famously suggested by Parker (2009). But it is arguable that adequacy for a purpose requires researchers to take into account additional epistemic values, in particular, if simulations are supposed to explain some phenomena. Beisbart thus considers the notion of what is called structural validation and discusses reasons for and against allowing an appeal to additional standards in validation. Beisbart concludes by presenting a scheme for defining validation.

Chapter 3 is written by William L. Oberkampf and titled “Simulation Accuracy, Uncertainty, and Predictive Capability: A Physical Sciences Perspective”. The author expounds and defends a conception of validation which he calls “the restrictive notion of model validation”. Validation in this sense is an assessment of model accuracy relative to available experimental data. It is necessary, but not sufficient for showing the predictive capability of simulations. A simulation has predictive capability if it can foretell, within an estimate of uncertainty, the response of a system for conditions that have not yet been observed in nature. This capability is extremely important for applications of simulations in engineering and simulation-informed decision-making. As Oberkampf explains, what is most important for validation in the restrictive sense is the assessment of the accuracy and of uncertainties. As he further points out, validation in this sense needs code verification and solution verification and the estimation of uncertainty. The chapter details what these activities are and how they should be accomplished in practice. In this context, Oberkampf defines crucial terms needed for carrying out these activities, e.g., model form error and validation metrics. He stresses that model validation should not be confused with model calibration. He further explains the definition and use of model validation experiments. All in all, the chapter describes a powerful approach to validation that has been proven extremely valuable in the physical sciences. It is quite demanding though because it requires the availability of appropriate data, and it is restricted to areas in which simulations are built upon the known principles that govern the target system. In his conclusions, Oberkampf argues that these conditions are not met in the social sciences. As a consequence, the notion of validation defended in this chapter is not applicable in the social sciences.

In Chap. 4, titled "Verification and Validation Principles from a Systems Perspective", David J. Murray-Smith offers a broader perspective on validation and verification. He calls it a systems perspective because the real-world target of the simulation is regarded as a system, which is often composed of many elements. Murray-Smith's focus is on so-called lumped parameter models, which typically model each element of the system using one characteristic. The models then use ordinary differential equations or a combination of ordinary differential and algebraic equations to model the interconnections between the elements. Note though that most of the notions and principles developed in the chapter can be extended to other simulations. As does Oberkampf in his Chap. 3, Murray-Smith stresses that both verification and validation of the simulation are required, where the former is characterized as internal, while the latter is said to be external. The chapter explains the principles that guide verification and validation. As far as the latter is concerned, the focus is on accuracy, and Murray-Smith points out that the latter should be assessed not only by comparing simulation outputs and measured data, but also by using methods such as identifiability and parameter sensitivity analysis. What is distinctive about the approach by Murray-Smith in comparison with Oberkampf is, maybe, that the former is more friendly towards face validation, a method which is crucially based upon expert judgment. He thinks that face validation is particularly helpful during the early stages of the development of a simulation.

Two basic notions that are decisive for both validation and verification are those of error and uncertainty. Both validation and verification can be described as attempts to deal with various kinds of errors that can arise in computer simulations. But what exactly are these errors? How can they be classified? And how can researchers appropriately deal with them? These are the questions at the center of Chap. 5 by Chris Roy, titled "Errors and Uncertainties, Their Sources and Treatment". He defines error as the difference between a specific number included in the output of a simulation and the corresponding true value. If the true value of a quantity is not determinate, this gives rise to uncertainty. Roy distinguishes between aleatoric and epistemic uncertainty: While the former arises from random processes in nature, the latter derives from our ignorance. Because the true value of a quantity is often not known, the related error is uncertain and can only be estimated. Errors are often classified according to their sources. For instance, round-off errors arise because digital computers cannot properly represent irrational numbers and effectively work with rounded numbers. Roy explains how the various kinds of errors can be tested for and estimated. He also discusses how the estimates of the various types of errors can be combined as to yield a total error for a simulation output.

Part II turns to foundations of validation in a slightly different sense. The broad idea is to understand validation in terms of frameworks that philosophers of science have proposed for scientific method and inference. Very roughly, the scientific method comprises those rules and recipes that scientists qua scientists should follow in their research. At a fine-grained level, we find several methods such as observation, experiment, thought-experiment, or computer simulation, each coming with its own methodological rules. But some philosophers have thought that scientific method, in particular, scientific inference, can be characterized at a more general level. If this is correct, then the general rules of scientific method apply to computer simulation too. The question of this bundle of chapters thus is what prominent general philosophical accounts of scientific method and inference imply for validation and whether this is useful advice.

The first account under scrutiny is falsificationism or the Popperian philosophy of science. Popper‘s approach is in fact in high esteem in many scientific circles (see, e.g., Godfrey-Smith 2003, p. 57). Although mostly concerned with empirical science, Popper was convinced that scientists cannot, and should not, inductively infer theories, since he took induction not to be rational. What is crucial for Popperian science is instead the attempt to subject hypothesis and theories to severe testing. Once a theory has been shown to be false (or falsified) because it contradicts what we observe, it should be rejected. Popper did not consider computer simulation, specifically, but if we apply his outlook to simulation science, then the focus should be on invalidation rather than on validation, as Keith Beven and Stuart Lane point out in Chap. 6 titled "Invalidation of Models and Fitness-for-purpose: A Rejectionist Approach". Indeed, testing simulation models for all kinds of problems is common practice. However, as the authors show further, strict invalidation is often not possible because simulation models involve many uncertainties. Beven and Lane thus explain methods with which uncertainties can be handled while sticking with the idea that invalidation should be the aim. They admit that invalidation becomes more subjective in this way, but point out that being explicit about the assumptions of analysis will help facilitate communication with practitioners and help avoid the use of simulations that are not fit-for-purpose in practical applications. They ultimately recommend invalidating computer simulations in science and argue that this opens up new avenues of research (see also Chap. 27 by Roache for a discussion of the falsificationist perspective on model validation).

A falsificationist outlook in philosophy of science is often criticized because it does not allow for the positive confirmation of hypotheses as, e.g., suggested by simulations. The Bayesian approach to philosophy of science, also called Bayesian epistemology, by contrast, can describe how trust in scientific results is built. The basic premise underlying Bayesian epistemology has it that belief (or trust) comes in degrees and can be measured in terms of probabilities. The probabilities express how (un)certain an agent feels about a certain proposition. As new evidence comes in, the probabilities are updated, e.g., using Bayesian conditionalization. The question of Chap. 7 by Claus Beisbart ("Simulation Validation from a Bayesian Perspective") thus is whether Bayesian epistemology provides a useful framework for doing validation. After a presentation of the basics of Bayesian epistemology, the author argues that data-driven validation can be conceptualized using Bayesian updating. For Beisbart, Bayesian epistemology offers more than that because it can also describe how the prior credibility of the conceptual model underlying a simulation impacts on the credibility of the whole simulation. A problem in this respect is, however, that a computer simulation deviates from the conceptual model due to approximations. Beisbart argues that this problem can, too, be solved in Bayesian terms. In the final discussion, the Bayesian account of validation developed in the chapter is systematically assessed. Beisbart admits that some general objections against Bayesian epistemology, e.g., the so-called old evidence problem, are relevant to a Bayesian approach to validation. However, he argues that Bayesians are in a better position to conceptualize the inferences implicit in validation than are falsificationists.

In the second half of the twentieth century, philosophy of science underwent what is often called a historical turn. Following, in particular, Thomas Kuhn, philosophers turned their attention to the history of science. Kuhn himself distinguished three types of phases of development within science, viz., pre-paradigmatic research, so-called normal science, and scientific revolutions. In Chap. 8, titled "Validation of Computer Simulations from a Kuhnian Perspective", Eckhart Arnold discusses the consequences of Kuhnian philosophy of science for validation. After a short introduction to Kuhn's outlook, Arnold argues that we should resist the temptation to call the development of the method of computer simulation revolutionary in Kuhn's sense. But it might still be argued that computer simulation requires a new paradigm of validation. Arnold rejects this idea, too, and stresses that the validation of computer simulation is not in principle different from attempts to confirm a theory. In the last part of his paper, Arnold has a closer look at simulations in the social sciences. He observes that some agent-based simulations are not properly validated and that there is not even much of an interest in validation in related fields. He thus concludes that these fields from the social sciences are pre-paradigmatic in Kuhn's sense.

A fourth philosophical approach that is applied to validation is hermeneutics (Chap. 9). The latter is sometimes thought to be the art, or the theory, of understanding with a clear focus on the understanding of texts and other products of human culture. But hermeneutics can also be seen as a philosophical movement or school, with Dilthey, Heidegger, and Gadamer being its most famous proponents. At first sight, a hermeneutical approach to validation may not seem very promising because hermeneutics appears more appropriate to provide foundations for the humanities, which have human culture and its products as their main subject matter. But in an often-quoted paper, in which Kleindorfer et al. (1998) consider several philosophical positions to obtain advice for validation, they come to sympathize with a hermeneutical outlook on validation. This is motivation enough for Nicole J. Saam ("Understanding Simulation Validation—The Hermeneutic Perspective") to have a closer look at the potential benefits and pitfalls of a hermeneutic approach to validation. She distinguishes between hermeneutics *in* validation and hermeneutics *of* validation. While the former tries to use methodological advice from hermeneutical scholarship within the practice of validation (as Kleindorfer et al. suggest), the latter tries to understand validation itself as a practice. Saam argues that hermeneutics in validation is a failure; what hermeneutics implies about the validation of computer simulation is at best vague, if not inappropriate. A hermeneutics *of* validation, by contrast, is found to be viable. Saam argues that the hermeneutics of validation may even be useful for working scientists. She thinks that such a hermeneutics can proceed via interdisciplinary dialogue in which historically influential “prejudices” about validation in certain disciplines may be corrected.

1.3.2 Methodology (Parts III–VI)

Parts III through VI turn to the methodology of validation itself. The basic aim of these parts is to give an overview of the various strategies and techniques that are available for validation and to discuss their viability. Part III considers steps that have to be taken prior to validation proper, but that are very important for validation.

A lot of simulations are built upon independent models such as, e.g., the Ising model. In such cases, the accuracy and the credibility of the results of the simulations depend on how credible the models are. Chap. 10 by Axel Gelfert, titled “Assessing the Credibility of Conceptual Models”, analyzes the credibility of such models. His questions are what credibility of a model amounts to and how credibility can be achieved. He proposes to understand the credibility of models following the credibility that we assign to people in everyday talk. It is common that a person is a credible source of information for certain questions and in certain contexts, while less so in others. Likewise, Gelfert suggests, a model may be credible in some contexts, while not in others. He thus stresses that the credibility of a model crucially depends on the purposes and research questions that researchers have in mind when working with a model. A similar point has been made by Parker (2009) when she has suggested that models and simulations should be evaluated following their adequacy for purpose. Gelfert distinguishes two broad classes of purposes, viz., the representation of real-world systems and the exploration. Concerning the first aim, he argues that fit with empirical data and the faithful representation of mechanisms underlying a range of phenomena in the target system are important criteria for model credibility.

The best model does not help if a computer simulation fails to yield approximate solutions to it. If there is a bug in the program, for instance, then the simulation program does not provide what it is supposed to. This is where so-called verification steps in. Roughly, verification aims to ensure that the computer simulation does yield approximate solutions to the model that the researchers intend to use. While Oberkampf in Chap. 3 and Murray-Smith in Chap. 4 cover the basic principles that guide verification, Part III includes two chapters that deal with the verification of a special type of simulation. The reason is that verification is particularly difficult if the original model (often called the conceptual model) consists of systems of ordinary or partial differential equations (ODEs or PDEs, for short). Such equations involve variables that may vary in a continuous way and differential operators such as the first derivative in time. Due to their digital nature, computers cannot correctly represent such differential operators. Consequently, the ODEs and PDEs need to be discretized, which is to say that the differential operators are approximated. This leads to characteristic errors often called discretization errors (cf. Chap. 5 by Roy in this volume). The chapters that we have commissioned about verification for Part III deal with related problems, which arise if models consist of ODEs and PDEs.

In Chap. 11 ("The Foundations of Verification in Modeling and Simulation"), William J. Rider introduces the basic techniques of verification that are relevant in this setting. He draws a basic distinction between two steps of verification: code verification and solution verification. Code verification is aimed at making sure that the output of a simulation program does approximate the solutions to the model under investigation. As Rider points out, this can be done by showing that the outputs of the simulation program converge to a correct solution, as the so-called resolution is increased. To show this, researchers run the program with several resolutions (e.g., with different time steps) and compare to a benchmark solution. The difference between the output and the benchmark is called error, and the mark of code verification is that the rate with which the error becomes smaller is consistent with a certain theoretical expectation. Rider considers various benchmarks and stresses that they need case-specific documentation. Solution verification, by contrast, is supposed to determine the errors associated with the output of a computer simulation. Rider emphasizes that, during solution verification, typically, no benchmark solutions are available. This is to say that there are more unknowns. The basic idea then is to fit a simple numerical model for the convergence to the output obtained. Both the methods of code and solution verification are illustrated using a concrete example. Note that Oberkampf (Chap. 3) and Roy (Chap. 5), too, comment on both steps of verification.

Chapter 12 ("The Method of Manufactured Solutions") by Patrick J. Roache is focused on a novel technique that can be used during code verification. As just mentioned, the latter is based upon comparing the outputs from a simulation with correct solutions to the equations of the model implemented in the simulations. For complicated ODE or PDE models, such solutions are barely available. This is where the method of manufactured solutions can step in. As Roache points out, the central idea is to assume that an arbitrarily chosen function solves a variation of the model equations. To obtain the modified model equations, a source term is added that compensates for the fact that the equation does not hold for the chosen function. The advantage is that the chosen function is known to be an exact solution of the modified equations. In this way, the usual tests constitutive of code verification can be applied. In his chapter, Roache describes the method in detail by considering simple examples. As he points out, it is important that all terms from the original model equations are exercised by the chosen function. By contrast, it is not an issue whether the chosen function is realistic regarding the applications. The last few sections of the chapter discuss the broader significance of the method.

The chapters of Part III explain further preparatory steps on which the practice of validation is based. These steps include, for example, the choice of a validation metric. This is the topic of Chap. 13 by Robert E. Marks "Validation Metrics: A Case for Pattern-Based Methods". A validation metric is a distance measure that quantifies how far the outputs from a simulation are from measured data from the target system. The choice of a validation metric is of utmost importance because the measure determines which kinds of deviations between simulation output and data are taken to pose a problem for the simulation. In the simplest case, a validation metric may just be the modulus of the difference between the output value, say for pressure at some time, and its measured value. But it is often more appropriate to construct a measure that takes into account not just one instance of time and not just one characteristic such as pressure. Further, both the simulation outputs and the measured data are subject to uncertainties. If the latter are expressed using probabilities, then probability models arising from the simulation output and the measured data need to be compared. In his chapter, Marks first lists properties that have been proposed for validation metrics. He then distinguishes between several families of measures that are defined for the comparison of probabilities. His focus is on what he calls "pattern-based measures". One subgroup of the pattern-based measures is based upon information theoretic terms, e.g., Shannon entropy. A second group is constituted by strategic state measures. Marks' focus is on the state similarity measure that he has suggested elsewhere. Its application is illustrated using a simple agent-based model for interacting brands. A very comprehensive treatment of validation metrics can be found in Oberkampf and Roy (2010, Chap. 12). A short overview of validation metrics is also contained in Chap. 17 by Saam in this volume.

Before we can compare the output of a simulation with data using a validation metric, we have to make sure that the simulation output is properly produced and well understood. This is particularly an issue for stochastic simulations. In Chap. 14, "Analyzing Output from Stochastic Computer Simulations: An Overview", Christine Currie explains how to run stochastic simulations properly and how to preprocess the output. She distinguishes between terminating and nonterminating simulations. As the name suggests, terminating simulations are finished if a certain type of event occurs. This requires some care to set the initial conditions. Nonterminating simulations often run into what is called a steady state. If only the steady state is of interest, then a so-called warm-up period of the simulation output has to be removed. Currie details two possible methods. She further explains how often a stochastic simulation has to be run to obtain a certain level of significance. The chapter is self-contained as all required statistical notions are defined.

Results of computer simulations are usually validated by comparing them to various types of reference points. The chapters of Part IV describe reference points that are commonly used as well as related techniques. The reference points considered include data, stylized facts, and the users' judgements. In addition, there is a chapter that proposes to conceptualize the comparison with reference points in terms of the terminology known from benchmarking.

Data are clearly the most important reference point, thus David J. Murray-Smith, in his Chap. 15 ("The Use of Experimental Data in Simulation Model Validation"), considers the use of experimental data for the validation of deterministic simulation models. Of course, many computer simulations have target systems, on which the researchers cannot run experiments. But even if the target system cannot be experimented on, experimental data from other systems that resemble the target system may provide insight. Murray-Smith discusses graphical and quantitative methods of comparing data sets from the target system with output from the simulation model. But his main emphasis is on the design of experiments to obtain suitable data. He argues that so-called identifiability analysis (which is focused on the possibility of estimating model parameters using data) can provide valuable information for experiment design. Murray-Smith emphasizes that well-designed validation test specifications should include initial conditions and all boundary conditions, the form of inputs and the measuring equipment used. All relevant operating conditions and the full range of parameter values must be considered. It is also important to specify and to take into account the accuracy of the data. Murray-Smith uses a model of the pulmonary gas exchange processes in humans to illustrate these issues. While Chap. 15 is concentrated on the validation of deterministic models, this volume contains several chapters that deal with stochastic simulations, e.g., Chap. 13 by Marks, Chap. 14 by Currie, and Chap. 26 by Mättig.

Besides experimental data, there are other kinds of data, in particular, observational data and historical data. There are some other chapters in this volume which deal with the use of observational data, in particular, Chap. 29 by Theis & Baldauf and Chap. 30 by Rood who describe the validation of weather forecasts resp. climate models. Chapter 33 by Seibert and co-authors demonstrates the use of observational data in the validation of hydrological modeling. The peculiarities of historical data are addressed by Köstlbauer in Chap. 36.

Matthias Meyer (Chap. 16, “How to Use and Derive Stylized Facts for Validating Simulation Models”) considers stylized facts, which have become prominent in economics and other social sciences. This concept was introduced by economist Kaldor who suggested that theorists “should be free to start off with a stylised view of the facts – i.e. concentrate on broad tendencies, ignoring individual detail” (Kaldor 1968, p. 178). Stylized facts may also be called patterns, empirical regularities or statistical properties of a phenomenon (Grimm et al. 2005). Meyer clarifies the concept and proposes to define stylized facts as broad, but not necessarily universal generalizations of empirical observations that describe the supposed essential characteristics of a phenomenon that require explanation. Meyer argues that stylized facts can not only provide a reference point for model construction ex ante but also for the purpose of model validation. In addition, he addresses the question of how to obtain stylized facts. Four approaches to establish stylized facts are presented and assessed, among them a new, particularly transparent approach which has been developed by Meyer and his colleagues.

Sometimes judgments are recommended as reference points for the validation of simulation. Most often, these are supposed to be expert judgements, but in certain areas of research, the users’ judgements may matter too. An example of such an area is so-called action research, which is considered in Chap. 17 by Nicole J. Saam (“The Users’ Judgements—The Stakeholder Approach to Simulation Validation”). This is a type of research in the social sciences which is supposed to foster social change and thus involves the participation of citizens. The researcher is not a detached observer anymore, but instead an active participant. In this area, a so-called stakeholder approach has been suggested for simulation validation. Saam uses the validation of socio-ecological simulation models as a typical example of action research that involves the stakeholders in simulation validation. A model of a socio-ecological system is typically supposed to represent the stakeholders themselves, in particular, their behavior and their tacit knowledge, as well as the environment in which they live. Saam argues that action researchers have to consider the stakeholders' judgements as an indispensable point of reference for simulation validation, if they submit to a constructivist view of social reality. Not only are stakeholders needed to ensure that their tacit knowledge is reflected in the simulation; rather, they have also to accept the simulations and to act on the results, if the simulations are to be successfully used. To obtain the stakeholder's judgements in such a framework involves repeated efforts of communication, which requires a strong background in qualitative methods of empirical social research as well as gaming simulation.

The last chapter in this row is not about a specific reference point but rather introduces a general framework for thinking about the comparison with reference points, viz., the terminology of benchmarking. Benchmarking is well known from, e.g., management, but has occasionally been used and discussed in the context of validation too. In Chap. 18, titled "Validation Benchmarks and Related Metrics", Nicole J. Saam discusses validation in the terms of benchmarks and benchmarking. Benchmarking in general is defined as the evaluation of a performance that crucially involves comparison. Saam distinguishes between benchmark variables, which capture those aspects that are of interest for the evaluation, and proper benchmarks, i.e., determinate threshold values of the benchmark variables. The focus of the chapter is on a descriptive account of benchmarking in simulation validation. Saam offers a typology of benchmarks and argues that benchmarking allows for some flexibility such that standards of varying degrees of strictness can be accounted for. She also stresses that benchmarking often involves a social component: It is frequently groups of people who have to agree on benchmark variables and benchmarks proper, if some benchmark is supposed to make a difference for a practice. Saam distinguishes validation metrics and benchmarking metrics, which both can be used as a measure that takes the distance with respect to some benchmark.

Validation deals with various kinds of uncertainties. This raises a lot of questions, e.g., how the uncertainties should be expressed and assessed and what sorts of inferences they still allow. There are several mathematical approaches from statistics that answer these questions. The chapters from Part V explain these approaches and apply them to validation.

Hypothesis testing using frequentist statistics is a method that is well known in many branches of science and used for instance in the analysis of experimental results. In his Chap. 19, "Testing Simulation Models Using Frequentist Statistics", Andrew Robinson discusses this method for the validation of computer simulation. He first reviews the method in general terms. In the simplest variety of the method, a null hypothesis is formulated, and a rejection region for some characteristic is defined. If the null hypothesis is true, then it is very unlikely that the characteristic is measured to be in the rejection region, and thus the null hypothesis is rejected if the measured characteristic takes a value in the rejection region. This method is often supposed to spell out the idea of Popperian attempts at falsifying a hypothesis (cf. Chap. 6 by Beven and Lane). During validation, it is natural to propose a null hypothesis according to which the simulation output matches measured data to a certain accuracy. However, as Robinson argues, this way of applying the method leads to problematic results; for instance, a computer simulation may be rejected too quickly if the data is sparse. Robinson thus proposes a different way to apply frequentist statistics to simulations. The essential trick is to turn to so-called equivalence tests, under which the role of the null hypothesis is reversed. The content of the latter then is that the computer simulation is invalid. Robinson details the application of his preferred method using two examples.

Hypothesis testing can also be done using Bayesian methods. In Chap. 20 ("Validation Using Bayesian Methods"), Xiaomo Jiang and co-authors detail related methods for simulation models. The crucial idea is to accept or reject a hypothesis on the basis of quantities well known from Bayesian statistics. This idea moves beyond mere updating one's probabilities, which is in the focus of Chap. 7 by Beisbart. To decide on a hypothesis, the so-called Bayes risk is formed. It is something like a negative expected utility and based upon information on what the costs are for taking certain decisions. This function needs then to be minimized. It turns out that this is equivalent to checking that the so-called Bayes factor is larger than a certain value. The chapter considers interval hypothesis testing for the uni- and the multivariate case and summarizes a lot of technical results that are useful in this regard. In particular, the Box-Cox transformation is introduced to deal with non-normal data. Further, Bayes networks are introduced. Jiang and co-authors use several examples to illustrate Bayesian hypothesis testing in validation.

Both frequentists and Bayesians use probabilities that are assumed to have precise values. But in the literature about modeling, Frigg et al. (2014) have argued that it is at best misleading to handle so-called structural model error (i.e., errors about the structure of the model equations) using probabilities. A possible way out are imprecise probabilities. In the literature about validation, they are in fact referred to by Oberkampf and Roy (2010). The aim of Chap. 21, "Imprecise Probabilities" by Seamus Bradley is to introduce imprecise probabilities and to discuss their application to validation. As Bradley points out, the basic idea is that events or propositions are assigned a set of probabilities, typically an interval. Bradley explains how such imprecise probabilities are dealt with mathematically and how they may be interpreted. He further points out problems with interval probabilities. One problem is that it is not quite clear how we can make decisions on the basis of interval-valued probabilities. Bradley's conclusion is nevertheless that imprecise probabilities provide a promising framework for thinking about validation.

In Chap. 22, titled "Objective Uncertainty Quantification", Ed Dougherty and co-authors propose a framework that is useful if validation proves infeasible. This is so if there are significant uncertainties about the simulation model itself. The challenge then is to quantify these uncertainties. The authors take the perspective of engineering where the aim is to achieve a certain goal with a system that is insufficiently known. The uncertainties arise as uncertainties about values of model parameters that describe the system of interest. The framework they propose is Bayesian because subjective probabilities are used to express the uncertainties. The main idea is to define a cost function and to find the best way of achieving the goal on average, relative to the cost function, given the probabilities. It is then possible to express the costs of the uncertainties. Dougherty and co-authors explain their method using the example of regulatory gene networks and experimental design. Note though that their framework is not one for validation itself.

The validation of simulation models is likely to become a permanent task in many disciplines. The chapters in Part VI "The Organization and Management of Simulation Validation" describe legal prescriptions as well as administrative and procedural activities related to simulation validation as a permanent task.

Based on experience from environmental meteorology, Heinke Schlünzen (Chap. 23, "Standards for Evaluation of Atmospheric Models in Environmental Meteorology") explains how a shared standard on simulation validation can be obtained, starting from developing evaluation guidelines. In the field of environmental meteorology, evaluation was a long-standing issue, since in this field model results are used to take decisions relevant for humans and the environment. In the process of guideline development, it turned out that verification, validation and evaluation were differently understood. Securing agreement on definitions of these terms was important. An evaluation *guideline* helps to harmonize different approaches and aims at determining what counts as sound practice. However, such a guideline is not mandatory or legally binding and thus cannot be enforced. It is only a standard that serves as a norm. Schlünzen describes how such a standard was obtained in environmental meteorology. Relevant stakeholders were involved to achieve an agreement supported by a broader community. Schlünzen highlights that developing a standard often requires compromises between desiderata based upon science and practical needs and practicability. She proposes a generic outline for developing a model evaluation guideline: First, the application area has to be specified. Second, the model developer has to define evaluation steps. Finally, the model user has to work on these evaluation steps. This three-step procedure is detailed using examples from environmental meteorology.

If simulations become very complex, then their validation becomes so challenging that they need professional management. Validating such simulations needs tremendous efforts involving numerous steps and even several teams of scientists. Validators have to consider various types of physical devices and models coming from different areas. Models may be distributed on different computers. To guide users to manage the validation of complex simulation systems professionally, Fei Liu and Ming Yang, in their Chap. 24 ("The Management of Simulation Validation"), present principles for simulation validation and a management framework. For instance, important principles state that validation of a model is conducted with respect to its purpose, and that validation must be conducted throughout the whole life cycle of a simulation system. These and further principles are introduced and discussed. In their management framework, Liu and Yang adopt a process-oriented view. They distinguish four components: process, scheme, metrics, and tools—each addressing verification and validation (V&V). In particular, Liu and Yang argue in favor of V&V schemes based on optimization techniques, a tree-like V&V metric system distinguishing top metrics, bottom metrics, performance measures, and evaluation values, and the computer-aided management of simulation V&V.

The validation of complex simulation models is also the background of the next chapter titled "Valid and Reproducible Simulation Studies—Making It Explicit"

(Chap. 25). Oliver Reinhardt and co-authors emphasize that a multitude of data sets and a lot of additional information accompany the validation process, e.g., output from several runs of the simulation program, data used during calibration, input data, information about test cases, and so on. Also, as simulation models are rarely developed from scratch, but often reuse existing models, there is information on these prior models and their validation. Reinhardt and co-authors summarize all these different kinds of information and data as "artifacts". To make these artifacts and their mutual relationships accessible, Reinhardt and co-authors propose the use of a declarative formal modeling language for simulation, as well as a declarative language for specifying and executing diverse simulation experiments. They emphasize that all information that is important for the validator to establish trust in the model should be made explicit. To this purpose, they propose a provenance model. Such a model does not only cover the artifacts but also the processes through which they were obtained. The authors present the development and validation of an agent-based model of migration from Senegal to Europe to illustrate their argument.

1.3.3 Validation at Work—Best Practice Examples (Part VII)

The aim of Part VII is to present best practice examples that demonstrate how the methods and techniques of validation are applied in various disciplines and with different types of simulation models. Depending on the discipline and the simulations, the examples cover different aspects of the validation process. While some chapters concentrate on the use of experiments (Chaps. 26 and 27), others focus on standards for simulation validation (Chap. 27), model-to-model comparison (Chap. 28), idealized test cases (Chap. 29), and on the validation culture of a whole research field (Chap. 30), or on techniques for calibration, estimation, input, and output validation (Chap. 31). But the chapters, too, show significant differences in the broader outlook on validation.

The first chapter in this part comes from the scientific investigation of the smallest constituents of the physical world, viz. particle physics. Experimental particle physics has seen spectacular breakthroughs in the last few years; in particular, the so-called Higgs particle was discovered using the ATLAS and CMS experiments at the Large Hadron Collider at the CERN. Experimental physics is heavily based upon computer simulations. Thus, in Chap. 26, "Validation in Particle Physics Simulation", Peter Mättig describes validation in this field using the ATLAS experiment as an example. As he points out, the simulations are primarily used to obtain model predictions that can be compared to data. The simulations have two crucial components, viz., the so-called physics generators that model the collisions between the particles and their decays, and the simulation of the detector. Mättig explains in detail how these components are validated, for instance, using data from previous precision measurements. The validation of computer simulations is particularly interesting because it is an integral part of experimental research. The experiment itself needs validation, and computer simulations are used for this validation. So, to some extent, the experiment and the simulations are validated by cross-comparing them to each other. For Mättig, there is nothing problematic about this.

The next chapter, Chap. 27 by Patrick J. Roache ("Validation in Fluid Dynamics and Related Fields"), turns to fluid dynamics. Since fluid dynamics describes flows of liquids and gases, it has a huge range of applications, including aerodynamics, weather forecasting, ocean currents, blood flow in artificial hearts, etc. In all these applications, a medium (e.g., ocean water) is described as a continuum due to a clear separation of scales between molecular motion, on the one hand, and the continuum flow, on the other. The latter is described in terms of aggregated quantities such as velocity, pressure, density, and temperature. Classical fluid dynamics is based on the conservation of mass, momentum, and energy, expressed in integral or differential form. This leads to the Navier–Stokes equations, a set of partial differential equations, which most often do not have closed-form solutions. These equations are included within much more complicated models as used in astrophysics (Chap. 28), weather forecasting (Chap. 29), and climate science (Chap. 30). Thus, the validation concepts described by Roache in this chapter are of broader relevance; they are in fact useful for all simulations that use PDEs. In the main part of his chapter, Roache explains ASME (2009), V&V 20-2009, which is an American National Standards Institute Standard document. Current validation practice in fluid dynamics is based upon this standard. The focus of V&V 20-2009 is on "unit problems", which isolate one simple physical system rather than complex systems. Roache also considers a new paradigm of experiments designed specifically for validation. This paradigm recognizes that requirements for validation are distinct and that validation experiments are much easier than traditional experiments in some respects, but more demanding in others.

Astrophysics is a typical field in which experiments cannot be done on the target. Alan C. Calder and Dean M. Townsley (Chap. 28, "Astrophysical Validation") present two studies aimed at validating components of Flash, a freely available, parallel, adaptive mesh simulation code used for modeling astrophysical phenomena and other applications. They first present a study of validating the hydrodynamics routines in Flash with experiments that replicate the high energy density environments of astrophysics in a laboratory. Calder and Townsley stress that a quantitative comparison between the simulations and the experiments and the determination of the uncertainty required close collaboration between experimentalists and theorists. The second study addresses thermonuclear combustion in supernovae explosions. The validation is to some significant extent based upon a model-to-model comparison: A simplified model is tested against higher fidelity models for a given physical process. Calder and Townsley clarify that this type of comparison combines elements of verification and validation.

Weather forecasting is a very prominent application of computer simulations for making predictions. As citizens, farmers, companies, etc. need weather forecasts all over the world and every day, there is an immense experience about validating weather forecasts. In their chapter on the validation of weather forecasts (Chap. 29, "Validation in Weather Forecasting"), Susanne Theis and Michael Baldauf explain that the atmospheric model that underlies the simulations can be partitioned into various parts, in particular, the dynamical core—i.e., partial differential equations and their numerical solver—and the parameterizations. During validation, first the dynamical core is tested via idealized test cases, then parameterizations are added. Finally, the output of the whole program is compared to observed weather. Theis and Baldauf explain how meteorologists deal with forecast uncertainty. Some of the sources of forecast uncertainty are known. They are sampled from a realistic range of options. In particular, it is well known that initial conditions are a major source of uncertainty. This is due to the chaotic nature of the atmosphere. Another source of uncertainty are imperfections of the model. Meteorologists have established the practice to transform the outcome of the ensemble forecast into probabilistic forecasts. The quality of these forecasts is then assessed in a statistical manner.

Climate science is another field in which experiments cannot be done on the target. A major difference to astrophysics is that climate models are highly relevant for political decision-making. Politicians need validated models and results to decide on societal attempts at mitigating climate change or to plan measures of adaptation. Validation of climate simulation models is considered in Chap. 30 by Richard B. Rood ("Validation of Climate Models: An Essential Practice"). As he notes in his introduction, his field was challenged by the paper by Oreskes et al. (1994), which argued that numerical models of geophysical phenomena cannot be validated. Rood doesn't agree with this verdict and replies that it is meaningful only in an abstract, philosophical sense. It is at odds with evidence for the successful use of models and their ubiquitous and successful applications in society. In the main part of the chapter, he gives a most encompassing description of validation practice in his field. Rood describes how different aspects, such as validation criteria, independent observational data, validation metrics, and statistical models as well as evaluations of physical consistency are interrelated in the practice of climate model validation or evaluation, as some climate scientists prefer to call it. He emphasizes that verification and validation processes are not purely quantitative. Rood describes evaluation as an iterative, deliberative process. He even uses the concept of deliberative validation to highlight the role that expert judgments have in this practice. Rood also stresses the emergence of intercomparison projects which have promoted the development of shared standards of evaluation. And he addresses issues of the management of simulation validation, such as validation plans and protocols. Rood emphasizes that a culture of verification and validation has emerged in the climate-modeling community, and that validation/evaluation has been established as an essential practice in climate science.

Whereas the previous chapters from this part have basically dealt with models using nonlinear partial differential equations, the next chapter, Chap. 31, by Giorgio Fagiolo and co-authors ("Validation of Agent-based Models in Economics and Finance"), turns to a different type of model. Agent-based models are becoming increasingly popular in economics and other social sciences. These models are not based upon approximations of differential equations. Thus, crucial parts of verification as understood and practiced for ODE and PDE models are neither possible nor necessary. In their chapter, the authors first introduce agent-based models. They emphasize that the complex microeconomic interactions and the presence of ubiquitous nonlinearities (even in the simplest models) do not allow one to obtain closed-form solutions. Then they elaborate a theoretical framework for validating agent-based models that is in stark contrast with validation as conceived of in mainstream economics, where a falsificationist outlook prevails. For Fagiolo and co-authors, validation is different from falsification and does not fit into a binary framework of rejecting versus not rejecting the model. Instead, validation comes in degrees, which also allows researchers to judge to what extent a model performs better in comparison to another. This allows for choosing among alternative model specifications. Notwithstanding the differences between calibration, estimation (which are not the focus of this volume), and validation, Fagiolo and co-authors discuss all three methods and related techniques. They distinguish between what they call input validation and output validation. The former addresses the assumptions of the model as well as the initial conditions and assesses the impact of different parameters on the dynamics of the model. This is needed because agent-based models are not based on a well-confirmed conceptual model. By contrast, output validation, as referred to by Fagiolo and co-authors, corresponds to what is simply called validation in many other chapters of this volume. The chapter reviews numerous techniques for both ways of validating agent-based simulation models.

1.3.4 Challenges in Simulation Model Validation (Part VIII)

The chapters of Part VIII cover important practical challenges that simulation scientists face when applying the methods and techniques described in Parts III through VI. These challenges go back to peculiarities that are specific of either some type of simulation models or some discipline.

The first challenge addressed in this part has been called equifinality. Equifinality means, roughly, that simulations that differ in some respects produce very similar output and thus fare equally well as regards observational data. In his Chap. 32, “Validation and Equifinality”, Keith Beven discusses equifinality using examples from the environmental sciences. Results from many computer simulations with a variety of hydrological models have shown that equifinality is generic to modeling in this area. Beven explains the peculiarities of the inexact sciences with respect to errors and uncertainties in the input and observational data, and points out that this limits the use of traditional statistical hypotheses testing. Beven then introduces the Generalized Likelihood Uncertainty Estimation (GLUE) method which is related to Bayesian methods of validation, in particular, to Bayesian updating. This method rejects the idea of one single optimal solution. Rather, an evolving ensemble of models considered acceptable is identified. This ensemble is assumed to be useful in prediction and can be refined as new information becomes available over time. The basic idea behind the GLUE analysis is, roughly, to start from a prior set of models, parameters, and variables. Simulations are run and solutions obtained. For every single simulation, parameter sets are rated according to the degree to which they fit observed data. The ensemble of models is then divided into two ensembles, one with non-acceptable solutions and a second one with acceptable solutions. Eventually, as more observations become available, the parameter sets in each ensemble can be updated.

GLUE allows that each acceptable solution is assigned a likelihood measure. If its output is considered to be unrealistic, the simulation is rejected as having zero likelihood. In this way, researchers obtain a discrete joint likelihood function for all the models, parameters, and variables. Beven argues that GLUE allows for testing models as hypotheses because models that do not provide simulated values within the limits of acceptability will be rejected. Finally, the ensemble of acceptable models is used to produce likelihood weighted simulations. Beven emphasizes that it is the *set* of parameters that will be considered as acceptable or not within a given model structure. Thus, within this framework, model structures can be compared and combined, or one model structure can be chosen over another.

The next chapter resumes the discussion of equifinality and discusses its relation to underdetermination, over-parameterization, as well as other problems such as over-fitting. In philosophy of science, underdetermination is used to describe situations in which there is insufficient evidence to decide between different theories. Over-parameterization is present when more parameters are used in a model than can be identified based on the available information. In mathematical terms, an example is when a polynomial with more than n free parameters is fitted to n data points. Jan Seibert and co-authors (Chap. 33, "Validation and Overparameterization—Experiences from Hydrological Modeling") explain that the term is used a bit differently in environmental modeling, where it refers to situations in which a model contains more free parameters than are identifiable with confidence. As it happens, most environmental models suffer from over-parameterization. It is a serious problem in environmental modeling, as this means that a model may work well for the wrong reasons. Seibert and co-authors discuss different ways to validate models that simulate hydrological processes at the catchment scale. They point out that the balance between model testability and over-parameterization has to be considered.

The next challenge is posed by long-term predictions. Obviously, such predictions cannot be confirmed in terms of empirical data now and are very uncertain. But how can the uncertainties be assessed? In their Chap. 34, "Uncertainty Quantification Using Multiple Models – Prospects and Challenges", Reto Knutti and co-authors consider coordinated model intercomparisons which have been established by the climate-modeling community. In these intercomparison projects, models from ensembles are evaluated against each other. The ensembles are used to explore uncertainties either by testing the robustness of projections or as a basis for statistical methods that estimate the uncertainty about future climate change. Here, a model projection is called robust if it is produced by most models in the ensemble (where of course robustness does not imply accuracy). Knutti and co-authors point out weaknesses of this approach. One problem is that it treats all models as independent and equally plausible. Speaking figuratively, each model has one vote, as have citizens in democracies, which has led to talk of "model democracy." Knutti and co-authors argue that model democracy becomes harder to justify. For instance, often parts of models are reused in other models, which leads to a violation of the independence condition. As a solution, Knutti and co-authors suggest reweighting all models of the ensemble for performance and dependence.

In the remainder of this part, we turn to challenges that pertain to specific disciplines. Some social sciences (e.g., sociology, political science) and historiography face particular difficulties already when developing formal models. In all these disciplines, there are no fundamental laws of (social, political or historical) dynamics. Also, many social scientists do not assume realism, i.e., they do not think that they describe a reality that is independent from their theories and models. They also reject the idea that the social sciences are based upon an epistemology that is modeled after the natural sciences. For instance, they do not think that social sciences should aim at theories, at mathematized models and predictions, etc. (see, e.g., Kertész 1993). They opt for a pluralism in perspectives how the same objects, i.e., social phenomena, should be scientifically investigated: Approaches from the social sciences differ considerably in their ontological, epistemological, and methodological assumptions. Concerning validation in the social sciences, they hold what Nickles has claimed about theory competition in science studies: "There is not just one thing in dispute here, of course, but a whole thicket of nasty problems involving social construction, psychological construction, the balance of nature and nurture, realism, relativism, the relation of reason or thought to interests, whether justification is ultimately social or internal (and conventional or natural), to what extent the world is intelligible to human beings, whether or not the world is ultimately messy and uncodifiable; (…), and so on. Moreover, all of these terms are multiply ambiguous!" (Nickles 1989, p. 245). An immediate consequence for this volume is that its related chapters cannot fully reflect the pluralism present in the social sciences.

In this chapter on challenges to validation in the social sciences, Michael Mäs (Chap. 35, "Challenges to Simulation Validation in the Social Sciences. A critical-rationalist perspective") concentrates on models of social influence dynamics in networks and on one epistemological perspective, viz., critical rationalism. Note, however, that his argument is meant to apply more generally to the validation of models in social science. Mäs identifies five challenges to validation: social-scientific theories are based on many obscure concepts, many social-scientific concepts are latent (i.e., refer to an unobservable realm), the representation of time is unclear, various processes influence the dynamics in parallel, and context dependencies limit the development of general models. He then formulates four recommendations for future theoretical and empirical research that will help tackling these challenges: Modelers should compare models and identify critical assumptions, defend their assumptions, explore the scope of a model and its boundaries, and do more validation.

The last challenge addressed in this volume is the extreme case of a n = 1 type problem. The paradigmatic example of a discipline dealing with n = 1 type problems is historiography. It is no surprise then that historians have been slow to apply computer simulations, and in his chapter about validation (Chap. 36, "Validation and the Uniqueness of Historical Events"), Josef Köstlbauer first needs to consider the possible use of simulations before turning to challenges of validation, more specifically. In the philosophy of the science, there is a distinction between idiographic and nomothetic research. While the latter aims at deriving general laws that explain types or categories of objective phenomena, the first is directed at understanding the meaning of contingent, unique and sometimes also subjective phenomena, e.g., in culture and society. Historiography, Köstlbauer explains, belongs to this latter group of disciplines. Since computer simulation is perfectly suitable for nomothetic research, but not so much for understanding and for unique phenomena, there is a problem in applying computer simulations fruitfully in historiography. Historiographers do not formulate general laws nor do they rely on deductive-nomological approaches. Köstlbauer argues that this should not keep historians from exploring the potentials of computer simulations as far as this is possible. In particular, he considers three uses that even go beyond computer simulations to also encompass social simulation and games: simulations of the big-data/longue durée type which operate on a macro-level, microhistorical research using agent-based models, and digital games and simulation games. The latter facilitate reflections on the various options, and thus potential futures, that historic figures faced during their deliberation. History video games have been developed to demonstrate both historical causalities as well as the fundamentally undetermined character of history. Köstlbauer argues that in all cases, validation has the potential to make historians reflect more on evaluative assumptions, and on the ways in which they pose questions and explain processes.

1.3.5 Reflecting on Simulation Validation: Philosophical Perspectives and Discussion Points (Part IX)

Part IX of this book steps back from the various difficulties encountered in validation. It opens the perspective and offers more general philosophical reflections on validation. To some extent, this is to take up loose ends that the other chapters have left here and there. To another part, the point is to explore the significance of validation for science and its understanding in a broader perspective. While, so far in this book, philosophical inquiry has most often been employed to clarify fundamental concepts (Part I) and to frame thinking about validation (Part II), philosophers of science will now write about topics that are hotly debated in their own field. The last decade or so has in fact seen the emergence of a field that may be called the philosophy of computer simulation, and we have asked some philosophers from this area to survey crucial debates in this field and to connect them to validation.

One question that has been much debated in the philosophical literature is what type of method computer simulation is. The challenge is not so much to characterize computer simulation using a definition, as have done Hartmann (1996, Sect. 2.2) or Humphreys (2004, p. 110). The question is rather whether computer simulation qualifies as a species of another method, e.g., experimentation or thought experimentation, or whether simulation is at least very similar to one of the methods, as has sometimes been suggested. Clearly, if something like this is the case, it should have consequences for the way we think about validation of simulations, in particular, since the term "validation" has been used for some other methods too. In

Chap. 37, titled “What Is a Computer Simulation and What Does This Mean for Simulation Validation?”, Claus Beisbart considers various proposals to account for computer simulation by referring to other methods and discusses the consequences for conceptualizing validation. The idea that computer simulations are experiments is quickly dismissed by Beisbart, although authors like Parker (2008) have noted significant similarities between the validation of experiments and of simulations. A more convincing account of computer simulation takes them to be thought experiments. As Beisbart points out, this account isn’t very telling for validation because there are no established principles for validating thought experiments. He finally considers models and argues that simulations can be in many ways regarded as models. But once more, this doesn’t have telling implications for validation, because the validation of models isn’t well developed independent from simulations; quite often, model validation is in fact facilitated by computer simulations. It is not without irony when Beisbart concludes that the validation of experimentation offers the most illuminating perspective on the validation of simulations.

In Chap. 38, titled “How Do the Validations of Simulations and Experiments Compare?”, Anouk Barberousse and Julie Jebeile take a much closer look at the validation of experiments and the validation of simulations. Their question is how both sorts of validation compare to each other? The authors first propose a notion of validation that is common to both methods, the main idea being that the results of the methods comply with requirements on the part of the users. They further assume that both experiments and simulations try to represent a target system in the real world either in terms of an experimental setup in a lab or using a computer code. The main claim of the chapter is twofold: First, the methods of experiments and computer simulations differ, in particular, because the former involve “materiality” with respect to their target, while the latter do not. This is to say that the system experimented on and the target often share material properties, which is not so in simulations. But this difference, and this is the second claim, does not play out for validation. Barberousse and Jebeile use recent experiments and simulations from evolutionary biology to illustrate their findings. These findings are consistent with Parker’s (2008) diagnosis that experimenters and simulationalists use the same type of strategies to validate their results.

In Chap. 39, titled “How Does Holism Challenge the Validation of Computer Simulation?”, Johannes Lenhard takes up the issue of holism. Holism about testing or confirmation is a very important claim in philosophy of science, advanced by, e.g., P. Duhem: Many hypotheses or theories from empirical science cannot be tested or confirmed by observation, if taken in isolation. The reason is that auxiliary hypotheses are needed to connect the theories to observation, e.g., hypotheses about the working of measurement devices. So, it is really a whole of several hypotheses that is up to empirical scrutiny. Since the validation of computer simulations is concerned with confirming the results of the simulation, it is natural to ask whether holism is an issue for simulations too. Lenhard argues that this is so, the idea being that simulations form wholes that consist of many assumptions. This would not be too much of a problem if the simulation program was built up in a modular way. The whole could then be decomposed in an orderly manner. But according to Lenhard, there are two systematic reasons why modularity is threatened, even if a program was initially modular. One is that researchers tune parameters that are operative in some modules to the performance of the whole simulation, as Lenhard shows using many examples. Further, in software development, all kinds of bugs in coding are often not properly corrected; rather their effects are compensated using ad hoc fixes, so-called kluges. The effect is that the modules are interconnected with each other in a very intransparent way to form a whole that is not easily understood. As a result, it is difficult to improve a simulation program if it does not perform as intended. Lenhard further argues that holism threatens the neat separation between validation and verification that is often postulated (e.g., by Oberkampf in Chap. 3).

Validating a computer simulation is certainly a matter of evaluating them. But what exactly are the values to which we may appeal when validating a simulation? And how exactly can such values be used? These questions are addressed by Gertrude Hirsch Hadorn and Christoph Baumberger in Chap. 40 ("What Types of Values Enter Simulation Validation and What are Their Roles?"). The authors draw on a general discussion on how values shape choices faced by scientists, and apply crucial insights from this debate to computer simulations. In this way, they focus on a very broad notion of validation, which may not coincide with narrower uses of the term in, e.g., engineering. Hirsch Hadorn and Baumberger distinguish between three types of values, viz., epistemic, cognitive, and social ones. While epistemic values, in particular, empirical accuracy, determine the degree to which a result is credible, cognitive values fix how useful a simulation is. For instance, a simulation is practicable, if it is easily run, and it is useful in the sense of relevant for, e.g., explanation, if it has explanatory power. Since usefulness can become important for assessing the credibility of a simulation, some cognitive values are related to epistemic ones. While appeal to epistemic and cognitive values is quite uncontroversial, it is less clear whether scientists may draw on social values when they make choices. Hirsch Hadorn and Baumberger examine a famous argument by Rudner (1953) that favors the appeal to such values and find it convincing, even as far as computer simulations are concerned. However, as they point out, social values may only be used in a higher order function, viz., to fix the level of credibility that is needed for acceptance of a hypothesis, and not for determining how credible it is. All in all, Hirsch Hadorn and Baumberger paint a pluralistic picture of evaluating computer simulations.

Some influential authors, e.g., Oberkampf in his Chap. 3, stress that simulation results should be assessed for accuracy (in the sense of a blind prediction) when new experimental data become available. But what exactly is the epistemic value of simulation predictions that later turn out to be true? This question is particularly challenging if prediction is contrasted with calibration. A model is calibrated to some data, if model parameters are adjusted, or tuned, to that data. The question then is: Is there epistemic surplus value when new predictions of a model turn out to agree with observations, or is the epistemic significance just the same as in a case in which a model accommodates the data to which it has been tuned? This is the question of Chap. 41 by Mathias Frisch ("Calibration, Validation, and Confirmation"). The author uses Bayesian epistemology to answer his question. This first leads into a problem, because, in a very straightforward application of Bayesian updating, known data can never confirm a model. This is known as the old evidence problem. Frisch thus turns to solutions that Bayesians have proposed for the problem. He argues that the solutions do not lead to a special value of predictions. But he does finally find a setting in which the success of predictions has epistemic surplus value. This is so, very roughly, if it is not known whether a correlation between certain characteristics extends beyond a certain range of applications that has been covered by data so far.

As some previously mentioned chapters from this volume make clear, e.g., Chap. 3 by Oberkampf or Chap. 4 by Murray-Smith, verification of a simulation is crucial for validation. But what exactly is the relationship between verification and validation, both of which are often combined in what is called „V & V“? Some practitioners, e.g., Oberkampf in his Chap. 3, argue for a clean separation between verification and validation. Some philosophers, by contrast, notably Winsberg (2010), but also Lenhard in his Chap. 39 in this volume, have challenged this view to some extent. Chapter 42 by Claus Beisbart ("Should Validation and Verification Be Separated Strictly?") thus discusses the distinction between verification and validation and the relationship between both methods. His first point is that both methods are clearly different from a conceptual point of view since they have different aims. But this doesn't exclude that, e.g., one method is used in order to apply the other. Since verification is concerned with the relation between the prior model upon which a simulation is based (conceptual model) and its implementation in the computer (computational model), Beisbart argues that we should distinguish between the validation of the conceptual and the computational model. An immediate consequence is that validation of the computational model can in principle be achieved by verifying the solutions and by validating the conceptual model independently. But the required independent validation of the conceptual model is most often impossible. This means that the validation of the conceptual and the computational model go hand in hand. Beisbart reconstructs this in terms of inferences. He stresses that some prior credibility in the conceptual model and in the verification of the simulation are necessary to validate both types of models. In the final part, the author discusses Winsberg's doubts about the distinction. He concludes that they point to some qualifications but do not fundamentally threaten the results obtained earlier in the chapter.

A further issue about simulations that has been discussed in philosophical circles is whether computer simulation is significantly novel and whether it challenges fundamental ideas entrenched in present-day philosophy of science. This is of course a question that is most important for philosophers because it is related to their daily business. Philosopher Winsberg has argued that simulation is in fact in an interesting sense novel (e.g., Winsberg 2001). In our context, it is interesting to note that his argument is related to the justification of results from simulations and thus to validation. For Winsberg, validation is novel because it is downward, autonomous and motley. This view was challenged by Frigg and Reiss (2009). In Chap. 43, titled "The Multi-Dimensional Epistemology of Computer Simulations: Novel Issues and the Need to Avoid the Drunkard's Search Fallacy", Cyrille Imbert discusses the novelty of validation of computer simulation. He adopts a theoretical framework from philosopher Goldman (1999), the broad idea being that computer simulation is part of belief-generating processes in a social setting. He then discusses the ways in which such a process can go wrong. He mentions not just issues that are by now familiar to readers of this introduction, but considers also, e.g., the use of random numbers, the replicability of simulations and the access to simulation results. He stresses that the reliability of simulations is contingent on the practices of simulation scientists and encourages his fellow philosophers to take the whole belief-generating process seriously. In this way, novel issues arise, and Imbert warns his fellow philosophers not to commit the drunkard's fallacy, which would invite them to stick with those epistemological aspects of simulations that they are familiar with, or that lend themselves to an investigation in terms of their favorite concepts, methods, or questions.

1.4 Outlook

Although we hope that this volume significantly contributes to a better understanding of the validation of computer simulations, we feel that more research is needed. Let us thus make a few suggestions for avenues of future research. They are to some extent based upon our experience to put the volume together and to systematize previous literature about validation. To some part, the suggestions draw on our readings of the chapters. Needless to say that the suggestions are preliminary and not meant to be complete in any sense. Our focus is on general issues about validation.

One fundamental problem for general research about validation is that the term itself (along with related terms) is used differently in different circles. However, the disagreement is not as pervasive as to suggest that people are not talking about the same kind of thing. As Chap. 2 by Beisbart in this volume shows, various definitions of validation can be related to each other, and there is a common ground on which different views about what validation is can be compared. But there is substantial disagreement too. This is a problem first because it hinders cross-disciplinary exchange about validation, which is beneficial as it can disseminate knowledge about validation. One threat is that researchers from one field take their notion of validation as given and misunderstand what they adopt from other disciplines. A different problem is that researchers from one area may come to think that work about validation in some other area is not relevant to them because the notions of validation differ too much. Since it seems unlikely that researchers from different disciplines will quickly agree on a common understanding of validation, we propose that at least a "translation manual" be created that helps to understand each other. Such a manual would be particularly helpful given that unease about the term "validation" has led many researchers to propose alternative terms such as "evaluation" or the neologism "evaludation" (Augusiak et al. 2014).

Presumably, the most significant disagreement about what validation is, or should be, centers around the standards inherent in validation and their strictness. Very roughly, there is the following divide: In some areas, notably in applied fields, in which safety is a vital concern, people require a comprehensive evaluation of a simulation. A quantitative comparison between simulation outputs and measured data is required to obtain reliable information about the accuracy of a simulation.

Ideally, validation experiments should be designed to test the simulations (see, e.g., Chap. 3 by Oberkampf in this volume). In other fields, the mathematical rigor and the strict requirements implicit in this account are not an option. One reason is that the results of simulations are often meant to provide qualitative rather than quantitative information. As a result, the notion of accuracy does not seem to be applicable. Additionally, in the social sciences, predictions turn often out to be false because, when they become known, people change their behavior. This means that a lot of care is needed if model predictions are compared to data to validate a simulation. Moreover, data is typically quite sparse and restricted to some macro-level. Validation experiments are out of reach, so it is not possible to test the simulations rigorously using empirical data, and there is often not a well-established theory that can be used to argue for the validity of the model underlying the simulation. It is, therefore, no surprise when researchers from these fields resort to face validation and expert judgment.

The description of this divide may seem something like a caricature, and it is a variation on a well-known theme, which is roughly the difference between the exact natural sciences and the social sciences. But there is more than a grain of truth in this picture. The divide is real, and it raises a significant problem: Researchers that deal with the challenges of sparse data, etc., have an interest to make their simulations credible to at least some degree and thus have a legitimate interest in validation. When they talk about the validation of their simulations, researchers from the "other side" will likely become suspicious because they think that efforts to make, e.g., social science simulations more credible should not be called validation simply because this activity lacks rigor. In our view, this calls for research about whether, and if so, to what extent and how, the more rigorous approach to validation may be extended to fields for which it does not seem to be applicable at present. The challenge here is that there is not just one dimension in which the fields differ, but rather several, e.g., basis in mathematical theory versus no such basis, well-confirmed versus contested model assumptions, quantitative versus qualitative results; data at the micro level versus data at the macro-level, and so on. Although simulations in some field will often be at the same end of the spectrum on several of the dimensions, this is not always the case. Each dimension thus needs its investigation, the main question being whether, and if so, how, the more rigorous account may be sensibly extended and generalized. If an extension proves impossible in one or the other dimension, then we should ask what alternative ways there are to make a case for the results of simulations and what this means for our understanding of validation (cf. Chaps. 9 and 17 by Saam in this volume).

When it comes to standards and requirements, a sensible point to make is that their strictness should depend on the intended uses of a simulation. This is true, but only pushes things one step further: What we need is a classification of the possible uses to which computer simulations may be put and a mapping from the various classes to standards for validation. The classification may go as far as to suggest what sorts of validation metric may be chosen for each class. This would help to systematize, and to develop existing validation metrics further. What is interesting concerning the plurality of uses too is that there are several "qualified" notions of validation that look at specific aspects of a simulation, e.g., structural validation (see Chap. 31 by Fagiolo). The proposed classification may help to clarify these notions.

Apart from these “large-scale” issues about validation, there are of course some more specific problems that call for more research. We name but two of them.

First, the use of statistics in the comparison between simulation output and data as well as the representation of uncertainties need further scrutiny. It is telling in this respect that both the chapters about frequentist statistics (Chap. 19 by Robinson) and Bayesian epistemology (Chap. 7 by Beisbart) indicate some unease about the standard methods used in both camps. There are further strong arguments to the effect that so-called epistemic uncertainties can often not be dealt with appropriately using probabilities. Instead, something like imprecise probabilities is needed (see Chap. 21 by Bradley in this volume). Conceptual as well as mathematical work is required to elaborate imprecise probabilities and to show how they can be used in decisions.

Second, it is often a problem that researchers are interested in simulation predictions about system S and aspects A, but that only data about a slightly different system S’ and/or different aspects A’ are available (see Chap. 34 by Knutti et al.). The question then is how they can evaluate their simulations for S and A by using data concerning S’ and A’. Technically, the question is how a validation metric can be chosen that works on S’ and A’ but measures how well S is represented in respects A. The answer depends of course on how S and S’, and A and A’, are related to each other, but the general question is how these relationships play out in the choice in the validation metric.

There are also philosophical questions that need more scrutiny. For instance, talk about simulations and their validation often sounds thoroughly pragmatist: Uses and usability seem to be the most important desiderata. Despite that, we do not find pragmatist or instrumentalist accounts of simulation validation. More research is needed to see whether simulation validation makes a case for a pragmatist or instrumentalist approach to the philosophy of science. We should note though that there are some deep philosophical disagreements about how science works (e.g., between those who want to shun induction and those who do not take this to be required). It is unlikely that parties that disagree in this profound way will agree on validation. Conversely, validation will not make a substantive difference to their debate.

In any case, there are fascinating research questions about validation that require cross- and interdisciplinary efforts. We hope that this volume helps to put them on the agenda for future research.

References

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to ‘evaludation’: A review of terminology and a practical approach. *Ecological Modelling, 280,* 117–128.

Godfrey-Smith, P. (2003). *Theory and reality. An introduction to the philosophy of science*. Chicago: University of Chicago Press.

Feinstein, A. H., & Cannon, H. M. (2003). A hermeneutical approach to external validation of simulation models. *Simulation & Gaming, 34,* 186–197.

Frigg, R., Bradley, S., Du, H., & Smith, L. A. (2014). Laplace's demon and the adventures of his apprentices. *Philosophy of Science, 81*(1), 31–59.

Frigg, R., & Reiss, J. (2009). The philosophy of simulation: Hot new issues or same old stew? *Synthese, 169*(3), 593–613.

Ghetiu, T., Polack, F. A., & Bown, J. (2010). Argument-driven validation of computer simulations–A necessity rather than an option. In *VALID 2010. The Second International Conference on Advances in System Testing and Validation Lifecycle*, August 22–27 (pp. 1–4). Nice, France, IEEE Press.

Goldman, Alvin I. (1999). *Knowledge in a social world*. Oxford: Clarendon Press.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., et al. (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science, 310,* 987–991.

Harding, A., Keegan, M., & Kelly, S. (2010). Validating a dynamic population microsimulation model: Recent experience in Australia. *International Journal of Microsimulation, 3*(2), 46–64.

Hartmann, S. (1996). The world as a process: Simulations in the natural and social sciences. In: R. Hegselmann, U. Müller, & K. G. Troitzsch, (Eds.), *Modelling and simulation in the social sciences from the philosophy of science point of view, theory and decision library* (pp. 77–100). Dordrecht: Kluwer.

Herskovitz, P. J. (1991). A theoretical framework for simulation validation: Popper's falsificationism. *International Journal of Modelling and Simulation, 11,* 56–58.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. New York: Oxford University Press.

IPCC. (2014). *Climate change 2013–The physical science basis working group I contribution to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge: Cambridge University Press.

Kaldor, N. (1968). Capital accumulation and economic growth. In F. A. Lutz & D. C. Hague (Eds.), *The theory of capital* (Reprint ed., pp. 177–222). London: Macmillan.

Kertész, A. (1993). *Artificial intelligence and the sociology of scientific knowledge*. Frankfurt/M.: Lang.

Klein, E. E., & Herskovitz, P. J. (2005). Philosophical foundations of computer simulation validation. *Simulation & Gaming, 36,* 303–329.

Kleindorfer, G. B., O'Neill, L., & Ganeshan, R. (1998). Validation in simulation: Various positions in the philosophy of science. *Management Science, 44,* 1087–1099.

Murray-Smith, D. J. (2015). *Testing and validation of computer simulation models: Principles methods and applications*. Cham: Springer.

Nickles, T. (1989). Integrating the science studies disciplines. In S. Fuller, M. de Mey, T. Shinn, & S. Woolgar (Eds.), *The cognitive turn. Sociological and psychological perspectives on science* (pp. 225–256). Dordrecht: Kluwer.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge: Cambridge University Press.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science, 263,* 641–646.

Parker, W. S. (2008). Franklin, holmes, and the epistemology of computer simulation. *International Studies in the Philosophy of Science, 22*(2), 165–183.

Parker, W. S. (2009). Confirmation and adequacy-for-purpose in climate modeling. *Aristotelian Society Supplementary, 83,* 233–249.

Roache, P. J. (2009). *Fundamentals of verification and validation*. New Mexico: Hermosa Publishers.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science, 20,* 1–6.

Schlesinger, S., et al. (1979). Terminology for model credibility. *Simulation, 32,* 103–104.

Walker, D. C., Hill, G., & Wood, S. M. et al. (2 more authors). (2004). Agent-based computational modeling of wounded epithelial cell monolayers. *IEEE Transactions on Nanobioscience*, *3*(3), 153–163.

Winsberg, E. (2001). Simulations, models, and theories. Complex physical systems and their representations. *Philosophy of Science, 68,* S442–S454.

Winsberg, E. (2010). *Science in the age of computer simulation*. Chicago: University of Chicago Press.

Part I
Foundations—Basic Conceptions in Simulation Model Validation

Chapter 2
What is Validation of Computer Simulations? Toward a Clarification of the Concept of Validation and of Related Notions

Claus Beisbart

Abstract This chapter clarifies the concept of validation of computer simulations by comparing various definitions that have been proposed for the notion. While the definitions agree in taking validation to be an evaluation, they differ on the following questions: (1) What exactly is evaluated—results from a computer simulation, a model, a computer code? (2) What are the standards of evaluation—truth, accuracy, and credibility or also something else? (3) What type of verdict does validation lead to—that the simulation is such and such good, or that it passes a test defined by a certain threshold? (4) How strong needs the case to be for the verdict? (5) Does validation necessarily proceed by comparing simulation outputs with measured data? Along with these questions, the chapter explains notions that figure prominently in them, e.g., the concepts of accuracy and credibility. It further discusses natural answers to the questions as well as arguments that speak in favor and against these answers. The aim is to obtain a better understanding of the options we have for defining validation and how they are related to each other.

Keywords Evaluation · Model · Code · Truth · Accuracy · Credibility · Adequate representation · Adequacy for purpose · Explanation · Data-driven validation · Test

C. Beisbart (✉)
University of Bern, Bern, Switzerland
e-mail: Claus.Beisbart@philo.unibe.ch

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_2

2.1 Introduction

This volume is built on the premise that the validation of computer simulations needs more attention. Here, very roughly, validation comprises the efforts to show that a computer simulation represents its target appropriately. Validation needs more attention *in practice*, because we can only rely upon results from computer simulations both for scientific purposes and in applications, e.g., in engineering or policy advice, if the results have been shown to be genuine. As a matter of fact, however, validation is often done sloppily and superficially, if it is done at all (see Chap. 8 by Arnold for examples). Validation needs more attention *at a theoretical level* because researchers complain that validation is not well-understood (see, e.g., Ghetiu et al. 2010, p. 1 for testimony).

A first step forwards to better understand validation is to address the question of what validation is. Ideally, we answer this question with a full-fledged definition of validation. Since "validation" is a general term that denotes a concept, it is natural to think that the definition explains, and thus clarifies, the *concept* (or the notion, as we will also say) of validation: The definition unpacks what is thought under the label of "validation", and thus what the term "validation" means, or should mean.

Of course, even if we have explained the concept of validation in this way, this will not answer all questions we may have about validation, e.g., how it should be done in practice. But a clarification of the concept at least helps to avoid misunderstandings about validation. Moreover, answers to more substantive questions about validation depend on what we mean by "validation", so a clarification of the concept is the first necessary step to address other questions.

The aim of this chapter is to make progress in clarifying the concept of validation of computer simulations. This chapter is philosophical in nature because the clarification of concepts is a genuinely philosophical task, at least if the concepts are as fundamental as is the one of validation. To explain this point in very simple terms, we may say that concepts are tools that we need in our thinking, and that philosophers try to improve our thinking by explaining fundamental concepts. Such a project, it seems, can be undertaken without much recourse to empirical data and is thus not part of the empirical sciences. The so-called conceptual engineering, i.e., the critical appraisal and development of our concepts, does in fact figure prominently in present-day philosophy (see, e.g., Blackburn 1999, p. 2 for a programmatic statement and Cappelen 2018 for a recent account of this endeavor).

As it turns out, clarifying the notion of validation is a challenge, since we find various proposals for a definition in the literature. While some differences between the definitions are merely verbal, there are also substantial disagreements (see below for evidence), and so the question arises: Which definition should we adopt? In this chapter, we will not argue for one specific definition or concept of validation. Our main aim is more modest: We will compare and discuss prominent proposals to define validation and classify the differences between them. To this purpose, it will be useful to clarify notions that are prominently mentioned in definitions of validation, e.g., that of accuracy. We hope that our comparative investigation yields at least two benefits: First, we obtain a systematic overview and a closer understanding of the options on the table. Second, we can better appreciate what substantial controversies about validation there are. The reason is that different understandings of validation have different consequences. At the same time, they reflect different stances on scientific inquiry more generally, and it is worthwhile to see how these stances manifest themselves in our thinking about validation.

The most comprehensive and interesting discussions of the notion of validation so far are due to Roache (2009)[1] and to Oberkampf and Roy (2010, Chap. 2). An often quoted article about the notion is provided by Oreskes et al. (1994). We will engage with this literature in what follows. Note though that our perspective is broader than the one taken by Roache, Oberkampf and Roy: While these authors focus on engineering and the physical sciences, we also want to take serious other disciplines. Also, the way in which we proceed is more modeled after the methods of a philosophical inquiry.

This chapter is organized as follows: We start with some preliminaries in Sect. 2.2 For instance, we detail how we think about concepts and clarify our focus on the validation *of computer simulations*. We list some definitions of validation in Sect. 2.3. The definitions are compared and discussed in detail in Sect. 2.4, which contains the main work of this chapter. We draw conclusions in Sect. 2.5 by summarizing our clarifications using a scheme for a definition of validation.

2.2 Preliminaries

In this chapter, we are interested in the *concept* of validation. For our purposes, we need not elaborate on what concepts are (see e.g., Margolis and Laurence 2014 for an overview of philosophical work on concepts). Suffice it to say that concepts are general and can apply to several particular things, which, in turn, instantiate the concept. We should not confuse concepts with words. The latter can stand for, or denote, concepts. Note though that ambiguous words can stand for several different concepts, depending on the context.

In the philosophical literature, various desiderata on the clarification of concepts have been proposed (see e.g., Carnap 1950/1962, Chap. 1 for a famous classic in this respect). One important standard may be called descriptive accuracy: If a definition of a concept is proposed, it should capture the way in which the concept is in fact understood and used in a community of speakers, or at least approximate it as closely as is possible. In the terms used by Carnap, the concept specified in the definition should be as similar as possible to the one used in the community. How a community thinks of a concept can be found out by observing the way in which competent people use the term denoting the concept. So we have to make sure that a clarification of the notion of validation does not move too far from the way in which the term "validation" is used in relevant circles (although nothing hinges on the word as such).

[1] This article is reprinted with an addendum as Chap. 3 of Roache (2013), see also Roache (1998).

The quest for descriptive accuracy runs into trouble though, if a notion is not always understood in the same way. This condition is likely met for the use of "validation" among scientists. This at least is indicated by the fact that different definitions have been proposed. To choose between the various definitions, we need to appeal to different desiderata, and recourse to such desiderata will have the consequence that the proposed clarification is to some part stipulative, because it regiments existing uses of the concept. The most important additional desiderata considered in the philosophical literature comprise exactness, fruitfulness, and simplicity (see Carnap 1950/1962, pp. 5–8; see Brun 2016 for analysis and elaboration). One important aspect of *exactnes*s is that vagueness be removed in the following sense: There should be as few cases as possible in which it is not clear whether the concept applies or not (Brun 2016, p. 1222). A concept is *simpler* if it is more easily defined (Carnap 1950/1962, p. 7); finally, we can call a concept more *fruitful* if it facilitates theory building, e.g., by allowing for more generalizations (ibid., p. 7). Whether a certain clarification of a concept is fruitful or not, often depends on background knowledge. Thus, in what follows, our argument will often be based upon views about what can sensibly be achieved during validation. Clearly, a concept of validation would not be fruitful if it could never be instantiated.[2]

After these general reflections about the clarification of concepts, we can turn to validation more specifically. The word "validation" is derived from the term "to validate" (which, in turn, derives from "valid"). The term is slightly ambiguous in the following sense: "validation" may either refer to a type of activity or to its results. For instance, when we say that validation has been achieved, we mean that a certain result has been reached. This result may also be described by saying that validity (in some sense) was attained. There is a close connection between this validity and the practice of validation, and we might define validity to be the result of successful validation qua activity, or the other way round. But these ways of defining validity, or validation, are not illuminating because they move us in a circle that does not really explain what the whole business of validation is about. So we have to break the circle and to define at least one of the concepts in terms of different concepts, and this will also clarify the other concept. In this chapter, our efforts into clarification will be targeted at the *activity* of validation. We thus assume that the term "validation" denotes a type of practice (which is, of course, also denoted by "to validate"). The concept, or the type is instantiated by several concrete activity tokens that are grouped together.[3]—One reason to focus on the practice of validation is simply that this book is about the steps that need to be taken during validation qua practice. Also, it is not so clear what "validity" really means in this context.[4]

[2] We have roughly, but not in every detail, followed Carnap in specifying the desiderata.

[3] This is not the place to clarify the relationship between concepts and types. For our purposes, it suffices to note that both are general because they can be instantiated by several particulars or tokens.

[4] In logic, there is a clear-cut definition of validity: An argument is valid if, and only if, it is impossible that the premises are true and the conclusions false. It is not clear though how this notion of validity should be applied to simulations. True, a simulation or a model involves an inference from the model to the target system, and this inference is supposed to be a reasonable argument. But there

Validation has always an *object* that is validated. In the context of this book, it is worth noting that experiments, models and computer simulations are often considered as objects of validation. Certain types of software and programs or codes are also often said to be validated. What is most important in the context of this book is, of course, the validation of computer simulations, and we only want to clarify what this amounts to. But as we will see, there is a tight connection with models, so we need to consider the validation of models too. Likewise, there is a close relationship to computer codes and software, so will also briefly consider their validation (and what is called their "verification", cf. Chap. 11 by Rider and Chap. 27 by Roache in this volume). Note too that, if computer simulations turn out to be experiments, then the validation of the former can be expected to be more or less the validation of experiments, but we will not pursue this line of thought here any further (see, however, Chap. 37 by Beisbart and Chap. 38 by Barberousse and Jebeile in this volume).

By "computer simulation" (or just "simulation", for short) we mean a method that crucially involves the execution of a computer code, which traces at least approximate solutions to equations that are essential parts of a model of the dynamical evolution of some real or imagined target system under real or counterfactual conditions. In this way, the dynamical behavior within the target system is imitated using a digital computer. This definition is supposed to cover simulations based upon (ordinary or partial) differential equations and agent-based simulations as well as cellular automata. Monte Carlo simulations that trace the stochastic dynamics of a system are included too. What is further decisive according to our working definition of simulations is that the equations to be solved have an interpretation in terms of a target system. As detailed in our working definition, this is often a real system, but there are computer simulations for which a real-world target system is missing; e.g., when simulations consider a world of point particles in which gravity behaves differently from gravity as we know it. The equations that are (approximately) solved in the simulation are supposed to provide a model of the target system; either in the sense that they represent a real-world system or in the sense that the equations define the dynamics of a merely imagined system. Finally, the definition excludes computations that (approximately) solve equations from a model that does not describe a dynamical evolution (for instance, a model may only represent the spatial distribution of some objects at one instance of time). At least the philosophical literature about computer simulation has mainly reached a consensus that simulations imitate a time evolution of a system (see e.g., Hartmann 1996, Sect. 2.2; Humphreys's definition in his 2004, pp. 110–111 is an exception). But the restriction to representations of a time evolution does in fact not make any substantial difference to validation. Thus, in what follows, the restriction is not necessary to our argument, even though our jargon ("tracing evolution") and our examples center on simulations of some dynamical behavior.[5]

is most often no chance that the argument is logically valid (because it is, e.g., analogical). Other uses of the term "valid", e.g., in "valid rules" are even more remote from simulations.

[5] There are simulations that do not use a digital computer, e.g., analog simulations. Our focus is entirely on simulations done with a digital computer.

Our working definition of computer simulation is meant to make clear what method we refer to when talking of simulations, but it is not supposed to answer deeper philosophical questions as to what computer simulations really are (see Chap. 37 by Beisbart in this volume). Thus, later in this chapter, further discussion will be needed to get clear about what object is validated.

With these preliminary clarifications in mind, we can now consider prominent definitions of the validation of simulations.

2.3 Influential Definitions of Validating Computer Simulations

In the 1970s, the Society for Modeling and Simulation International (SCS) instituted the SCS Technical Committee on Model Credibility, headed by S. Schlesinger. This committee proposed the following definition of model validation (Schlesinger et al. 1979, p. 104):

> Val-SCS "Substantiation that a COMPUTERIZED MODEL within its DOMAIN OF APPLICABILITY possesses a satisfactory RANGE OF ACCURACY consistent with the intended application of the model." (uppercase letters as in the original; they indicate terms for which definitions are given)

Here the range of accuracy is the

> *demonstrated* agreement between the COMPUTERIZED MODEL and REALITY within a stipulated DOMAIN OF APPLICABILITY. (ibid., emphasis theirs)

The American Institute of Aeronautics and Astronautics (AIAA) has recommended the following definition of validation, which was later adopted by the American Society of Mechanical Engineers (AIAA 1998; ASME 2006; see Oberkampf and Roy 2010, Sect. 2.1.4, pp. 26–31 for the history):

> Val-AIAA "The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model" (here quoted after Oberkampf and Trucano 2008, p. 719)

This definition is also adopted by Oberkampf and Roy (2010, p. 32). Roache (2009/2013, p. 79) starts with this definition too and elaborates it as follows:

> Val-Roache "Validation: The process of determining the degree to which a model with its associated data is an accurate representation of the real world as determined by experimental data, the metrics of which are chosen from the perspective of the intended uses of the model."

Since the authors who contribute to this book come from various fields, we have also collected a survey of the definitions they propose for validation. Of course, many of the definitions (e.g., the ones adopted by Gelfert, Chap. 10 and Frisch, Chap. 41) draw on the ones already mentioned. But here are some examples that do not do so in an obvious way: Murray-Smith (Chap. 4 in this volume, abstract) defines validation as follows:

> Val-MS "The word 'validation' is used to describe procedures for establishing whether the model fidelity is adequate for the purposes of the given application."

Saam (Chap. 18, Sect. 1) quotes the following definition due to Caldwell and Morrison (2000, pp. 202 f.):

> Val-CM "*Validation is a proactive, diagnostic effort to ensure that the model's results are reasonable and credible*" and "*to assess whether the model's outputs are reasonable for their intended purposes.*"

Social scientist Mäs (Chap. 35, Sect. 1) defines validation as follows:

> Val-Mäs: "In the present chapter, the term 'validation' describes the process of confronting a theory with empirical evidence with the ultimate aim of developing a sound explanation of the empirical phenomenon."

It is finally worth quoting a definition that is concerned with validation of software more generally (not just related to computer simulations): The Institute of Electrical and Electronics Engineers (IEEE) has set standards for validation and verification of software and proposed the following definition of validation of software (IEEE 2012, p. 11):

> Val-IEEE "(**A**) The process of evaluating a system or component during or at the end of the development process to determine whether it satisfies specified requirements. (**B**) The process of providing evidence that the system, software, or hardware and its associated products satisfy requirements allocated to it at the end of each life cycle activity, solve the right problem (e.g., correctly model physical laws, implement business rules, and use the proper system assumptions), and satisfy intended use and user needs."

2.4 Discussion of the Definitions

The definitions just quoted and most other definitions, as, e.g., used by other authors in this book, have a lot in common, but differ in a couple of crucial dimensions. In the next subsection, we will stress important common traits. We will then turn to a systematic discussion of the dimensions on which the definitions differ.

2.4.1 Commonalities

The definitions that we know of concur in that validation is, or at least involves, an evaluation or assessment. Sometimes, this is very explicit in the wording of the definition (see e.g., Chap. 31 by Fagiolo et al., Sect. 2.1; Chap. 40 by Hirsch Hadorn and Baumberger, Sect. 2 in this volume), sometimes it is more implicit. For instance, Val-SCS speaks of a substantiation that a computerized model is sufficiently accurate. Since accuracy is value-laden (more accuracy being better) and accuracy is supposed to be substantiated, it is clear that validation is taken to imply an evaluation. True, there are definitions of validation that only speak of a comparison (e.g., Rood in his Chap. 30, Sect. 2 in this volume), but we can safely assume that the comparison is not merely descriptive, but crucially involves evaluation too.

A more interesting issue is whether validation goes beyond evaluation or assessment. Some authors suggest that this is so, e.g., Lenhard in his Chap. 39 in this volume assumes that validation includes the improvement of a model as a reaction to an evaluation. The term "substantiation" used by Schlesinger et al. (1979) is not entirely clear on this issue and may be understood as encompassing efforts to improve a computer simulation too. But given the fact that most definitions of validation do not include such efforts, we will not consider them any further in this chapter. This is not to deny that, in practice, evaluation and model development are often much intertwined. But if a simulation code is first evaluated and then improved as to better match measured data, this is not just validation of the former version of the code; rather, we may want to call the whole process calibration (see e.g., Chap. 3 by Oberkampf for this notion).

A lot of definitions, e.g., Val-AIAA, let the assessment implicit in validation refer to some real-world system. It is likely that this is also meant by other definitions that do not explicitly mention the real world. In what follows we will thus assume that a real-world system is decisive for the assessment implicit in validation. Now, this makes only sense if the target system of a simulation is a real-world system (which may be considered under counterfactual conditions though). So we will assume that the simulations to be validated do have such a real-world target. This excludes other simulations, which have a merely imagined system as their target (which is allowed by our working definition of simulation). This exclusion is not a problem. It is, of course, true that such simulations should be assessed in terms of their target too. But a related assessment can only make sure that the simulation properly reflects the imagined target. Since the latter is defined in terms of modeling assumptions about the imagined scenario, the assessment has to make sure that the modeling assumptions are properly reflected in the imagined system. Now, this is a question that is answered during so-called verification of a simulation. Very roughly, the verification of a simulation has to make sure that the results of the computer simulation approximate solutions to the model to a sufficient accuracy (see Chap. 11 by Rider and Chap. 13 by Roache in this volume). In this sense, one may say that simulations with a merely imagined target are subject to some sort of validation, but that the latter reduces to what is known as verification in this particular case.

2.4.2 Difference 1: The Object of the Validation

If validation is crucially an evaluation, the next natural question to ask is: What exactly is evaluated? This question is obviously the same as to what the proper object of validation is—a question that has already emerged above in our preliminaries. We are of course interested in the validation of a computer simulation, but this does not really specify what the proper object is because a simulation may be thought to be several things. In fact, the definitions that we have collected differ in this regard. Many of them put the (computer/simulation) *model* at the center of validation, very often with some restriction, e.g., that the focus be on the intended applications of the model. There are exceptions though; Val-CM assumes that the *results* of a simulation are evaluated, while Rood, in his Chap. 30, mentions the *code* as object of evaluation. This is also implicit in Val-SCS, since Schlesinger et al. (1979) assume that the computerized model is a code. Some authors even show sympathy to the idea that the practice of running computer simulations is assessed, when it comes to validation (see Chap. 40 by Hirsch Hadorn and Baumberger in this volume).

In what follows, we will not assume that, in validation, the practice of running simulations is a primary object of evaluation. One reason is that only a few authors take this to be the case. A second reason is that the practice of doing computer simulations is clearly aimed at producing an adequate simulation model or code and related results. So the practice is first and foremost to be evaluated using the models, codes, and results that are produced, which brings us back to the other candidates for proper objects of validation. What needs closer scrutiny then to become clearer about the proper object of validation is the question of how the other candidates, i.e., simulations, their results, the models, and the code, relate to each other.

We take it that the *results* of a simulation arise from the outputs that were produced when the simulation code was run once or several times (so the results are meant to be *actual* results). Very often, what is called a result from a simulation consists in claims about the target system, e.g., that the mean global surface temperature raises only moderately for a certain emission scenario considered in a climate simulation (see Beisbart 2012, 2017 for a detailed account of how such results arise from the output). In philosophical parlance, such results are propositional; they can be expressed in terms of assertive sentences. We assume that they are about the target system. Results may be formulated about a model too, but then there is no need for validation in the sense assumed here. There are arguably also results that are no claims, most importantly viewgraphs and animations. But these kinds of results are only of interest to the extent to which they represent the target system. For the purposes of this chapter, we can assume that their representational content can be expressed in claims too, so it is propositional too.[6]

Results in this sense are obviously a subset of the *possible/potential* results one may obtain from the computer simulation in a broader sense (e.g., its code or the underlying model). Thus, if the simulation in this latter sense is evaluated, all results it can yield for a certain range of applications are up to assessment. We can thus conclude that the assessment of a whole simulation includes the assessment of its *actual* results, but significantly goes beyond this. Accordingly, assessing a model/code is more interesting, but also more demanding, than assessing a finite series of results. While actual results of computer simulations can often be fully stated in a small

[6] This is not meant to answer the general question of whether we can fully capture the content of paintings, etc., in terms of language. The answer is probably no; but our focus is here on viewgraphs, etc., that are used for scientific purposes, and it seems more likely that their content is fully reflected in claims expressible in language.

number of claims that can each be checked empirically, this is not possible for all the potential results a simulation may yield. The question of whether the assessment of a computer model or code is exhausted by the evaluation of its results will be tackled below.

Turn now to *computer codes*. Clearly, every computer simulation involves a code, and, very often, computer simulations are identified and individuated using the code. For our purposes, it is most appropriate to think of the code as an interconnected set of instructions written in the machine language such that a suitable computer hardware can carry out the instructions. For the most part of our discussion, this code need not be distinguished from programs written in higher level programming languages such as C or from an algorithm.

Consider finally *models*. It is too clear that there is a close connection between computer simulations and modeling, for "to simulate" means roughly "to model", and according to our working definition, a computer simulation yields at least approximate solutions to equations from a model. But things are a bit more complicated (see e.g., Beisbart 2014 for closer analysis). As various authors, e.g., Winsberg (1999), have stressed, there is, very often, not just one model involved in a computer simulation. The reason is that models are typically changed and further developed when they are implemented on a computer. This is particularly clear for simulations that are based on differential equations, e.g., models of fluids that employ the Navier–Stokes equations. Here, typically, the differential equations have been proposed as a model independently of computer simulations. To implement them on the computer, scientists have to discretize them, which is to say that the differential equations are approximated using, e.g., difference equations. It then is natural to say that the latter equations form the core of a second model that is distinct from the initial one.[7] More generally, it is often useful to distinguish between the *conceptual* and the *computational* model (Oberkampf and Roy 2010, p. 38; cf. Schlesinger et al. 1979, p. 103). While the conceptual model is made of assumptions that reflect the scientific understanding of the target system independently of, and prior to, a simulation (at least up to some simplifications), the computational model is implicit in the computer code in the sense that the latter provides exact and correct solutions to it. In this way, the dynamics of the computational model is uniquely defined in terms of the code. By contrast, there is some leeway as to what exactly the conceptual model is; it is a model that is somehow implemented in the computer code, but not in every detail or with 100% accuracy. How we identify the conceptual model in a concrete example of a computer simulation is a matter of interpretation.[8]

[7] In the example of the Navier–Stokes equations, the models do not differ in their ontology, i.e., in the range of things they assume, but only in their equations that govern the dynamics. But there are in fact approximations that involve a change in the ontology.

[8] A general problem about identifying models is that ordinary talk is often quite imprecise on what exactly a model is. For instance, sometimes, a set of differential equations with a free parameter is called a model, while, other times, the parameter is assumed to have a specific value to yield one model. So what we call models differ in the degree in which they are specific. In this way, some of the types of models that Winsberg (1999) uses to understand computer simulations differ because they add assumptions to the assumptions of other models. In this chapter, do *not* assume that the

Since the *computational model* and the *simulation code* are closely connected, so are their respective validations. We can in fact legitimately claim that the validation of a simulation code is no more than the validation of the computational model. This is not to deny that a code is a different sort of thing than a model. A code is a set of instructions, while a mathematical model is defined in terms of assumptions that are independent of language. But in the context of validation, the computer code is only of interest qua implementing a model, and we need not differentiate between the validation of the simulation code and of the computational model.[9]

By contrast, the validation of the *conceptual model* differs from that of the computer code simply because the conceptual model is in general only implemented to some approximation in the code. So we can only think of a simulation in terms of a conceptual model if the results of both the conceptual and the computational models are close enough for the purposes of some inquiry, such that the outputs of the simulations can, in fact, be used to constrain the solutions to the conceptual model, although the outputs specify strictly speaking solutions to the computational model only. As far as validation is concerned, if a researcher is interested in evaluating the results of a conceptual model, she can only use the computer output if the latter faithfully reflects the behavior of the conceptual model. Activities that show that this condition is fulfilled are called verification (see Chap. 3 by Oberkampf, Chap. 4 by Murray-Smith, Chap. 11 by Rider and Chap. 27 by Roache in this volume).

Since verification can be quite difficult, it may be suggested that researchers should concentrate on the computational model if they engage in validation. But there is a problem with this suggestion. As defined above, the computational model is implemented in the computer code such that running it delivers correct solutions to it. This model is first difficult to grasp because the computer code deviates from the equations at the core of the conceptual model. For instance, the execution of the code leads to roundoff errors that are not naturally described in terms of simple equations about the model. Second, the working scientist may be mistaken about the computational model, for instance, if the code contains an unknown "bug" going back to a mistake in the programming. Thus, verification is needed even if researchers do not consciously change the model with which they start, when they implement it on the computer (this is so for many agent-based models and cellular automata for which no approximations are needed). The focus of verification is then not on approximations but rather on a test whether the code really solves the equations constitutive of a certain conceptual model that scientists intend to use. We leave the

conceptual and the computational model differ in their degree of specificity. Rather, both types of models are assumed to have roughly the same degree of specificity, but to differ in their model assumptions.

[9] Since a computer code forms the software, we can briefly consider Val-IEEE at this point, which was given for software validation more generally. The definition stresses the fulfillment of certain requirements (part A), or the satisfaction of intended use and user needs (part B). This covers the validation of computer simulation codes if, during the validation of simulations, the requirements mentioned in Val-IEEE are just the requirements that are set on codes in the validation of *simulation* codes, and if we make an analogous assumption about user needs. In this sense, the validation of simulation codes is a special case of validation of codes.

discussion of whether verification is part of, or rather a precondition of, validation to Chap. 42 by Beisbart in this volume (see also Chap. 27 by Roache in this volume).

Altogether, our discussion leaves us with two distinctions that can be used to classify the possible objects of validation: First, we need to distinguish between the results of a simulation and a simulation itself. While the former comprise claims that were obtained from a small number of runs of the simulation code and thus focus on a couple of trajectories within the target system, the simulation as such is more comprehensive, e.g., because it traces model behavior in the sense of many possible trajectories.[10] What exactly the computer simulation is, e.g., a program in a high-level language, a code or a model does not matter for our purposes, because the computer simulation matters for validation only to the extent it implements a model. What needs to be distinguished though—and this is the second distinction—are conceptual and computational models and their respective validations.

2.4.3 Difference 2: The Standard of Evaluation

If validation is an evaluation, another natural question to ask is what the standard of evaluation is. The definitions that we have quoted appeal to the standards of accuracy (Val-SCS, Val-AIAA, Val-Roache), adequacy for purpose (Val-MS), reasonableness and credibility (Val-CM). Val-Mäs mentions the ideal of sound explanations. What we find in the literature too sometimes is an appeal to truth. For instance, Naylor and Finger (1967, p. B93) define validation of a model thus:

> To […] validate any kind of model (e.g., management science models) means to prove the model to be true.

But they add immediately that truth can most often not be proven.

While some of the standards mentioned here and elsewhere, e.g., accuracy, are quite *specific*, others, e.g., reasonableness, are very *unspecific*. Many types of things can in some sense be called reasonable. For the purposes of this chapter, it is most fruitful to focus on more specific standards because they are more informative. Standards such as reasonableness, by contrast, would need interpretation in terms of more specific standards (see Lacey 1999, Chap. 2 for standards that are appealed to in theory choice).

When it comes to more specific standards, it is notable that some of them primarily apply to model results, while others do not. For instance, truth and credibility are characteristics that are first and foremost instantiated by claims or specific results. A code (qua series of instructions) or a model (e.g., qua simplified system) cannot be true or credible, properly speaking. It may only be true or credible in the derivative sense, e.g., because it produces true or credible results. By contrast, the standard of

[10] In Monte Carlo simulations, one run of the code typically yields an ensemble of trajectories in the target system, but the distinction between a few actual results and model behavior in a more comprehensive sense applies here too.

reliability (mentioned in Chap. 38, Sect. 2.1 by Barberousse and Jebeile) is primarily said to hold of processes and procedures that are meant to produce knowledge, and thus applies to simulations qua models or codes.

Let us thus first consider standards that are primarily relevant for the *results* of computer simulations, viz. truth, credibility and accuracy.

2.4.3.1 Truth, Credibility and Accuracy

It is clear that practicing scientists are interested in the *truth* of their results. The result that p (e.g., that there is an increase in precipitation in some region at some time) is true if, and only if, p is the case (i.e., if there is an increase in precipitation in this region at this time). This is a platitude, but it has proven surprisingly illuminating in philosophical inquiry on what truth is (see e.g., Künne 2003). There is no need to dig deeper into philosophical theories about truth, because almost nothing hinges on that for the purposes of this chapter. The reason is that the truth of a claim can only be shown with an appeal to other standards (see below), so it will only be important as a background ideal.

Credibility, as it is understood here and in much philosophy, is related to truth as follows. That a claim is credible or trustworthy (to some degree) is meant to say that it deserves belief (to that degree; degrees of beliefs form a basic notion within Bayesian epistemology, see Chap. 7 by Beisbart in this volume). In this vein, Oberkampf and Roy (2010, p. 12) state that credible "results of analysis […] are worthy of belief or confidence". Belief is an attitude that people take toward claims, and this attitude is aptly characterized as taking something true. So, the *credibility of a result is the worthiness to take it true*. How worthy of belief a result depends on how well it is justified.[11] That is, it depends on how much it is supported by the evidence that speaks in favor of it. For instance, the claims of string theory or another fancy physical theory of quantum gravity are not particularly credible at present since we lack evidence for them.[12]

To many people, the term "credibility" sounds very subjective. And indeed, if credibility were just the degree to which people find something plausible, then it would be subjective in the sense that it differed between people. But credibility is not apparent plausibility or something like this, but rather worthiness of belief, and it is at least arguable that the degree to which a claim is credible in a certain context (given specific evidence) can be determined in an objective way. The idea is that it is a function of the available evidence: The stronger the latter speaks in favor of a claim, the more credible is the claim. In ordinary talk, we often assume that we agree on how strong the evidence is for a claim and how worthy of belief it is. There are some

[11] Justification is a key term in epistemology (the philosophical study of knowledge), one important idea being that knowledge is justified, true belief.

[12] To support the claim that results of a simulation *about a real-world target in nature* hold true, the evidence needs to be empirical. Thus, in this chapter, evidence is meant to be empirical evidence. This does not preclude that validation is to some part built on mathematical proof. But such proof cannot replace the recourse to observation.

philosophical reasons to doubt that (see e.g., Chap. 7 by Beisbart in this volume), but we cannot settle this issue here. So let it be noted that credibility is meant to be as objective as it can be. Note also that, in this chapter, credibility is not meant to suggest a fairly low degree of credibility, as it is often in the ordinary talk (compare “this is certain” vs. “this is credible”).

As we have just noted, at present, string theory is not particularly credible. But it may nevertheless be true. Conversely, a claim may be supported by considerable evidence, but still be false. So truth and credibility can come apart. If so, which aim should we give priority? It is clear that our fundamental interest is in truth. But whether or not some results are true cannot be read off from them, as it were, and we need evidence to assess whether they are true. Weighing the various pieces of evidence that speak for and against the truth of some claim determines how credible the claim is. Once we have determined the credibility of a claim, what should we say about its truth? Well, roughly, we should take the claim true if, and only if, it is sufficiently credible. Consequently, from the first personal perspective of truth-seeking people, truth and credibility cannot pull in different directions.[13]

The notion of *accuracy* is best explained using the example of measurement as follows (cf. Humphreys 2004, p. 16): Consider a characteristic, e.g., the mass of a certain body or the pressure of air at a certain location. Suppose further that a measurement device yields a value for this characteristic, say, a mass of 1,334 kg. Assume further that, as a matter of fact, the mass of the body takes a certain value, say, 1,335 kg. Call this value its true value of mass. The value that is output by the measurement or the simulation is the more accurate the closer it is to the true value, and the degree of accuracy may itself be measured by taking a distance measure between the measured and the true value.

This account of accuracy can be generalized to computer simulations: An output value of a characteristic, for e.g., mass, is the more accurate the closer it is to the true value. The resulting account of accuracy is consistent with the definition quoted from Schlesinger above, if “range” means degree and if the term “disagreement” refers to the level with which the value obtained in the simulation differs from the true one.[14]

This understanding of accuracy presupposes that two conditions are met: first that there is a true value, and second that deviations from the true value can at least be ordered, if not quantified in terms of numbers. Both conditions are fulfilled in our example with mass, as they are regarding many simulation results, e.g., about the mean global surface temperature. However, some computer simulations contain parameters for which the assumption of a true parameter value is taken to be false,

[13] A conflict between truth and credibility may arise if we consider the evidence which *other* people have to take something true. Their evidence may favor a claim that we think to be false.

[14] Reference to a true value is innocuous here. The true value of a characteristic is just the value that the characteristic takes as a matter of fact (if it takes one, see below for this assumption). Note too that true values are not restricted to the present time. We thus assume that we can talk now about the true value of temperature at noon tomorrow. Predictions that a certain characteristic takes this and this value at noon tomorrow are true if, and only if, the value given coincides with the true value for tomorrow at noon. So, if we want to say that predictions of this type can be true, we need to allow for a notion of a true value at some future time.

for instance, because the parameter is only well-defined in an idealized model of the target system. Further, some characteristics take values that cannot be ordered in a natural way, so neither can be deviations from the true value. For instance, the characteristic of the preferred musical style of a person takes values such as “classical music”, “jazz”, and “rock”, and it is dubious whether these can be ordered in a nonarbitrary way. For the time being, we will assume that both conditions on accuracy are fulfilled.

The notion of accuracy is closely related to that of error (cf. Chap. 5 by Roy in this volume). In a setting as described above for measurement and simulation, the deviation of the value obtained from the true one is called error. If a distance between both values is meaningful, the error is quantified in terms of the distance, where it is natural to use the same distance measure as for accuracy. A value obtained in a simulation is thus the more accurate the smaller the error is. *Accordingly, accuracy can informally be described as closeness to truth*.

How accurate a specific output from a computer simulation is can in principle be determined by comparing the output and the true value. But in practice, the true value is most often not known, and there is no information about the true value independently of the computer simulation. There is thus uncertainty about the error. The best thing that scientists can do then is to constrain the size of the error in some way, e.g., via knowledge about how the computer simulation functions.

There are least two basic ways in which the sizes of errors, and thus accuracy, may be constrained: In some rare cases, exact bounds on the errors can be demonstrated using mathematical proof. This assumes that the emergence of the errors can be fully described in terms of mathematics, as is the case when the errors are due to mathematical approximations. In other cases, a probability model over the errors can be specified. For instance, on the basis of experience, researchers may know that a computer simulation produces errors that follow a Gaussian probability model with a certain mean and variance. On the basis of such a probability model, researchers can specify a range of errors such that the true error is within this range with a certain probability, say of 95%. There are more sophisticated ways to express the uncertainty about an error, see e.g., Chap. 21 by Bradley in this volume. In what follows, for ease of presentation, the term “estimate” is used as an umbrella term for any way in which information about the error is specified. Obviously, “estimate” is not an estimator in the statistical sense.

Errors are generic in computer simulations (see Chap. 5 by Roy in this volume). Thus, every serious report about outputs from computer simulations is accompanied by an estimate of the errors (the same is true about measurements). In fact, what we call results of simulations incorporate, or should at least do so, estimates of errors. For an illustration, suppose that a computer simulation has output a value of 15.4910159137125 °C for the mean temperature at some place at some time in the future. The output number is affected by many errors, e.g., roundoff errors that arise because the precision with which the computer can represent irrational numbers is limited, such that there is almost no chance that it is true. What has the chance of being true is only a statement to the effect that the temperature is within a certain range of values around the output value 15.4910159137125. Indeed, some essential part of converting the numbers that are output from a simulation into propositional results is to determine estimates for the errors or their bounds. If the accuracy of an output number is taken into account in formulating the result of a simulation, e.g., by saying that the value of some characteristic is in this and this range, then the result itself, and not just an output, can be said to have a certain accuracy, viz., the accuracy that is specified. In what follows, we assume that results come with estimates of the (bounds of) errors and with a specified level of accuracy, as is common in measurement too.

Of course, to state a result into which a specific level of accuracy has been inscribed is strictly speaking no more than making a claim. The question is whether this claim is true. The answer depends on whether the accuracy that is specified is too high. Since the accuracy of a result of a simulation is most often not directly known, what scientists can at best do is to provide *evidence* for the claim that the results are such and such accurate. Accordingly, such claims become more or less credible. Thus, what researchers can at most obtain in typical cases are *more or less credible claims about the accuracy of results*. As credibility of accuracy is a very important standard, we call it cred(acc). It is important because it is the closest we can get in direction of truth.[15]

When researchers investigate a result for its cred(acc), there is a tradeoff between accuracy and credibility: Claims that the errors are smaller, or that accuracy is higher, are logically stronger and thus more difficult to justify and so less credible. Accordingly, the more accurate you want your result to be, the less credible is your claim. This point can well be illustrated using statistical methods, e.g., the confidence interval (see Chap. 19 by Robinson in this volume). This interval comprises a range of values such that the confidence to find the true value in it can be quantified using a probability. The smaller the interval, and thus the higher the accuracy, the smaller is the probability. Note though that this tradeoff does not create any real conflict, because a researcher can well formulate various claims about accuracy with different credibility. For instance, she can say that the error is smaller than 1 m with this and this probability, and smaller than 2 m with a higher level of probability. These statements do not conflict with each other. A conflict arises only if the researcher has set predefined values for both credibility and accuracy and if there is no way to reach both. Then at least one of the requirements for accuracy and credibility needs to be relaxed.[16]

So far, our discussion is premised on the assumptions that (i) there is a true value and that (ii) there is a range of possible values that can be ordered. Both assumptions can be relaxed. Turn first to outputs for which no true value exists, for instance, because the output is a parameter that presupposes an unrealistic idealization. Obvi-

[15] Although we refer now, strictly speaking, to evidence for accuracy, and not for truth, we are not appealing to a different sort of evidence. The idea is rather that we have only one type of evidence and that it typically only supports a claim of the sort that a characteristic takes such and such value with this and this accuracy, and not the claim that the characteristic takes such and such value full stop.

[16] Accuracy is also an issue in the construction of models and may then be balanced against various other desirable features. But since our focus is on the validation of an existing model and not on model construction, we need not comment on such tradeoffs.

ously, claims to the effect that such a variable takes this and this value cannot be assessed for accuracy and credibility, but nor is there any interest in their accuracy or credibility. Thus, such outputs do not matter for validation in the way other outputs do. This is not to say that the values of such parameter cannot be assessed in different terms.

Turn second to results that do not assign a characteristic a value from a range of ordered values. These are most often *qualitative* results, e.g., that a thunderstorm takes place during the next week. Often, such qualitative results arise in a simulation not because a variable in the simulation program traces whether there is a thunderstorm or not; rather, the thunderstorm in the simulation is identified on the basis of the values of other variables, e.g., differences in pressure. Drawing on work by Bogen and Woodward (1988), we may say that a (simulated) *phenomenon* is constructed from given (simulated) *data*. Unlike the claim that some temperature has a value of 15.4910159137125 °C, the claim about the thunderstorm has a reasonable chance of being true. So the most natural way to handle such claims or results during validation is to say that such claims are as accurate as they can be and that validation is only concerned with the credibility of this 100% accurate result.

To sum up our discussion about standards of validation so far: The contenders truth, credibility, and accuracy are related to each other and not really alternatives. What researchers really want is truth or, at least, that the outputs from their simulation code come closest to the true values of the characteristics of interest. But what they can only establish realistically is the credibility of claims about the accuracy of the outputs (cred(acc)). This suggests that references to truth, credibility, and accuracy in the definitions of validation merely stress different aspects of this main idea. In what follows, we will thus take truth, accuracy, and credibility together under the label of “cred(acc)”.

So far, we have dealt with truth, credibility, and accuracy as applied to *single results* (claims). But can the standards be applied *to a whole simulation* too? In principle, it is at least conceivable that a simulation model produces *true* results for all intended applications, and that all of its model assumptions are true of the target system. Likewise, a simulation model may yield results that are all *accurate* to a certain degree, and the underlying model assumptions may also have a certain degree of accuracy. And all this may become more or less credible due to suitable evidence. But all this does not happen in practice. Models are based upon simplifications, e.g., abstractions and idealizations. This means that they are never fully true of their targets, and there will be limits to the accuracy of their results. Thus, overall truth or accuracy is not a sensible standard on models (see e.g., Bailer-Jones 2003 and Parker 2009 for this argument). There is talk about the accuracy of models, but what is often meant by this is only that a certain, limited range of results from a model has a certain degree of accuracy. Such general claims may be argued for if there is sufficient knowledge about the relationship between a model and its target. We can quantify accuracy overall by first defining the accuracy of one single simulation run by, e.g., adding the squares of the accuracies in single characteristics. Defining such a measure is crucial for setting what is called a validation metric (see Oberkampf and Roy 2010, p. 68 for a definition and ibid., Chap. 13 for many examples; see also

Chap. 13 by Marks and Chap. 18 by Saam in this volume). By generalizing over all possible runs of the simulation code, we may obtain a measure of the accuracy of the code or a related model (be it the computational or the conceptual one).[17]

2.4.3.2 Adequacy for Purpose

Apart from truth, accuracy, and credibility, the main contender for a standard of validation is adequacy for purpose. A prominent paper arguing that models (be it conceptual or computational ones) should be assessed for adequacy for purpose rather than for truth is Parker (2009).

At first sight, adequacy for purpose seems quite different from truth and cred(acc) considered so far. But there are in fact connections, and to some extent, cred(acc) and "adequacy for purpose" can be combined. A first thing to note in this respect is that the purposes to which researchers can appeal in validation are limited. For instance, the possible purpose to produce beautiful animations with a simulation code does not matter for validation.

What then are legitimate purposes that researchers can appeal to when validating simulations? A prime example is prediction in a very broad sense: The computer simulation is supposed to provide information, which is typically not known otherwise. This information may refer to the future, but we will also allow for information about the present and the past. The information may be quantitative or qualitative; it may refer to a particular token event or to a type of phenomenon. Whatever the details, the purpose of obtaining information obviously leads us back to cred(acc) as follows. The purpose is fulfilled, if we obtain correct information, i.e., if the claims we obtain as predictions are true. Since there will be errors, which are not known (recall our argument above), the best we can hope for are accurate predictions with a high credibility. So, for a very broad range of purposes of computer simulations, adequacy of purpose boils down to cred(acc) of some claims.

Another purpose that is quite common for computer simulation is exploration (see e.g., Chap. 10 by Gelfert). The idea is to explore phenomena that might occur if the target system was changed. For instance, scientists may wish to explore how traffic might flow if certain speed limits were introduced. There are two ways to understand this exploration: Either scientists are interested in what is possible "as a matter of fact", i.e., given the laws of nature, etc. Then they are interested in true claims about the target system, but in claims that are different from the claims that we have investigated so far. The claims now hold that a certain characteristic *can* take this and this value. But if we assume that there are true values for various possibilities or counterfactual scenarios (and we have to make this assumption if we want to say that the claims under consideration are truth-apt), such claims can be dealt with in the same way as before. Alternatively, scientists may be interested in what is possible

[17] In defining the overall accuracy of a (simulation) model, researchers may also want to include the accuray of the assumptions underlying the model. Whether or not this is feasible will be discussed in Sec. 4.6, where we consider so-called structural validation.

given their knowledge about a target system (this is called epistemic possibility). But then the question is simply what is compatible with some assumptions (viz., the assumptions taken to be knowledge about the target system). In this case, validation is not an issue anymore because the simulation is not supposed to represent a real-world target as it is. All in all, exploration does not imply that validation has to go beyond cred(acc).

So important purposes of simulations lead back to cred(acc). Conversely, there is a route from cred(acc) to a plurality of purposes. The reason is that a typical computer simulation traces a lot of characteristics (positions, velocities of particles; the intensities of various fields at various locations…) and thus produces a lot of output. Researchers pick some characteristics as relevant, while neglecting other characteristics, when they assess cred(acc) of the simulation. For instance, climate scientists may use a simulation to predict the temperature in some region and neglect its predictions for precipitation, and validate the simulation accordingly. Now very often, a simulation that is good at predicting temperature is not good at predicting precipitation, and vice versa (see e.g., Baumberger et al. 2017b, pp. 4–5). So depending on whether a simulation is validated with an emphasis on temperature or with an emphasis on precipitation by appeal to cred(acc), the assessment will be different, and researchers have to decide which characteristics they take to be relevant to validation according to their purposes. It follows that validation is inherently purpose-relative even if it is focused on cred(acc).[18]

What emerges then, so far, is that cred(acc) and adequacy for purpose can get along with each other. For one thing, the basic purpose of obtaining information leads us back to cred(acc). This is the standard that matters for individual results of simulations as far as the purpose of gaining information is concerned. For another thing, truth and accuracy cannot sensibly be demanded of a model or simulation overall, so the implications of the model or simulations that are assessed for their cred(acc) have to be chosen according to purposes.

In fact, if cred(acc) is restricted to certain results of a model, it spells out what adequate or accurate representation in virtue of this model is (cf. Val-AIAA and Val-Roache). For to claim that a model delivers an adequate representation of its target is to say that we can infer a lot of information about the target by using the model (this is e.g., manifest in Suárez' 2004 account of modeling). Now, information is specified in terms of true claims, and the more accurate the claims are, the more information they give us. Since truth and accuracy cannot be read off from claims, what we can assess only is their credibility. And because a model is not supposed to represent its target in every respect, cred(acc) is restricted to some claims that can be inferred using the model.

However, there are different scientific purposes that may be pursued with simulations. The most important candidate is explanation (mentioned, e.g., in Val-Mäs):

[18] In practice, the specification of the purpose and the intended applications of a computer simulation is often less than precise. This seems to allow that a simulation is later applied to systems that are quite different from the target system that was originally intended. But such an application would need new validation.

Researchers may want to explain some phenomenon, say a certain chemical reaction, using simulations, and explanatory aims pursued with simulations may matter in validation. To see whether this introduces additional standards in validation, we need to answer two questions: 1. Can the purpose of explanation be spelled out in terms of cred(acc) too? 2. If not, i.e., if different standards become important, is it still a matter of validation to assess simulations following these standards? These questions are relevant not just for simulations, but also, for e.g., theories, but we will here discuss them with a focus on simulations. Our discussion is concentrated on simulations that are supposed to explain something.

To address the first question, consider a social scientist who tries to explain an event at the level of society, say a revolution, in terms of an agent-based model that takes into account the degree to which individual people are willing to use violence, the degree of their dissatisfaction with the political system, and possible causes thereof. As a result of a computer simulation study, she comes up with a certain explanation, e.g., that the revolution was due to corruption in the system and the dissatisfaction that it produced.[19] What might a validation be that takes this explanation into account? A first answer is that explanations consist of claims too, e.g., that the dissatisfaction with the political system was large. Such claims may be thought of as results of the simulation and then be validated in terms of truth or, since this is not possible, cred(acc), or so the first answer is.

But there are problems with this answer. Explanatory claims are quite special as far as evidence for them is concerned. For instance, causal claims are explanatory, but as philosopher David Hume famously insisted, we cannot observe that A causes B (e.g., Hume 1748, Sect. VII, Part I, 50). True, researchers do have methods to establish explanatory, in particular causal claims (see e.g., Pearl 2000 about causal inference), but the question is whether related methods introduce different standards than cred(acc).

There is a case for a positive answer to this question. Researchers need often to choose among competing candidate explanations to settle on the real explanation. They will, of course, pick the best explanation. They then infer that the assumptions introduced by the best explanation are true. This inference is called inference to the best explanation (see Harman 1965 for the introduction of this term and Lipton 2000 for discussion). Now it is arguable that explanations are, other things being equal, better when they are simpler or more elegant. As far as computer simulations are concerned, some simulation may incorporate explanatory assumptions that are simpler and less complex than those from another. Simplicity and elegance are clearly standards that differ from cred(acc).

But the appeal to standards such as simplicity and elegance to infer explanatory statements is controversial. For instance, van Fraassen (1980) is a famous skeptic about inference to the best explanation quite generally. It is also debatable whether simplicity and elegance are really features that render an explanation more likely

[19] See Grüne-Yanoff (2009) for a philosophical analysis of explanatory claims obtained from agent-based simulations.

to be true. So a definite answer to our first question above needs to await closer philosophical investigation.

Turn now to our second question. Even if elegance, simplicity etc. count as virtues that allow to infer explanatory claims, they are only relevant for validation if computer simulations as such are in the business of picking the correct explanation and if conclusions about the best explanation form part of the results from a simulation. But it is very dubious whether this is so. Computer simulations can clearly yield the so-called how-possibly explanations: They can show that a certain combination of factors would lead to a certain type of phenomenon. They can do so by drawing the consequences from assumptions to the effect that some combination of factors obtains. But to judge that a particular how-possibly explanation is better than another (e.g., because it is simpler or more elegant) is not a matter of running computer simulations, nor is the inference to an explanation that is judged best. Such inferences go beyond simulations and should thus not matter for validation.

All this is not to deny that computer simulations, i.e., the underlying models and their results, can be assessed for their explanatory power. But the question is whether such an assessment is part of validation. There is a case for a notion of validation that is restricted to some purposes and standards that do not go beyond cred(acc), as applied to some aspects of a simulation. First, a lot of authors think that the relevant standards in validation comprise accuracy and credibility, but not simplicity or elegance. This is evident from the most influential definitions of validation. In the terms familiar from conceptual engineering, we here appeal to similarity, i.e., to the way a concept is used as a matter of fact. Second, a more restricted notion of validation is likely more fruitful than a broader one. As argued above, cred(acc) can be used to spell out that something is a good representation. Note also that cred(acc) straightforwardly applies to results of simulations, while elegance and simplicity do not. So it is arguable that the assessment of a simulation in terms of cred(acc) is closely related to an assessment of the results, whereas standards such as simplicity, etc., are different because they only concern the simulation program or the conceptual model as a whole.

This is not the place to decide the issue about which standards should matter for validation. A lot of authors from this volume, e.g., Oberkampf (see Chap. 3 in this volume) would resist attempts to respect standards that move beyond cred(acc). A few other authors, e.g., Saam in her Chaps. 9 and 17 as well as Hirsch Hadorn and Baumberger in their Chap. 40 seem more sympathetic to a broader understanding of validation that includes standards such as simplicity. Instead of settling the question, our aim here is to provide some understanding of the issues involved.[20]

In any case, every decision on the issue has consequences for elaborating the understanding of validation. If several different standards are involved in validation, the question emerges whether validation is exhausted by an assessment with respect

[20] There are other purposes that are typical of scientific inquiry and that may be postulated to matter for validation too. For instance, to the extent that understanding goes beyond explanation (see Baumberger et al. 2017a for a recent overview of the debate about understanding), it may be claimed that validation appeals to values that matter for understanding. Such values may be discussed in the same way as we have discussed explanatory virtues.

to each of the standards, or whether the results of this exercise are combined to arrive at an overall assessment of a simulation or its results. The term “validation” suggests that such an overall assessment is reached, because validation seems oriented after validity, which seems to be one standard. The question then is how the various standards are prioritized or weighted to obtain an overall assessment (an analogous question is important for theory choice too where various standards compete, see e.g., Kuhn 1977) . This is not so much of a problem if the focus is on cred(acc) because accuracy and credibility combine in a certain way.

After this quite involved discussion about the standards of evaluation, we can now turn to other dimensions on which the definitions differ. Here, the discussion can be much briefer. To simplify the discussion, in what follows, our focus is mainly on cred(acc) applied to a restricted part of the possible results.

2.4.4 Difference 3: Type of Evaluation

Some definitions of validation take it that validation is about determining the extent to which the standards constitutive of validation are met. The idea is to place a simulation on the scale of all possible values of, say, accuracy. The expression “determining the degree” in Val-AIAA refers to this idea (Val-AIAA refers to intended uses too, but the intended uses are meant to determine aspects in which the representation is adequate, and not a threshold for a pass-fail-criterion). Other definitions, by contrast, assume that the point of validation is to check whether or not some predefined requirements are fulfilled by certain results or by a simulation. In this way, e.g., Val-SCS speaks of a “satisfactory RANGE OF ACCURACY”. The difference can succinctly be characterized in terms of the type of verdict that the evaluation leads to: Is this verdict of the type “The score of the simulation (result) on the relevant scale is such and such” (this assumes that the scale can be measured in terms of numbers) or of the type “The requirement set has (not) been fulfilled.”?

Both answers are compatible with the idea that the basic standard(s) constitutive of validation may be fulfilled to various degrees. The difference between the answers is rather that, under the first, but not the second, requirements, e.g., thresholds, have been set. This leads to a binary set of outcomes of validation depending on whether the requirement has been satisfied or not.

How may we argue for one or the other view? One argument goes as follows: Validation is about validity, but validity is not a matter of degree. So, validation should involve a binary test. But this argument is very dubious because it is not clear what sort of validity is the aim of validation. It is clear, for instance, that we are not talking about logical validity. The claim that validation is about validity needs elaboration, and our attempts of specifying the standards constitutive of validation have led to standards that may be fulfilled to various degrees.

A different argument is that binary tests have an epistemological significance that evaluations without such a test lack. This point may be supported with the following intuition: If a simulation (result) passes a test with a predefined requirement, then it has achieved something. There was a risk that the test was not passed. No such risk is undertaken if a simulation (result) is just placed on a broad spectrum of possible outcomes. This thought accords well with a Popperian outlook of science, which stresses the testing of hypotheses (see Chap. 6 by Beven in this volume, see also below). But it is dubious whether it is indeed advantageous to frame validation as a binary test. Simulations can be put to many uses, and whether they prove useful in one or the other regard depends on the extent to which a standard is fulfilled, e.g., the extent to which a simulation (result) is accurate in a certain respect. If we place a simulation (result) in the space of possible values of accuracy and find that it is here and there, it follows that it is apt for some uses (i.e., those that require no more accuracy), but not for others. This means that we may reject the simulation for some uses. It is not clear what the added value is if we frame validation in terms of a test with a single requirement. Rather the opposite, we are more flexible if we do not do so (see Roache 2013, p. 68 for this argument).

If validation is conceived of as a binary test, an additional question emerges: Does validation of a computer simulation (result) imply that the test is passed? Or is validation the test independent of the result? If validation entails that the test is passed, "validation" is a so-called success term, while mere attempts at validation count as validation, if no passing of the test is required. There is some case for the former view (see Roache 2013, p. 69 following I. Celik), since it would be strange to say that a simulation was validated, if it has not passed the test. This point may be extended to an argument in favor of the "testing view" discussed above: It would be strange so state that a simulation was validated if it was assessed and if a low degree of accuracy was found (see ibid.). But it is not clear whether this point about the understanding of a very specific phrase ("has been validated") should outweigh other considerations, e.g., flexibility in use.

2.4.5 Difference 4: Cogency (Degree of Credibility)

Some of the definitions of validation stress that a certain degree of accuracy (or the fulfillment of an alternative requirement) has to be *demonstrated* (cf. the definition of accuracy used in Val-SCS) or shown (see Chap. 6, Sect. 1 by Beven in this volume). This requires something like a proof that a certain accuracy is reached. This requirement is quite strong (see below), and it may be relaxed e.g., by saying that a claim about the degree of accuracy needs only to be justified to some extent or that a certain case be made for it (cf. Chap. 37, Sect. 2 by Beisbart in this volume). Let us thus say that the definitions vary regarding the degree of *cogency* that is required for the verdict reached by validation.

Now if a definition of validation requires that *accuracy* of the simulation be shown with this and this cogency, this is in effect a statement about cred(acc): The claim that the results of the simulation have a certain accuracy is supposed to have a certain degree or level of credibility. If other standards figure in a definition of validation, then cogency is the credibility of the claim that a simulation complies with the standards.

This is at least so if the evidence that bears on the credibility is the same evidence that creates cogency, as should be the case.

Whether or not cogency is cred(acc) or the credibility of other claims, the question that emerges for the definition of validation is whether a certain minimal degree of credibility is required for validation (e.g., reasonable plausibility, strong or even conclusive reasons). Note that, in case cred(acc) is the only relevant standard for validation, this question is a special case of the question considered in the last subsection. But the question is different from the one considered in the previous subsection, when we talk about different standards. The question then reads how credible it needs to be that a different standard is realized to such and such degree.

If some definitions require high credibility, e.g., by demanding that a certain accuracy be *demonstrated*, this is well understandable. High credibility is a general standard in the sciences. It is often expected that the sciences deliver knowledge, which needs a high degree of credibility. It thus seems more than reasonable to keep up the standards. However, there is also the question of which degree of credibility can realistically be attained. This is a big theme in Oreskes et al. (1994), who argue that typical results of simulations cannot be shown to be true.

The question of how much credibility is possible crucially depends on what precisely is the object of validation. We can get as close as is possible to something like a demonstration when we talk about one result with a specified accuracy, e.g., that the value of a temperature is in a certain range. If we have a measurement of this temperature, we can check whether the temperature is in the range specified by the simulation result. Even if this is the case, this does not amount to a proof that the result is accurate enough, since the measured value of the temperature may be wrong due to errors in the measurement. This is why measurement errors have to be taken into account when simulation results and real data are compared. Consequently, what can at most be attained is a high credibility of the claim that a simulation result and a measurement outcome are compatible within some specified accuracy. This is not proof, but it comes as close as one can get to show that the results are as accurate as has been specified.

But a situation in which we can directly compare simulation results with measurement results in the way described is neither particularly interesting nor common. It is not interesting, because simulations are barely needed if we have measured results about the characteristics of interest (at least if the measurements are highly credible). It is not common because we often lack measurements corresponding to the outputs of computer simulation. This is so if the results refer to the future or to unobservable characteristics. The problem is even more significant when we turn to the validation of whole computer simulation models, because they can yield a lot of results depending on the initial conditions set.

In such cases, the credibility of the results cannot be directly established by comparing with data. Rather, we need some argument to the effect that the results are such and such accurate. The argument will start from certain premises that support the conclusion that the result is such and such accurate. Depending on how credible the premises are and how strongly they support the conclusion, the claim that the result is such and such accurate is made more or less credible. It is no surprise then that frameworks inspired by argumentation theory have been proposed for validation. Baumberger et al. (2017b) develop a framework for arguing that projections from climate simulations get it right, but their account can easily be extended to other simulations. Saam in Chap. 18 in this volume uses Toulmin's model of argument to propose a framework for validation.

As far as the concept of validation is concerned, it seems prudent not to require a specific degree of credibility for successful validation. It is more appropriate to stipulate that validation determines the degree to which a simulation (result) is credible. Even results that are not highly credible may be useful, for instance, in the framework of Bayesian decision theory, or if the question is whether there is a certain risk at all (see Resnik 2003). Oreskes et al. (1994) are certainly right to stress that something as cogent as mathematical proof is not possible in validation. In this respect, the definition SCS-Val is too strict since it requires demonstration of a certain accuracy. To what extent arguments allow for a high credibility of simulation results is a philosophical question to which we return in the next subsection.

2.4.6 Difference 5: Empirical Methodology

Some definitions of validation, e.g., Val-Roache and Val-Mäs, constrain validation by requiring a certain method, viz., the comparison between outputs or results from the simulation with measured data. We assume that the measurements need not come from experiments proper (as suggested by Val-Roache), but may also stem from mere observations (e.g., astronomical observations). We further take it that the comparison between simulation outputs and measured data refers to a limited number of data points (i.e., characteristics at certain times and certain locations) that do not exhaust the actual results of a simulation, not to mention the possible results that may be obtained from running a code. Consequently, an inference is necessary that mediates between the comparisons between simulation output and measured data, on the one hand, and additional (possible) results from the simulation, on the other. This inference can be cast as an argument. It is inductive in the sense that the premises (which state a certain agreement between simulation output and measured data), even if true, do not guarantee that the conclusion (that additional results hold) are true. In typical cases, the inference runs from certain times (for which a comparison between simulation output and measured data has been made) to other times (for which a prediction is made); or from some types of situations (for which a comparison between simulation outputs and measurements was available) to others (which are described in terms of different initial conditions and parameter values). The inference is much more dangerous when it runs from some types of characteristics (that have been found to agree between simulation and measurements, e.g., temperature) to others (e.g., precipitation, for which predictions are of interest).

Is it a good idea to require per definition that validation be carried out via comparing simulation output and measured data, for short: that it is data-driven (see Roache 2013, pp. 71–72 for a brief discussion)? The answer depends on two further questions: 1. What is the significance of a comparison between simulation output and measured data? 2. Are there any alternative methods that can have a comparable significance for validation? Let us take both questions in turn.

1. Assume for the sake of the argument that simulation outputs and measured data agree within a certain range of accuracy to some credibility. To simplify the presentation, we will now suppress talk of accuracy and credibility of the agreement. What does the agreement mean for other outputs from the simulation? It is clear that there is no guarantee that other outputs from the simulation have the same accuracy.[21] Logically speaking, the situation is like this: We have a simulation or a broad set of (possible) simulation outputs. Both give rise to a broad set of claims (viz. that the outputs reflect the values of characteristics in the target system). This set of claims can be regarded as something like a theory, a model or a strong hypothesis from which a lot of specific claims about values of characteristics from the target system follow. Some of the latter claims have been found to agree with measured data, but the hypothesis has excess content. Inferring the hypothesis from the agreement would constitute an inductive inference.

The question of what can be rationally inferred in this situation is controversial in philosophical circles. Popperians shun inductive inference from science since they take it that Hume has shown induction to be irrational (see Beven, Chap. 6 in this volume for a Popperian outlook; see also Chap. 27 by Roache). Consequently, the agreement between the simulation outputs and the measurements for some data points has no significance. Popperians instead focus on the falsification of scientific hypotheses in terms of observations. The simulation (for instance qua underlying model) could have been falsified by comparing the simulation outputs and the measurements. In this case, the simulation should have been rejected. If falsification fails despite severe testing, a hypothesis or a simulation is said to be corroborated, but this is not supposed to mean that it has become more credible.

Other philosophers think that a hypothesis is to some extent confirmed if some of its consequences agree with measurements. Here, confirmation is much less than proof, the idea rather being that some case, however minimal, for the hypothesis has been made (see Hempel 1945, Sects. 1–2 for basic clarifications). This is expressed in quantitative terms in Bayesian epistemology, where credibility (for some person) is expressed in terms of probabilities. Using Bayesian updating, the credibility of a hypothesis is increased if it coincides with observed evidence (see Chap. 7 by Beisbart in this volume). For Bayesians it depends on the credibility of other hypotheses (or the agent's probability function over a whole set of hypotheses) how much the credibility of some hypothesis (here e.g., a simulation model) is raised. This seems

[21] This is even true for outputs within the so-called validation domain (see e.g., Oberkampf and Roy 2010, pp. 39–44). This domain is often constructed from the data points for which measurements and simulation outputs have been obtained. The idea is that these points sample a larger domain for which the simulation has been validated. As is rightly stressed, quite often a lot of intended applications are not from the validation domain. But the so-called validation domain too is only sampled using a limited number of data points. So there is no guarantee that the simulation works as intended even in the validation domain. This is not to deny that it is often reasonable to trust results obtained in the validation domain.

appropriate, because it often depends on background knowledge to what extent an inductive inference can legitimately raise one's confidence in the conclusion. This point is stressed too by J. N. Norton's account of induction (see e.g., Norton 2003). Very roughly, for Norton, particular inductive inferences must be underwritten by known facts to be legitimate. In a similar vein, Harman (1965) suggests that it is only reasonable to infer from a limited sample of data points to a more general hypothesis, if this is implied by the best explanation we have for the data points.

So independently of whether we follow Bayesians, Norton or Harman, the significance of the agreement between some outputs and measurements for a broader hypothesis (e.g., that a certain simulation model is accurate more generally) crucially depends on the context. The extent to which we can rationally take the broader hypothesis credible is a function of what we know otherwise. This is as it should be. For instance, it is a substantial question of whether a simulation the outputs of which match observed values of the temperature, can predict precipitation too. This question cannot be answered a priori but needs to be addressed to make an inference from agreement about temperature to predictions of precipitation and thus to validate a simulation more broadly.

What then is required to answer this question in practice? One option is knowledge that a model describes the interconnections between certain characteristic (here temperature and precipitation) appropriately because it represents the underlying mechanisms appropriately. Then, if there is strong evidence that the temperature is correctly described by the model, it is likely that the model describes precipitation well too (cf. Baumberger et al. 2017b). A slightly different option (cf. Chap. 41 by Frisch in this volume for a similar case) is some trust that a model describes a certain range of aspects of the target system well (here we assume that the range of aspects has been fixed e.g., by knowledge about how the model was constructed). If the model is then shown to agree with measured temperature, for instance, then we may say that the model as a whole has been confirmed, as far as the fixed range of aspects is concerned. In the terms suggested by Harman, we may say that the best explanation for the agreement is that the model does indeed work well for the whole range of aspects. If precipitation is one of these aspects, we can reasonably infer that the model can predict precipitation too (to some accuracy, with some credibility). It is, of course, a substantial question whether it is in fact part of the best explanation of the agreement that the model is good in the whole range of aspects.

To sum up then, if we don't turn Popperian and deny that inductive inference can bestow any credibility (and I have no hesitation to reject the Popperian view), then we should say that the significance of the agreement between simulation output and measured data depends very much on the context. What scientists can reasonably infer depends on what they know about the simulation and the target system.[22]

2. The second question is whether there is any alternative to validation in terms of comparing simulation outputs to measured data. In terms of arguments, the question is whether we can use different premises to make a case for results from a simulation.

[22] In Chap. 41 in this volume, Frisch addresses the question of how significant agreement between simulation outputs and measured data is if the simulation has been tuned to agree with the data.

There seems to be a case for a positive answer as follows: To make results of a simulation credible, we can argue that the simulations incorporate a model that has been confirmed a lot. Suppose for instance that we have a system of almost rigid bodies, for which Newton's laws and some force law are known to hold with high accuracy. Suppose further that we know very well the geometry of a container in which they are enclosed (i.e., the boundary conditions) and their initial positions and velocities. Then we have a strong argument from prior knowledge (or something close to it) to the accuracy of the conceptual model.[23] Suppose now further that this model has been implemented in a computer simulation and that the simulation can be shown to approximate the solutions to the conceptual model with sufficient accuracy. This is to say that the simulations have been verified. Then it follows that the computational model is accurate to some degree (see Chap. 42 by Beisbart for a more elaborate version of this argument). Altogether, we seem to have an argument for the accuracy of the simulations that is not built on a comparison between simulation output and measured data. So we seem to have validation that is not data-driven (cf. Parker 2008, pp. 170–171 for this strategy).

But there are a number of problems with the claim that validation may not be driven by data. First, the scope of such validation is very narrow. As Oreskes et al. (1994) remind us, hitherto unknown causes may interfere with the target system and e.g., change its geometry. This would invalidate the simulations. Also, very often, only some part of the assumptions that enter a simulation is known to hold of the target system, while others are only assumptions. If this is so, then the case suggested for the accuracy of the simulation crumbles. As a consequence, there are very serious limitations to validation that is not data-driven.

At a different, conceptual level, there are two other objections to the idea that validation may not be data driven. It may first be objected that claims to the effect that validation may not be driven by data are a cheat because the argument that is given for the accuracy of the simulations implies that validation has already been achieved before. But this objection is too quick. It is true that some validation has been achieved before, and in this sense, the proposed form of validation is derivative. But the validation that was in fact done before applies to the assumptions of the conceptual model and not to the simulations. Some crucial assumptions from the conceptual model may be drawn from a theory and thus have been confirmed using data from systems different from the target. The validation of the simulations consists of a new argument that takes the validation of the conceptual model as one premise and the verification as another. This is a new argument that validates a certain simulation.

The second conceptual objection is as follows: The type of validation that has been proposed as not driven by data just boils down to prediction. But again this objection is too quick. A prediction is just a claim. But the proposed form of validation is not just a claim but rather an argument that is supposed to support predictions.

The upshot then is that there is in principle a method that can be used to make simulations and their results credible and that is not driven by data. This is some case

[23] In the terms of Fagiolo et al. (Chap. 31 in this volume), showing that the model is adequate in this way would qualify as input validation.

for a notion of validation that does not always require the comparison between simulation outputs and measured data. This conclusion accords very well with the answer we have suggested for the first question, viz. that the significance of an agreement between simulation outputs and measurements is somehow limited because we need knowledge about the model and its relation to the target to infer the credibility and accuracy of other results. Despite this, it may still be argued that validation does need a comparison between simulation outputs and measurements, for instance, because this is a well-established understanding of validation (Carnap's similarity). In any case, whatever, we think about the possibility of a form of validation that is not data-driven, it is true that, in practice, the comparison between simulation output and measured data is most often absolutely needed for validation.

As far as data-driven validation is concerned, we may ask how far this method can be pushed. In particular, can the agreement between simulation output and measured data be used to show that not just further results, but also the model assumptions are true or sufficiently accurate? To raise this question is effectively to ask whether the validation of a simulation is exhausted by validation of its (possible) results or whether it can address the underlying model assumptions too.

Consider thus the following setting: A simulation, i.e., the conceptual or the computational model, assumes a certain structure for the target, i.e., certain components with such and such properties and such and such mutual relationships. Suppose that this structure cannot be compared to the structure of the target system because the latter is not fully known. For instance, an agent-based simulation assumes that a certain group of people have such and such beliefs and behave following these and these rules, although it is not known whether the rules hold of the target system. The hope is nevertheless to make a case that the model structure is correct, at least to some combination of accuracy and credibility, and thus to validate the model in the following way: The model is run using a computer simulation, some observable consequences of the model structure (e.g., about prices on a certain market) are taken and compared to measured data. The hope then is that agreement between the model output and measurements makes a case for the structure of the model being correct. Following Chap. 31 by Fagiolo et al. in this volume, we may call this structural validation.[24]

It is in principle correct that the confidence in a model structure is boosted in this way. For instance, in the terms of Bayesian conditionalization, it can be shown that the confidence in the model structure increases if the prior probability of the observed data is smaller than the probability of the data given the model structure. Since the latter probability is very high in our setting, there will most often be a boost in credibility. However, this boost in credibility may be very small and, often, the overall credibility reached for the model structure will not be very high if there are other model structures that produce the same effects about the data (cf. Chap. 32 by Beven in this volume). What is more realistic is that a small subset of model assumptions become quite credible if they were combined with very well-confirmed

[24] If the focus is on the conceptual model and its structure, then verification is crucial for this sort of validation.

model assumptions and if there was good agreement between simulation output and measured data.

It may be objected that structural validation can make a model very credible, if the model is inferred to be the best explanation of the data to which the simulation output has been compared. But what can be established with the simulation alone is at best a how-possibly explanation of some data, i.e., a possible way in which the latter may have been produced. It's a significant further step to argue that the explanation is best and thus true. In particular, it is arguable that theoretical virtues that go beyond cred(acc) are relevant for this step, and it is debatable whether such standards should matter in validation. The consequence is that it is debatable too whether or not structural validation should appeal to such virtues.

Altogether, structural validation is possible in the sense that agreement between simulation output and measured data can increase the confidence in assumptions that underly a simulation. But very often, structural validation will not reach a high level of credibility, and it is arguable that genuinely explanatory concerns that hinge on simplicity, etc., should not matter in structural validation.

2.5 Conclusions

To conclude, let me summarize the main results of our discussion. The definitions of validation of simulations that I have looked at concur in taking validation to be an evaluation. But there are disagreements about what is evaluated, what the standards of evaluation are, what type of verdict the evaluation results in, what sort of cogency is required and whether the evaluation proceeds only in terms of a comparison between simulation output and measured data.

We can put this together by proposing a scheme for definitions of validation. The scheme leaves certain lacunae that may be filled in different ways, which leads to various full definitions of validation:

> Validation of computer simulation is a ___________(3b) ___________(3a) evaluation of ___________(1) following the standards ___________(2) with cogency ___________(4) and using ___________(5)

The numbers follow the numbers of the differences noted in Sect. 2.4 above.

Lacuna (1) is to be filled with the proper object of simulation validation. Natural candidates are simulation results, the simulation codes, the computational or the conceptual model. The different ways to fill the lacuna do not really lead to substantial disagreement because they lead to different types of validation: validation of results, a code, etc., and it is clear that we want to leave space for the validation of results, code etc. To make this clearer we may define not the validation of computer simulation (as our scheme suggests), but rather the validation of ___________(1). But our discussion suggests that, what is really of most interest, is the validation of a simulation code or, equivalently, the computational model for some range of intended applications, where this is equivalent to the validation of a huge range of results that may be obtained using the code. What is also of interest is the validation of the conceptual model, which is typically different from the computational one. An interesting question is to what degree models, be it conceptual or computational ones can be validated not just concerning their predictions, but also their underlying model assumptions.

The lacuna (2) is for the standards appealed to in validation. The most natural candidate is what we have called cred(acc) in certain respects, i.e., credibility of accuracy regarding certain characteristics. What is often mentioned too is adequacy for purpose. I have argued that this standard can to some extent be unpacked using accuracy and credibility. But some people might want to include other standards such as simplicity.

The lacunae (3a) and (3b) are supposed to specify the kind of evaluation or, more specifically, the type of verdict that is reached. What is meant here is difficult to express in natural language, so let us coin technical terms. Let us say that evaluation is binary if it only allows for the outcomes “requirements met” and “requirements not met”; otherwise, it is called nonbinary. Lacuna (3a) is thus to be filled with “binary” or “non-binary”. If the “binary” is chosen, lacuna (3b) may be filled with “successful” or be left as it is. In the first case, the idea is that validation is only achieved if it the requirements are met.

Lacuna (4) allows one to require a specific degree of credibility. Otherwise, the lacuna is just canceled.

Lacuna (5) permits one to stipulate that all validation is done via a comparison with empirical data. Otherwise, the lacuna is dropped.

In this chapter, we have discussed several ways of filling the lacunae. This is not the place to decide between the options and to settle on a unique definition of validation. It is a task for future research to see whether a unique definition of validation can usefully be applied in different fields. The answer may well be no. It may turn out that some lacunae can only be filled relative to purposes that are specific to a discipline, a type of simulation, etc. This is not a problem at all as long as simulation scientists make clear in which sense they talk about validation.

Acknowledgements I am extremely grateful for extensive comments by William Oberkampf, Patrick Roache, and Nicole J. Saam. At a very different level, I wish to express my thanks to my mother, Bärbel Beisbart, who sadly passed away during the time when I was working on this chapter. I dedicate this chapter to her memory.

References

AIAA. (1998). Guide for the verification and validation of computational fluid dynamics simulations, AIAA G-077–1998. American Institute of Aeronautics and Astronautics, Reston, VA.

ASME. (2006). Guide for verification and validation in computational solid mechanics. American Society of Mechanical Engineers, ASME V&V 10-2006.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review, 97*(3), 303–352.

Blackburn, S. (1999). *Think*. Oxford: Oxford University Press.

Bailer-Jones, D. M. (2003). When scientic models represent. *International Studies in the Philosophy of Science, 17,* 59–75.

Beisbart, C. (2012). How can computer simulations produce new knowledge? *European Journal for Philosophy of Science, 2*(3), 395–434.

Beisbart, C. (2014). Are we sims? How computer simulations represent and what this means for the simulation argument, The Monist 97/3 (2014, special issue edited by P. Humphreys), S. 399–417.

Brun, G. (2016). Explication as a method of conceptual re-engineering. *Erkenntnis, 81,* 1211–1241.

Beisbart, C. (2017). Advancing knowledge through computer simulations? A socratic exercise. In: M. Resch, A. Kaminski, P. Gehring (Eds.), *The science and art of simulation I. Exploring–understanding–knowing* (pp. 153–174). Cham:Springer.

Baumberger, C., Beisbart, C., & Brun, G. (2017a). What is understanding? An overview of recent debates in epistemology and philosophy of science. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 1–34). New York: Routledge.

Baumberger, C., Knutti, R., & Hirsch Hadorn, G. (2017b). Building confidence in climate model projections: An analysis of inferences from fit. *WIREs Climate Change, 8,* e454. https://doi.org/10.1002/wcc.454.

Caldwell, S., & Morrison, R. J. (2000). Validation of longitudinal dynamic microsimulation models. Experience with CORSIM and DYNACAN. In: Mitton, L., Sutherland, H., & Weeks, M. J. (Eds.). *Microsimulation modelling for policy analysis. challenges and innovations* (pp. 200–225). Cambridge: Cambridge University Press.

Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering*. Oxford: Oxford University Press.

Carnap, R. (1950/1962) *Logical foundations of probability*. Chicago: University of Chicago Press.

Grüne-Yanoff, T. (2009). The explanatory potential of artificial societies. *Synthese, 169,* 539–555.

Ghetiu, T., Polack, F. A., & Bown, J. (2010). Argument-driven validation of computer simulations–A necessity rather than an option. In: *VALID 2010. The Second International Conference on Advances in System Testing and Validation Lifecycle* (pp. 1–4) August 22–27, 2010. Nice, France, IEEE Press.

Harman, G. H. (1965). The inference to the best explanation. *Philosophical Review, 74*(1), 88–95.

Hartmann, S. (1996). The world as a process: simulations in the natural and social sciences. In R. Hegselmann, et al. (Eds.), *Modelling and simulation in the social sciences from the philosophy of science point of view, theory and decision library* (pp. 77–100). Dordrecht: Kluwer.

Hempel, C. G. (1945). *Studies in the logic of confirmation (I.)* (Vol. 54, No. 213, pp. 1–260) Mind, New Series.

Hume, D. (1748). Essays concerning human understanding [now known as: An Enquiry Concerning Human Understanding], London: A. Millar, many new editions.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. New York: Oxford University Press.

IEEE. (2012). IEEE standard for system and software verification and validation. In *IEEE Std 1012-2012 (Revision of IEEE Std 1012-2004)* (pp. 1–223), 25 May 2012. https://doi.org/10.1109/ieeestd.2012.6204026.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In T. S. Kuhn (Ed.), *The essential tension* (pp. 320–339). Chicago: University of Chicago Press.

Künne, W. (2003). *Conceptions of truth*. Oxford: Clarendon Press.

Lacey, H. (1999). *Is science value-free? Values and scientific understanding*. London: Routledge.

Lipton, P. (2000). *Inference to the best explanation* (2nd ed.). London: Routledge.

Margolis, E., & Laurence, S. (2014). Concepts. In: E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/spr2014/entries/concepts.

Naylor, T. H., & Finger, J. M. (1967). Verification of computer simulation models. *Management Science, 14,* B92–B101.

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science, 70,* 647–670.

Oberkampf, W. & Roy, C. (2010). Verification and validation in scientific computing. Cambridge University Press.

Oberkampf, W. L., & Tucano, T. G. (2008). Verification and validation benchmarks. *Nuclear Engineering and Design, 238*(3), 716–743.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science, 263,* 641–646.

Parker, W. S. (2008). Franklin, Holmes, and the epistemology of computer simulation. *International Studies in the Philosophy of Science, 22*(2), 165–183.

Parker, W. S. (2009). Confirmation and adequacy-for-purpose in climate modeling. *Aristotelian Society Supplementary, 83,* 233–249.

Pearl, J. (2000). *Causality. Modeling, reasoning, and inference*. Cambridge: Cambridge University Press.

Resnik, D. B. (2003). Is the precautionary principle unscientific? *Studies in History and Philosophy of Biological and Biomedical Sciences, 34,* 329–344. https://doi.org/10.1016/S1369-8486(02)00074-2.

Roache, P. J. (1998). *Verification and validation in computational science and engineering*. Albuquerque, New Mexico: Hermosa Publishers.

Roache, P. J. (2009). Perspective: Validation—What does it mean?. *Journals of Fluids Engineering*; *131*(3), 034503–034503-4. https://doi.org/10.1115/1.3077134 (here quoted after the reprint in Roache 2013).

Roache, P. J. (2013). *A defense of computational physics*. Socorro, NM: Hermosa (revised printing).

Schlesinger, S., et al. (1979). Terminology for model credibility. *Simulation, 32,* 103–104.

Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science, 71,* 767–779.

van Fraassen, B. (1980). *The scientific image*. Oxford: Clarendon Press.

Winsberg, E. (1999). Sanctioning models. The epistemology of simulation. *Science in Context, 12,* 275–292.

Chapter 3
Simulation Accuracy, Uncertainty, and Predictive Capability: A Physical Sciences Perspective

William L. Oberkampf

Abstract Most computational analysts, as well as most governmental policy-makers and the public, view computational simulation accuracy as a good agreement of simulation results with empirical measurements. However, decision-makers, such as business managers and safety regulators who rely on simulation for decision support, view computational simulation accuracy as much more than agreement of simulation results with experimental data. Decision-makers' concept of accuracy is better captured by the term *predictive capability* of the simulation. Predictive capability meaning the use of a computational model to foretell or forecast the response of a system to conditions without available experimental data, even for system responses that have never occurred in nature. This chapter makes this important distinction by discussing the crucial ingredients needed for predictive capability: code verification, solution (or calculation) verification, model validation, model calibration, and predictive uncertainty estimation. Each of these ingredients is required, whether the simulation results are used in the generation of new knowledge, or for decision support by business managers, government policy-makers, or safety regulators.

Keywords Verification · Validation · Uncertainty quantification · Predictive capability · Model calibration

Acronyms

CFD	Computational fluid dynamics
MFE	Model form error
MMS	Method of manufactured solutions
PDE	Partial differential equation
SRQ	System response quantity
SQA	Software quality assurance

W. L. Oberkampf (✉)
Sandia National Laboratories - retired, Albuquerque, NM, USA
e-mail: wloconsulting@gmail.com

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_3

3.1 Introduction

Computational simulation plays an ever-increasing role in scientific research, business pursuits, and government activities. Combined with dazzling computer graphics and phenomenal temporal and spatial resolution, computational simulation is revealing details of physical processes and systems never seen before. From a historical perspective, however, computational simulation is newfound in its contributions to human activities and scientific knowledge, and therefore is not considered a mature and trustworthy provider of knowledge and information. While some may find this characterization unduly harsh, it can be argued that the history of the natural sciences, including both the theoretical and experimental traditions, required centuries of development and refinement to be considered trustworthy. Even then, some of the foundational concepts were occasionally destroyed and rebuilt on a more solid underpinning.

Computational simulation is extraordinarily interesting not only because of its potential capability in essentially every aspect of human activity, but also because it intertwines the modeling of real-world systems, mathematics, and high-performance computing. With the ever-increasing worldwide access to high-performance computing, the growth of computational simulation will only accelerate in the future. The issues addressed in this chapter, as well as other chapters in this book, deal with the accuracy and uncertainty of simulation results, the ability of simulations to predict future events (most never seen before), and how simulations are increasing their influence on decision-making in business, governmental policy, and regulatory safety. This chapter addresses the simulation of any type of inanimate (nonliving) physical system. Such simulations can deal with physics, engineering, chemistry, astronomy, earth sciences, atmospheric sciences, and oceanography.

Over the last four decades, many simulation communities have interpreted the accuracy assessment of computational simulation to be composed of two distinct activities: *verification* and *validation*. Most communities consider that verification deals with numerical solution accuracy and software correctness issues. A major exception to this perspective is the atmospheric sciences community. Their perspective will be discussed in Sect. 3.2: Foundational issues in simulation credibility. *Validation*, sometimes referred to as external model validation, has a wide range of interpretations; sometimes even contradictory interpretations between communities. For example, the computer software community has very different interpretations of the terms *verification* and *validation,* compared to many simulation communities. Section 3.2 discusses some of these diverse interpretations and the widespread confusion and conflict in terminology of verification and validation.

The root cause of the diverse perspectives of the terms verification and validation is most likely that they are both closely related to the concepts of truth and correctness. These concepts are certainly appropriate from a philosophical perspective, but these concepts are ineffective and unproductive from a practical perspective. Truth and correctness are absolute concepts; seldom of importance in the use of most simulations, such as design of technological systems and governmental policy questions. The vast majority of simulations occur in a setting where two practical issues are central. The first issue is the *adequacy* of the estimated accuracy of the simulation for the goal or task at hand. For example, "Is the accuracy of the simulation adequate to meet the requirements to answer a specific set of questions, such as a system design, safety, or performance issue?" The second issue is the *comprehensiveness* or *completeness* of the simulation. For example, "Does the simulation address the relevant issues for its intended purpose over a specified range of conditions, such as varying operating conditions, failure modes, or for a specified time into the future?" In a practical setting, both adequacy and comprehensiveness are fundamentally impacted by the issue of uncertainties in the inputs to the simulation, the uncertainty of the simulation results, the adequacy requirements, and the comprehensiveness requirements. Most fields of simulation have not tried to specifically separate all of these aspects. Section 3.2 discusses these issues and how many in the engineering community have not only separated these issues, but have also segregated the issues of model calibration and predictive capability.

Section 3.3 addresses the issue of verification, in the sense of numerical accuracy and software correctness, with a brief description on the topics of code verification and solution verification. Code verification deals with (a) testing whether numerical algorithms used in obtaining a numerical solution are correct and (b) assessing software quality assurance issues. Solution verification, also referred to as calculation verification, deals with the estimation of numerical error of a given discretized solution to a set of partial differential or integral equations. In many computational simulations of interest, the mathematical model describing the process or system of interest is given as a set of partial differential or integral equations.

Section 3.4 deals with model validation, as well as two closely related topics: model calibration and model predictive capability. *Model calibration* is defined as the use of empirical measurements to optimize or update parameters in the mathematical model. *Predictive capability* is the ability of the mathematical model to foretell the response of a system, including relevant uncertainties, for conditions which the system has never experienced. Whether we are concerned with a time-dependent system or a steady-state system, a prediction for the system means that there are no experimental data/observations available for precisely the same conditions that we are interested in at present. Some writers, such as (Silver 2012), find it useful to make a distinction between *prediction* and *forecast*, but that distinction is not useful here. It should be noted that most simulation results that are reported are of the retrodiction or postdiction variety, meaning that the experimental result or observation of the system under consideration (or something very similar) is available to the computational analyst beforehand.

The final section provides some concluding remarks dealing with the current state of code verification, solution verification, validation, and predictive capability in computational simulation. Observations are given concerning the general unreliability of simulation results, whether they are published in journals or provided to decision-makers, and what should be done to improve this unfortunate situation.

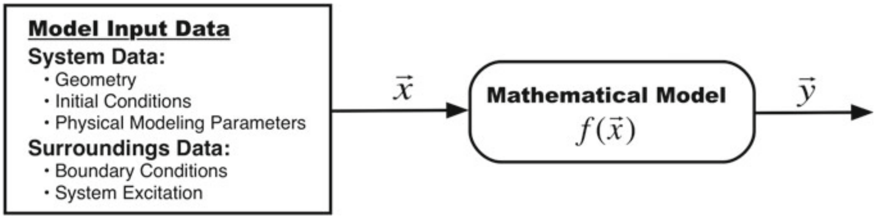


Fig. 3.1 Mapping of input data $\vec{x}$ through the mathematical model of the system of interest to obtain simulation results $\vec{y}$ Adapted from Oberkampf and Roy (2010)

3.2 Foundational Issues in Simulation Credibility

The process of the computational simulation of most physical systems can be thought of as a mathematical mapping of input data, through a mathematical model, to obtain the output data, as shown in Fig. 3.1. The input data can commonly be divided into two categories: *system input data* and *surroundings input data*. The system data can be further divided into geometric characteristics of the system, initial state conditions, and the physical modeling parameters of the system. Examples of physical modeling parameters are thermal conductivity, fluid viscosity, modulus of elasticity, and refractive index. The surrounding input data can be divided into boundary conditions and system excitation. System excitation refers to how the surroundings affect the system *other than* through the boundary conditions. Some examples of system excitation are gravity, electric fields (or magnetic fields) acting on a system, and vibrational excitation of the system from the surroundings; sometimes referred to as body forces and moments. The surrounding input data provide independent information on how the surroundings influence, and sometimes interact with, the system of interest.

All of the input data is represented as the vector $\vec{x}$, which can have a length of hundreds to thousands of elements. The mathematical model $f(\vec{x})$ is commonly given by a set of partial differential equations (PDEs) or a set of integrodifferential equations. For simplicity, both types of equations are referred to as PDEs. PDEs are presumed to be of such complexity that their solutions can only be obtained by an approximate numerical solution on a digital computer. The output data from the mathematical model is referred to as the system response quantities (SRQs) and represented by the vector $\vec{y}$. The SRQs can be local quantities within the solution domain on the PDEs, such as dependent variables of the PDEs. They can also be global quantities, such as an integral quantity, e.g., lift and drag of an aircraft, total strain energy within a structure, or net heat flux out of a system.

Simulation credibility deals with the assessment of the accuracy of $\vec{y}$ with respect to some true value or referent, whether or not it is knowable or measurable. As a result, there are three central issues in the quantitative assessment of simulation credibility: (a) how is $\vec{y}$ compared to the true value?, (b) what is regarded as the true value?, and (c) what is the requirement for the simulation result to be considered *credible* or *adequate* by the user or customer of the simulation? There are a multitude of both technical and practical complexities involved in comparison assessment, defining the true value, and specifying adequacy of the simulation. Most fields of simulation have developed their own approaches for dealing with some of these issues, but few have clearly separated and dealt distinctly with each of them. Although it is beyond the scope of this chapter to address how various fields address these issues, a few comments on each are included.

Quantitative comparison of simulation results with the true value can take many forms. In some simulation communities, this field of research is now referred to as the construction of *validation metrics*. Section 3.4: validation, calibration, and prediction discuss this topic, as well as other chapters in this volume (e.g., Chap. 13 by Marks and Chap. 18 by Saam in this volume). The central issue in the construction of quantitative comparison measures, or validation metrics, is "What features of the simulation result relative to the true value are important to the customer of the simulation?" Stated differently, the method of comparison of the simulation result with the referent should directly incorporate what features are important to the user of the information produced by the simulation. Some examples of simulation results that can be assessed with regard to the true value in increasing level of detail and difficulty are (a) global quantities for a boundary value problem, (b) global quantities for an initial-boundary value problem, (c) specified statistical features of the response, and (d) the time-dependent response of a local quantity in the solution domain. A major complexity that enters into the issue of comparison approaches is uncertainty in either or both the simulation result and the true value.

There are normally *many* true values because each true value (or function) is true for a specific situation of the physical system. As a result, the true value or referent is more complex than many people realize. For example, even assuming that a set of PDEs has a unique solution for each initial/boundary value problem, there is an infinite set of true solutions, one for each set of input data. Note that the infinite set occurs even if input data is deterministic, i.e., the input data, the mathematical model, and the solution are not even stochastic. If the true value is defined as the exact solution to the mathematical model, then the activity of comparison is referred to in most communities as *verification*. This type of accuracy assessment is entirely focused on the accuracy of the numerical solution of the PDEs and has *nothing* to do with ability of the mathematical model to replicate the physical system of interest. Some engineering communities have emphasized the importance of this activity, which is discussed in some detail in Sect. 3.3: verification activities, as well as other chapters in this volume. If the true value is defined as experimental measurements of the system, as discussed in Sect. 3.4 of this chapter, then most communities refer to this activity as *validation*. However, the atmospheric sciences community refers to this concept of validation as *forecast verification*. Jolliffe and Stephenson (2011) define forecast verification as "the process of summarizing and assessing the overall forecast quality of previous sets of forecasts."

If the true value is defined as measurements from an experiment, the uncertainty of the measurement of the true value is a never-ending issue. When measurement results are available, or any type of empirical observations of the system is available, these results can be used to assess the accuracy of the model. A more common simulation situation is when the input data to the model are adjusted to obtain improved agreement of the simulation with the empirical results. Adjusting the model or the input data, given observations of the system, is referred to as solution of the *inverse* problem. This topic, *model calibration*, is further discussed in Sect. 3.4 and in Chap. 41 by Frisch in this volume.

If the true value is an unknown quantity then, *no quantitative* assessment of accuracy can be obtained. This situation is the most common situation in the assessment of simulation credibility because most simulations address, at least in part, situations that have never been empirically observed. Similar situations to the simulation result may have occurred in recorded history, but not the specific situation that is simulated. Some examples include failure of a bridge design, impact of a new drug on human organs, long-term underground storage of high-level nuclear material, and the impact of CO_2 on global warming. When the true value is not known, the accuracy of the simulation is assessed in qualitative and subjective terms, such as face validation, believability, plausibility, and reasonableness. Even if comparisons are made between results of various competing models, this is still a qualitative assessment of credibility in the sense that there is no assurance that any of the models are correct.

The requirement for the simulation to be considered credible or *fit-for-purpose* for the intended use of the simulation is rarely addressed in most communities. This requirement must be judged relative to the accuracy of the simulation relative to the true value, even if the true value is also unknown or uncertain. In addition, the adequacy requirement should be set by the customer of the simulation; *not* the simulation analyst. In practice, if the customer does state a requirement, the accuracy requirement is not commonly satisfied initially. The customer may then loosen the requirement or use the simulation results in a different way than originally intended. The customer may also provide additional time and funding to improve the simulation results, obtain new experimental measurements, or obtain improved estimates of the true value.

Certain engineering and hydrology communities (Beven 2002; Refsgaard and Henriksen 2004; Rykiel 1996) have specifically separated the issues of accuracy assessment, definition of the true value, and simulation accuracy requirements. If the true value is unknown, then the adequacy requirement is entirely focused on the issues of uncertainty estimation in both the simulation result and the true value. Thus, as opposed to attaining the truth, the focus shifts completely to uncertainty estimation, physical-based principles, statistical inference, and adequacy for intended use of the simulation. Two engineering societies in the United States, the American Institute of Aeronautics and the American Society of Mechanical Engineers, have codified this practical framework into engineering standards documents (AIAA 1998; ASME 2006, 2009, 2012).

3.3 Verification Activities

Verification activities can be split into code verification and solution verification activities.

3.3.1 Code Verification

Code verification deals with two separate technical issues: (a) software quality assurance (SQA) practices and (b) testing the numerical algorithms used in obtaining a numerical solution (Oberkampf and Roy 2010; Roache 2009, 1972). SQA emphasizes determining whether or not the computational code is correctly programmed, e.g., coding errors have been removed and the code produces repeatable results on specified computer hardware and system software. The system software includes the computer operating system, compilers, libraries, data communications software, and data storage software. SQA focuses on the code as a software product and assessing its reliability and robustness from the perspective of software engineering. SQA procedures are particularly important during software development and software modification. Numerical algorithm verification addresses the reliability of the implementation of the numerical algorithms that affect the numerical convergence characteristics of the code. In other words, the numerical algorithm verification process focuses on accuracy and reliability of the numerical algorithms to approximate the solution to the original mathematical model of the system of interest, e.g., the PDE or integro-differential equations.

Numerical algorithm verification is fundamentally *empirical*, i.e., the activities are based on observations of performance, consistency, and convergence characteristics of the code for specific test cases. The goal of numerical algorithm verification is to accumulate sufficient evidence for a wide range of specific solutions to demonstrate that the numerical algorithms in the code are implemented correctly and functioning as intended. The most common technique for testing the numerical algorithms is to compare the numerical solutions obtained with the correct answer. (This technique, as well as others, is discussed in Knupp and Salari 2002; Oberkampf and Roy 2010; Oberkampf and Trucano 2008; Roache 2009.) In this sense, numerical algorithm testing is an error *evaluation* procedure as opposed to an error *estimation* procedure. The correct answer to the mathematical model is obtained from analytical solutions, manufactured solutions, or very accurate (i.e., benchmark) numerical solutions. As a result, the testing is case specific, meaning that the testing can only be exercised in a relatively small number of specialized cases. Therefore, these cases assume very important roles in code verification and should be carefully formalized and well documented in test plans for both the computational analyst and the customer of the simulation.

The most challenging aspect of code verification is the generation of accurate solutions for a wide variety of mathematical, usually nonlinear, models. Oberkampf and Trucano (2008) defined four categories of highly accurate solutions in terms of strong-sense benchmarks for code verification. Many researchers recognize that analytical solutions constructed by the method of manufactured solutions (MMS) can be the most powerful and effective approaches to generating benchmark solutions. Manufactured solutions are solutions to a class of mathematical models in the sense that the class is not restricted to any specific set of initial or boundary conditions. This results in significantly larger coverage of solution features as compared to traditional analytical solutions. An additional feature of manufactured solutions is that they are not restricted to mathematical models of physical reality. This occurs because the assumption of a manufactured solution happens at the very *beginning* of the process, as opposed to the traditional process of obtaining analytical solutions. Using this assumed solution, the right side of the original PDE is modified, resulting in nonphysical solutions. As long as these nonphysical solutions do not cause mathematical difficulties, such as negative temperatures or density, then these solutions provide an effective test. Some researchers, who are unfamiliar with MMS, view this shift to nonphysical solutions as making the method irrelevant to verification of their software. This view is erroneous because the purpose of code verification is testing the software and the numerical algorithms, *not* testing of the fidelity of the physical principles in the models. Thus, code verification is focused on testing of the software and the numerical algorithms that are used to solve a general set of PDEs. In the case of MMS, the right side of the original PDE is modified to achieve a much broader class of analytical solutions to a modified PDE that is closely related to the original mathematical model of the system of interest. (See also Chap. 11 by Rider and Chap. 12 by Roache in this volume.)

Trucano et al. (2005) noted that when solving complex systems of PDEs, a computational analyst finds it virtually impossible to decouple the distinct problems of mathematical correctness, algorithm correctness, computer hardware and operating system features, and problem-specific software implementation correctness. For instance, algorithms often represent non-rigorous mappings of the mathematical model to the underlying discrete equations. Whether such algorithms produce correct solutions to the PDEs cannot be assessed without executing the code on specific problems; the execution of the code is, in turn, coupled to the hardware and software implementation. One consequence of coupling mathematics, algorithms, computer hardware, and the software implementation is that the source of a numerical inaccuracy cannot always be easily identified. As simulations become more complex in terms of coupled physical processes and stochastic systems, the overlap between mathematical model complexities, discrete mathematics, SQA, and massively parallel computing will become more intertwined and much harder to dissect and understand.

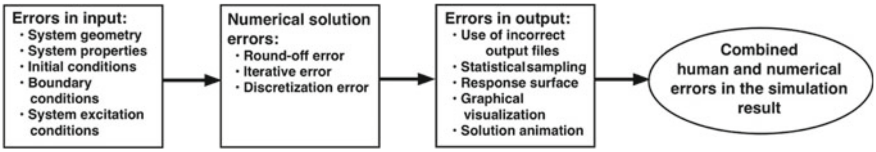


Fig. 3.2 Three types of error sources in solution verification

3.3.2 *Solution Verification*

Solution verification, also referred to as calculation verification or numerical error estimation, deals with the estimation of the numerical error of a given discretized solution to the PDEs. The goal of solution verification is distinctly different from that in code verification. That is, the primary goal of solution verification is to estimate the numerical accuracy of a discrete solution to the mathematical model of the system of interest for specific input data. Solution verification is dependent on the quality and completeness of code verification. Code verification could be described as the detailed exploration of solutions to a set of specialized mathematical models, whereas solution verification is the identification and estimation of numerical errors that are site specific to problems of interest to the computational analyst and the simulation customer.

Solution verification attempts to identify and quantify three sources of errors that can occur in the exercise of a simulation code (Fig. 3.2):

- Errors, blunders, or mistakes made by the computational analysts in preparation of the input data for the simulation code;
- Numerical solution errors resulting from computing the discretized solution of the mathematical model; and
- Numerical errors, blunders, or mistakes made by the computational analysts in any processing or presentation of the output data that is produced by the simulation code. (See also Chap. 5 by Roy in this volume.)

The first error source *excludes* errors or approximations made in the formulation or construction of the mathematical model. This important source of error will be discussed in Sect. 3.3: validation, calibration, and prediction. Human errors and unintentional misuse of simulation software packages are key components in the first and third sources of error. The human component is rarely discussed or addressed in simulation results, except in simulations conducted for high-consequence systems. Human errors can be difficult to detect, even in relatively small-scale analyses, because they are usually procedural or simulation-process errors. Fields that deal with these types of errors, such as human reliability assessment, have found that procedural or independent data checking methods are the most effective for identification of these types of errors. For example, if a solid mechanics simulation contains tens of computer-aided design (CAD) files, perhaps hundreds of different material models, and thousands of Monte Carlo simulation samples, then human errors, even by the most experienced and careful practitioners, commonly occur. (For a more detailed discussion of these error sources, see Neumann 1995; Oberkampf and Roy 2010; Reason 1997, 2008.)

The second error source shown in Fig. 3.2, numerical solution errors, includes errors that occur due to the computation of the discrete solution to the mathematical model on a digital computer. With modern computer hardware using a 64-bit word length, computer round-off errors rarely are a significant source of error. Iterative solution errors are those caused by incomplete iterative convergence of the numerical methods used to solve the discrete form of the equations of the mathematical model. For example, if one has a nonlinear boundary value problem as the mathematical model for the system of interest, then the iterative solution error is due to a nonzero residual of the discrete equations for a given spatial mesh resolution. Several reliable methods are available for estimating and controlling the iterative solution error. (See Duggirala et al. 2008; Ferziger and Peric 2002; Golub and Van Loan 2013 for a more detailed discussion.)

The most troublesome error source in the second block in Fig. 3.2 is the discretization error. This is troublesome not only because it is typically the largest source of error in solution verification, but also because it is difficult to estimate in most practical problems of interest. Discretization error is the numerical error due to the discrete approximation of the mathematical model for the specific set of input data characterizing the system of interest. This error exists regardless of what numerical method is used, e.g., finite element methods, finite difference methods, spectral methods, or mesh-free methods.

For the most common numerical methods, such as finite element and finite difference methods, the key issue is estimating and controlling the error due to spatial and temporal discretization of PDEs. The two basic approaches for estimating the discretization error are a priori and a posteriori error estimation techniques (Ainsworth and Oden 2000; Babuska and Strouboulis 2001; Oberkampf and Roy 2010; Roache 2009; Verfurth 2013). An a priori technique only uses information about the numerical algorithm that approximates the given PDE and the given initial conditions (ICs) and boundary conditions (BCs). An a posteriori approach can use all the a priori information as well as the computational results from previously obtained numerical solutions, e.g., solutions using different mesh resolutions or solutions using different order-of-accuracy methods. The last two decades have clearly shown that the only way to achieve a useful quantitative estimate of discretization error in practical cases for nonlinear PDEs is by using a posteriori error estimates.

A posteriori error estimation methods can be categorized as either higher order estimators (Type I) or residual-based estimators (Type II) (Roy 2010). The Type I methods involve post-processing of the solution (or multiple solutions) and include Richardson extrapolation, order extrapolation, and recovery methods for finite elements. The Type II methods employ additional information about the problem being solved and include discretization error transport equations, defect correction methods, adjoint methods, and implicit/explicit residual methods for finite elements. Regardless of the approach used for estimating the discretization error, the reliability of the estimate depends on the solutions being in or near the asymptotic mesh convergence region. Since complexity of the physical principles in the mathematical models has increased significantly over the last few decades, attaining the asymptotic convergence region or determining if you have attained it is extremely difficult. The spatial and temporal discretization resolution required to attain asymptotic convergence is computationally unaffordable on most complex problems, even on massively parallel computers. (For a more detailed discussion of solution verification, see Chap. 11 by Rider in this volume.)

3.4 Validation, Calibration, and Prediction

3.4.1 Model Validation

3.4.1.1 The Restrictive Concept of Model Validation

This section discusses two of the three issues presented in Sect. 3.2 that deal with simulation credibility: (a) the comparison of simulation results with the true value and (b) defining the true value. The restrictive concept of model validation is a perspective that has been developed, debated, and tested over the last two decades and has been found to be extremely useful in engineering and the natural sciences. The perspective has been adopted by various engineering societies in the US (AIAA 1998; ASME 2006, 2009, 2012) and some simulation communities in Europe. The hydrological and atmospheric sciences communities have also struggled with the concepts of validation, calibration, and predictive capability and have come to similar perspectives to that presented here (Anderson and Bates 2001; Beven 2002; Chiles and Delfiner 1999; Jolliffe and Stephenson 2011; Refsgaard and Henriksen 2004; Rykiel 1996; Wilks 2011).

Model validation, as defined here, is focused on the assessment of the error due to the approximations and assumptions made in the formulation of the conceptual and mathematical models. The conceptual model comprises all relevant information concerning the system of interest, modeling assumptions, simplifications, and approximations regarding the processes of interest within the system, as well as specification of the interaction of the surroundings with the system of interest. The mathematical model is the quantitative embodiment of the mental abstraction given in the conceptual model. The procedure for assessing the conceptual and mathematical modeling errors relies on the quantitative comparison of simulation results with experimental measurements from specially designed and executed experiments (see Fig. 3.3). Model validation, as depicted in Fig. 3.3, is usually referred to as the restricted concept of validation.

Figure 3.3 depicts a *model validation experiment* that has several key features that are different from other types of experiments (Aeschliman and Oberkampf 1998; AIAA 1998; ASME 2006; Marvin 1995; Oberkampf and Aeschliman 1992; Oberkampf and Roy 2010). A validation experiment can be conducted on many

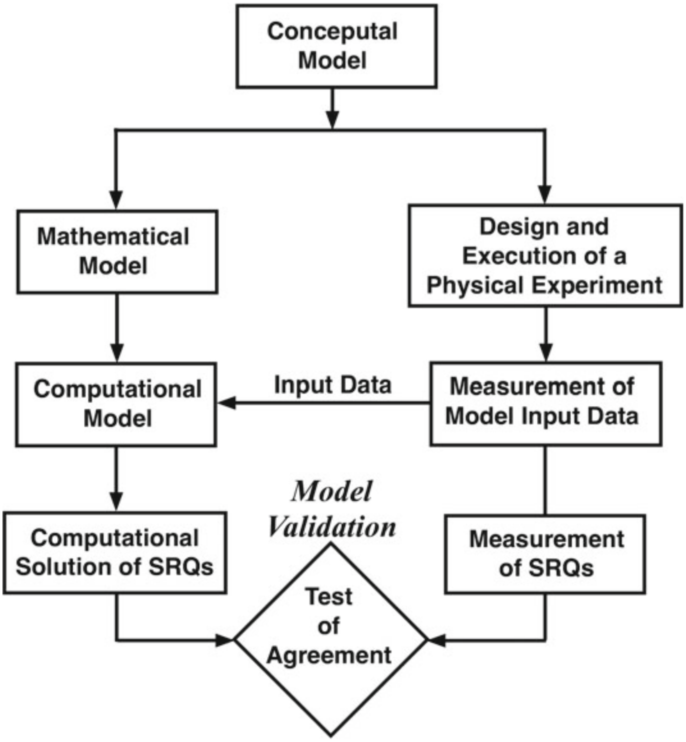


Fig. 3.3 Restricted concept of model validation

accessible systems of interest. For engineering systems, validation experiments can be conducted at the system level, subsystem level, component level, or down to the level of single physical processes. For example, the hierarchy of experiments that are conducted on a gas turbine engine is conducted on several different levels. These are listed from the top-level system to the simplest level: an experiment conducted on the complete engine at actual operating conditions; an experiment conducted on the compressor or the turbine sections; an experiment conducted on a single compressor blade or turbine blade; and a heat transfer experiment conducted on a flat surface. For large-scale natural systems, such as earth sciences (e.g., climatology, geology, and oceanography) and astronomy (e.g., cosmology, astrophysics, and planetary sciences), the concept of conducting validation experiments is unworkable because all of the input data needed for the model cannot be measured. Experimental measurements obtained for these types of systems would be referred to as *model building experiments* or *model calibration experiments*, both forms of model retrodiction that are discussed later.

Two important points regarding Fig. 3.3 stress the herein defined concept of model validation. First, the diagram assumes that code verification testing and solution verification activities have been adequately completed before model validation. Second, the computational analyst is attempting to predict the specific experiment conducted in the validation experiment. As a result, the experimentalist is expected to measure the model input data needed by the computational analyst to produce a solution. This expectation placed on the experimentalist is clearly *not* a traditional responsibility in experimental activities. For the experimentalist to measure the needed model input data, the experimentalist must be aware of what may be needed by computational modelers. An active partnership between the experimentalist and the analyst during the design and execution of the experiment is much preferred.

Recalling the depiction of the simulation process is shown in Fig. 3.1, the mapping of the input data through the mathematical model, $\mathscr{D}$, to obtain the system response quantity of interest, SRQ, can be written as

$$\mathscr{D}(\mathcal{G}, \mathcal{IC}, \mathcal{MP}, \mathcal{BC}, \mathcal{SE}) \rightarrow SRQ \tag{3.1}$$

$\mathcal{G}$ is the geometry of the system, $\mathcal{IC}$ are the initial conditions of the system, $\mathcal{MP}$ is the model parameters of the system, $\mathcal{BC}$ is the boundary conditions imposed by the surroundings, and $\mathcal{SE}$ is the system excitation imposed by the surroundings. If $\mathcal{G}$, $\mathcal{IC}$, $\mathcal{MP}$, $\mathcal{BC}$, and $\mathcal{SE}$ are left unspecified, then Eq. (3.1) is referred to as the weak definition of a model (Leijnse and Hassanizadeh 1994). A weak model definition *cannot* be validated because it is a general statement of the mapping of the model input to the model output, i.e., it lacks the specificity needed to obtain a solution to the model. If $\mathcal{G}, \mathcal{IC}, \mathcal{MP}, \mathcal{BC}$, and $\mathcal{SE}$ are all measured in an experiment, as indicated in Fig. 3.3, then the mathematical model $\mathscr{D}$ is referred to as the strong definition of a model (Leijnse and Hassanizadeh 1994). Measuring all of the input data means that parameters that exist in the model are also measured, for example, system material parameters as well as their variability. As indicated in Fig. 3.3, since the SRQ of interest is measured in the validation experiment, then the *model form error*, *MFE*, can be defined. Stated differently, given the measurement of the model input data, the SRQ, and their uncertainty, the only remaining error that exists is that is due to the conceptual and mathematical models. The model form error, also called the *model discrepancy error*, can be written as

$$\|\mathscr{D}(\mathcal{G}_E, \mathcal{IC}_E, \mathcal{MP}_E, \mathcal{BC}_E, \mathcal{SE}_E) - SRQ_E\| = MFE_E \tag{3.2}$$

Subscript E indicates a measurement of the quantity is made during the experiment and $\|\cdot\|$ indicates some type of norm that is appropriate for the type of functions involved. Furthermore, the subscript E on *MFE* indicates that the model form error is *site specific* for the conditions of the experiment. From a philosophy of science perspective, Fig. 3.3 and Eq. (3.2) state that based on empirical observations, the computational analyst can only make a very restricted evaluation of the accuracy of a mathematical model. Confidence in a broader sense of validity of a model can only be inferred based on the strength or difficulty of the test of the model and the accumulation of evidence over a wide range of conditions for which the model is tested.

If the input data to the model given in Eq. (3.2) is deterministic, then the simulation result will also be deterministic, e.g., a scalar value or vector field. However, if the input data is nondeterministic, i.e., stochastic, then the simulation should be stochastic, e.g., an ensemble of simulations is used to characterize the input data and the corresponding system response. Note that even if the input data were deterministic, MFE_E will be nondeterministic because of the existence of experimental measurement uncertainty in the SRQ_E.

The formal concept of a difference measure as indicated in Eq. (3.2) and at the bottom of Fig. 3.3 is a relatively new concept. The equation is entirely focused on quantitatively measuring the magnitude of the difference between simulation and experiment; *not* on the issue of "good or bad" or an "adequate or inadequate" agreement. Furthermore, it is a mathematical measure that can be intuitively interpreted as a distance between simulation and experiment, especially when given in terms of the dimensional units of the *SRQ*. The measure can be quantified at a point in space or time or over any domain of interest by use of an appropriate norm. The field of research that deals with the estimation of *MFE* is now referred to as *validation metrics*, a term coined by Trucano et al. (2001). The concept of model parameter estimation or optimization in order to minimize the magnitude of *MFE* in Eq. (3.2) is a much older concept. These concepts, and their relationship to the estimation of *MFE*, are discussed in the next section.

Equation (3.2) is entirely focused on an assessment of physical modeling errors, as opposed to a mixture of physical and numerical errors. If extensive code and solution verification have not been completed and documented, then the magnitude of *MFE* can be misleading. Since there are so many opportunities for numerical and human errors in simulation, the computational analyst can easily have the cancellation of the impact of these errors in various portions of the mathematical model and sub-models such that they obtain a small value of *MFE*. For example, a commonly occurring situation is that the modeler finds close agreement with experimental measurements and declares the model validated. Later, either the original modeler or another researcher who attempts to reproduce the earlier comparisons uses a more refined mesh or refined time step and discovers that now the model disagrees significantly with the experiment. With the ubiquitous use of numerical simulation in science and engineering around the world, and the generally poor quality of verification activities, this situation is unfortunately widespread.

3.4.1.2 Approaches to Validation Metrics

A wide range of approaches has been developed to implement Eq. (3.2). (In this volume, see Chap. 7 by Beisbart, Chap. 13 by Marks, Chap. 19 by Robinson, Chap. 20 by Jiang et al. and Chap. 23 by Schlünzen.) The three most common approaches for estimating *MFE* are statistical hypothesis testing, Bayesian estimation, and use of the area metric. In hypothesis testing, the accuracy assessment is formulated as a decision problem of whether or not the model's predictions are consistent with the available empirical data. The consistency between the model and the data is characterized as a probability, with low probability values denoting a mismatch of such magnitude that it is unlikely to have occurred by chance. Hypothesis testing is not well suited to the task of model validation because the primary goal of hypothesis testing is to identify statements for which there is compelling evidence of truth or falsehood. But, this goal is different from that of model validation, which is more pragmatic and more focused on the *degree to which* the model and experimental measurements concur. Two practical difficulties arise with the interpretation of the hypothesis test, or statistical significance, result. First, how should a model developer, project manager, or decision-maker interpret the result? It is not clear how to interpret a probability result as an accuracy measure. For example, how does one interpret the result, "There is an 80% probability that the model agrees with the experimental measurements?" The natural perspective of a design engineer, project manager, or decision-maker is to ask, "What is the physically interpretable difference between the model and the experiment and how does that compare with what is an acceptable tolerance for a particular application or decision?" Second, no matter what level of accuracy is specified, the model can be proven false at a given level of significance if more experimental data are obtained. That is, any model can be proven false, given enough data. (For an overview of hypothesis testing, see Lehmann and Romano 2005; Wellek 2010; Ziliak and McCloskey 2008.)

Bayesian model validation has received the greatest attention during the last two decades. Bayesian estimation is mathematically sophisticated and can be applied in very wide range of applications. However, it is extremely demanding in terms of the number of mathematical model evaluations, e.g., solutions of the PDEs that are needed for the process. Bayesian estimation could be described as a procedure to improve model predictions by updating uncertain model parameters when experimental data for the system of interest are available. The procedure to update the probability distributions of the model parameters uses Bayes' equation, the simulation result for the system of interest, and the available experimental data for the system. As a result, Bayesian estimation is a parameter optimization procedure operating under the constraint of the particular form of the mathematical model being used and the experimental data available. As the amount of experimental data increases, the probability distributions of each of the model parameters converge to probability distributions that allow the model result to yield the most accurate reproduction of the experimental data, conditional on the mathematical model. Note that the claim of convergence of the parameter distributions also relies on the assumption of a sufficient number of mathematical model evaluations. For computationally expensive solutions, Bayesian estimation resorts to using surrogate or response surface models to approximate the actual model evaluations.

Bayesian estimation has taken a much more important role in model validation since the transformative work of Kennedy and O'Hagan (2001). Kennedy and O'Hagan devised a statistical method to quantify what they referred to as the model inadequacy or discrepancy as part of the Bayesian model parameter updating procedure. They state, "we define a model inadequacy to be the difference between the true mean value of the real world process and the code output at the true values of the inputs." Their model discrepancy term, also called model *bias*, is closely related to the model form error defined in Eq. (3.2). Most Bayesian model validation procedures use a form of Bayesian hypothesis testing by incorporating the use of Bayes factor (or factors). The factor is interpreted as the ratio of the relative likelihood of the null hypothesis that the experimental data supports the model predictions divided by the alternative hypothesis that the data does not support the prediction. Bayes factors aid the decision regarding the acceptance or rejection of the null hypothesis test. As pointed out by some researchers (Ferson et al. 2008; Liu et al. 2011; Oberkampf and Barone 2006; Oberkampf and Roy 2010), however, it is not clear how a designer or policy-maker should interpret this factor in a decision context concerning how much error is incurred by using the model. For an overview of Bayesian model validation, see Babuska et al. (2008), Bayarri et al. (2007), Chen et al. (2008), Higdon et al. (2008), Li et al. (2014, 2016); Liu et al. (2008, 2011), McFarland and Mahadevan (2008), O'Hagan (2006), Rougier (2007), Trucano et al. (2006), Wang et al. (2009), Wilks (2011), as well as Chap. 7 by Beisbart and Chap. 20 by Jiang et al. in this volume.

The area validation metric is defined as the mismatch between the simulation and the experimental measurements as quantified by the area between the cumulative distributions functions (CDF) of each (Ferson et al. 2008). As a result, it is a straightforward difference measure that quantifies the left side of Eq. (3.2) directly in terms of the SRQ of interest. Stated differently, the area metric is a measure of distance between the stochastic simulation result and the uncertain experimental measurements in terms of the dimensional units of the particular SRQ of interest. As a result, the area metric is much more physically understandable to simulation analysts, designers, and decision-makers. The area metric uses the Minkowski L_1 metric to measure the difference between the two CDFs,

$$d(F, S_n) = \int_{-\infty}^{\infty} |F(z) - S_n(z)| dz, \tag{3.3}$$

where F is the empirical distribution function (EDF) of the simulation and S_n is the EDF of the experimental measurements. The EDFs are used, as opposed to the CDFs, because then the integral can be computed regardless of the number of samples of simulation and experimental data. Three advantageous features of the area metric should be noted. First, the metric can be computed regardless of how much data are available from the experiment or from the simulation. For example, if the simulation is computationally expensive such that only several realizations can be computed using a Monte Carlo simulation, the area metric can still be computed without resorting to approximations of the model response. Second, the metric generalizes deterministic comparisons between scalar values that have no uncertainty. That is, if the prediction and the observation are both scalar point values, the area metric is equal to their difference. Third, the area metric is the *evidence for the mismatch* between the simulation and the measurements, instead of the evidence for agreement. Thus, if there are only single realizations for both the simulation and the experiment, and the area metric is large, it does *not* necessarily mean that there is a large disagreement between simulation and experiment. This may mean that the difference between the two is entirely due to the random variability of each. As a result, the strength of the evidence for disagreement depends on the quantity of data available for both the simulation and the experiment. For a more detailed discussion of the area metric and how the concept has been extended to multivariate data, see Ferson and Oberkampf (2009), Ferson et al. (2008), Li et al. (2016), Liu et al. (2011), Oberkampf and Roy (2010), Voyles and Roy (2015) as well as Chap. 13 by Marks in this volume.

3.4.1.3 Difficulties in Model Validation

There are many situations where technical difficulties eliminate the possibility of comparing simulation results with model validation experiments, as depicted in Fig. 3.3. These difficulties can be generally grouped into inability to measure (a) the input data needed for the mathematical model and (b) the SRQs of interest. One example of this is the simulation of underground transport of pollutants or toxic chemicals, primarily due to water seepage through porous media. This requires knowledge of the porosity and permeability parameters of the subsurface material that appears as a random field in the PDEs. Although some characteristics of the subsurface material can be obtained by removing drill cores from the area of interest and conducting laboratory experiments, it is impossible to characterize the three-dimensional field over the entire underground volume. Another example from the field of solid dynamics is the simulation of the vibration of assembled structures, such as aircraft and automobiles, which requires knowledge of the mechanical stiffness and damping of every material connection in the entire structure. It is impossible, even on a single connection of two materials, to measure the local stiffness and damping *within* the surface connecting the two materials. The procedure that is used for vibrational simulation is to use a rudimentary model of the macro effects of the connection and then calibrate the parameters that exist within the sub-model of the material connection. As can be seen in both of these examples, this results in the requirement to solve the *inverse* problem. That is, given the measured response, and the given form of the mathematical model, the values of the model parameters to best reproduce the measured response are determined. The topic of model parameter calibration, which is actually a feature of model building, will be discussed in the next section.

The second common situation occurs when it is technically impossible to measure important SRQs of interest. An example of this situation is the process of hypervelocity impact of a particle with a target material. Typical experimental results are photographs of the impact crater or hole through a material after the impact. In some facilities, high-speed imaging of the penetration event is available. These data are useful in model validation, but they greatly limit the ability to quantitatively assess the accuracy of many SRQs predicted by the model. An example where it is conceptually impossible to obtain the needed experimental data for validation is in modeling physical phenomena with very long time scales or very large physical scales. Some examples are (a) long-term prediction of the effect of various contributors to global warming (b) response of the global environment to a large-scale atmospheric event, such as a volcanic eruption or the impact of a large asteroid, and (c) physical processes on planets, stars, and solar systems.

There are also many situations where it is impractical or not allowed to conduct validation experiments. Examples of impracticality include conducting validation experiments on the failure of a full-scale dam, major failure of a nuclear power reactor, and collapse of a skyscraper. Forbidden validation experiments include testing the physiological response of humans exposed to toxic chemicals or radiation, experimental drug testing that poses a high risk of major physical or mental impairment to humans, and hazardous or environmentally damaging weapon tests that are banned by international treaties. The most common method of partially assessing the predictive capability of a model for the situations discussed in this paragraph is to conduct experiments on subsystems or components of the system of interest, or to conduct experiments on surrogate systems, such as animals. It is important to note that even if an experiment is conducted on a system component or on a surrogate for the real system of interest, the concept of a model validation experiment depicted in Fig. 3.3 still applies. Therefore, the concept of a model validation experiment as discussed in this section applies *regardless* of the complexity or physical scale of the experimental system.

3.4.2 Calibration and Predictive Capability

3.4.2.1 Model Calibration and Bayesian Estimation

As discussed in Sect. 3.4.1, there is a strong conceptual distinction between model validation and model calibration. Model validation, as defined here, is the activity of quantitatively assessing model accuracy by way of comparison of simulation results with experimental measurements. Model calibration, or model tuning, is the activity of updating input parameters to the mathematical model such that improved agreement is achieved between simulation results and experimental measurements. Recall from Fig. 3.1 that input parameters can describe features of either the system of interest or the impact of the surroundings on the system. Usually, the input parameters are stochastic, e.g., given by a probability distribution, but they can also be deterministic scalars. A philosopher of science might ask, “Why is there such confusion and ambiguity concerning the terms *validation* and *calibration*, when the concepts seem so distinct?” One of the reasons is that computational researchers spend the majority of their efforts on improving and updating their models given experimental observations, as opposed to assessing the accuracy of their models relative to the observations. For example, as soon as any experimental observations are available, the researcher will *immediately* use the new data to improve the agreement of the simulation with the measurements. Improvement of the model, which is a type of retrodiction, can take many forms, such as rejection of the entire model and replacement with a better model, reformulation of the model assumptions so as to obtain better agreement with the observations, and replacement of some sub-models with other sub-models that produce better agreement with observations. However, by far the most common method of model improvement is optimization of the model parameters to obtain better agreement with the observations. The Bayesian framework is perfectly suited to this approach of model improvement.

Since the Bayesian approach is the dominant framework in model calibration, two fundamental issues are raised: (a) What is the scientific basis or defensibility for using Bayesian parameter estimation and model validation? and (b) Why is Bayesian estimation so effective and widely used in computational simulation? Concerning scientific defensibility, the Bayesian estimation of parameters can be justified for some situations, but *not* for others.

Bayesian updating, or some other parameter optimization procedure, is defensible for (a) physical modeling parameters that are *not* measurable outside of the context of the mathematical model of the system under consideration and (b) ad hoc model parameters. Ad hoc parameters are those that are introduced into models simply to provide a method for adjusting the results of the simulation to obtain improved agreement with empirical observations. Using the descriptive terminology of Lenhard and Winsberg (2010) for the evolutionary and adaptive nature of complex models, I would refer to these types of parameters as *kludges*. A third type of physical modeling parameter is one that is independently measurable *exclusive* of the context of the mathematical model of the system under consideration. Using Bayesian updating on these parameters, however, is *not* scientifically defensible. Essentially all models have all three types of parameters imbedded in them and modelers make no attempt to distinguish between them during parameter updating. As a result, model formulation/approximation errors, uncertainties in model input parameters, and numerical solution errors are all intertwined in the updating procedure. Because of this coupled nature, cancellation of errors and compensation for other errors and uncertainties in the updating process is inevitable. Furthermore, these parameters change every time new observational data become available, when other sub-models are changed, or when changes in mesh resolution occurs. According to Lenhard and Winsberg (2010, p. 257), “complex simulation models acquire an intrinsically historical character and show path-dependency. The choices that modelers and programmers make at time one about how to solve particular problems of implementation have effects on what options will be available for solving problems that arise at time two. And they will have effects on what strategies will succeed and fail.” Adjusting physical modeling parameters and ad hoc parameters is the lubricant that allows complex models to function adaptively.

Bayesian inference, which is a broader topic than Bayesian parameter estimation, is widely used and particularly effective in complex simulations. Although there have been hundreds of articles written on this topic over the last four decades, a few observations should be noted. For physical-based models, as opposed to other model construction approaches such as machine learning, Bayesian inference is quite effective for two reasons. First, it is able to update the stochastic parameters in some optimal sense, given the constraints of the physical-based model. For example, Bayesian inference is able to adjust probability distributions to obtain best agreement with available experimental data, given the constraints of conservation of mass, species, momentum, and energy. For relatively simple physical systems with few free parameters, specifically physical modeling parameters and ad hoc parameters, the impact of the physical constraints on Bayesian updating of the parameters is rather restrictive. For complex systems with hundreds or thousands of free parameters, however, there is *great* flexibility to optimize the stochastic parameters. There is commonly weak physical-based justification for this optimization and the resulting parameter correlation structure. (In this volume, see the chapter by Seibert et al.)

Second, a strong argument can be made that Bayesian estimation produces the optimal posterior probability density function (PDF) for the SRQs of interest, given as follows:

- The available empirical evidence;
- The constraint of the mathematical model;
- The assumed priors for the free parameters; and
- The ability to compute, or approximate, a very large number of solutions to the mathematical model.

According to an interview with James Berger, Bayesian inference is very widely used today (Wolpert 2004). Prof. Berger believes that widely available, high-performance computing and general-purpose software packages are two of the key reasons that have allowed technical analysts in essentially every field of physical, life, and social sciences to apply Bayesian inference. Additionally, Berger stresses, "that Bayesian statistics allows one to ask any desired question and obtain an answer to that question." To emphasize his point, he comments that many times he hears the following comment from avowed non-Bayesians: "I do not believe in Bayesianism philosophically, but the Bayesian approach (with Markov chain Monte Carlo) was the only way I could analyze this complex problem, so I used the approach." Combining high-performance computing, widely available software packages, and extraordinary effectiveness in answering just about any question result in an unstoppable technology.

3.4.2.2 Predictive Capability and Uncertainty

Over the last two decades, the term *predictive capability* has developed to better capture the concept of foretelling an outcome. Predictive capability stresses the notion of forecasting system responses based on the physical fidelity in the mathematical model, as opposed to a model *emulating* or *imitating* the response of the system. Although the concept of forecasting an event or outcome has been recognized for centuries in science, the increased emphasis on the basis of the prediction has been appropriate because of the questions raised about the ability of highly calibrated models to foretell events not seen before. Note that when the term *predict* or *foretell* is used, it not only refers to time but also steady-state responses of a system that have not been observed before. Included in the category of highly calibrated models are a wide range of models, including machine learning, data mining, Bayesian networks, and artificial neural networks that can emulate or imitate system responses. To clarify the distinction made here, consider two examples. The first example has no predictive capability, but the model can emulate the system, whereas the second example has predictive capability. First, suppose one constructs a surrogate model (or metamodel or emulator) of a stochastic physical process that is known to be a function of (only) N input quantities. Further, assume that a large number of random observations of the process over a specified domain of the N-dimensional input space have been used to construct the surrogate model. The domain of the input space where observations have been made is referred to as the convex hull of the input space. If the number of observations is sufficiently large, one could show that the surrogate model would have essentially perfect ability to predict the likelihood of the SRQ inside the convex hull. This type of model is referred to as a *descriptive model* (Bossel 1994).

Next, consider a stochastic system where the physical-based mathematical model is assumed to be perfect. That is, it is assumed that all of the relevant physical phenomena in the system are fully captured by the physical assumptions in the construction of the mathematical model. Furthermore, all parameters in the mathematical model are perfectly known, whether from theoretical considerations or from experimental measurements. A computational analyst could argue that this model could perfectly predict the likelihood of the SRQ over *any range* of the input data, as long as no new physical processes occur, other than those captured in the mathematical model. This type of model is referred to as an *explanatory model* because it represents the systems' detailed, coupled physical structure and the interaction of components such that it can predict future system behavior, even under conditions that have never been seen before (Bossel 1994). The present-day use of the term *predictive capability* emphasizes this type of explanatory capability of the model. Stating the inverse, even highly calibrated models cannot be expected to have predictive capability or reliable information content *outside* of the domain over which the model has been calibrated.

Physical-based models in engineering and science are situated between the two examples just described. The nature of physical-based models is that there is not only weaker confidence in the explanatory capability of our physical-based models, but also limited data with which to calibrate the model parameters. Figure 3.4 shows a two-dimensional input space defined by the two parameters α and β, each one characterizing some feature of the system or the conditions imposed by the surroundings. The validation domain is shown as the region in which various validation experiments have been conducted, denoted by V, i.e., the convex hull of V. The application domain shows the region where we intend to use the model from an application perspective, i.e., where predictive capability is desired. As is typical of operating conditions of a system, the corners of the operating envelope are specified in terms of pairs of coordinates (α_i, β_i), $i = 1, 2, \ldots 5$. The relationship between the application domain and the validation domain shown in Fig. 3.4 indicates that the validation domain is a subset of the application domain. For most complex engineering systems, there are *no* model validation experiments of the complete system. There may be, however, operational data of the system or of a similar system for certain operating conditions, such as normal or slightly off-normal operating conditions. Since it is either impractical or impossible to measure all of the input data needed for model validation of

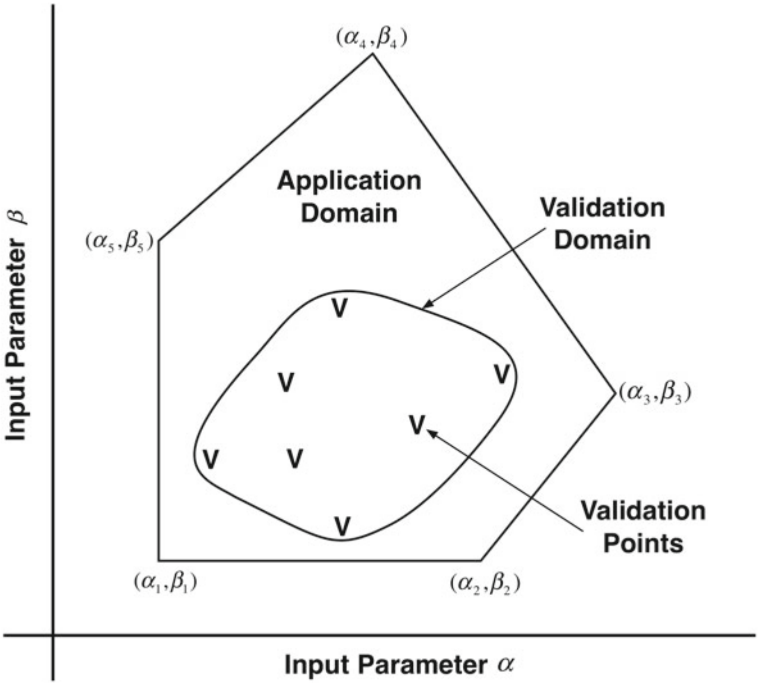


Fig. 3.4 Validation domain and application domain within a two-dimensional input space Adapted from Trucano et al. (2002)

a complex system, operational data for the system are used to calibrate model input parameters, as discussed earlier.

Presume that a validation metric result, i.e., an estimate of the model form error (*MFE*), has been computed at each of the conditions marked with a V in Fig. 3.4. A quantitative metric result for individual SRQs could be obtained from the model discrepancy term using either Bayesian estimation, or using the area validation metric, as discussed earlier. The boundary of the validation domain would represent the bounds where the model accuracy has been assessed. To account for model form error in a model prediction anywhere within the validation domain, some type of interpolation method could be used to estimate the model form uncertainty. The estimate of model form uncertainty could be added directly to the prediction of the SRQ where no data are available, in addition to the impact of model input uncertainty on the SRQ. Within the validation domain, the rigor of adding model form uncertainty to input uncertainty is dubious because they are two very different sources of uncertainty. Model form uncertainty is an epistemic uncertainty because it is lack of knowledge, specifically due to approximations and assumptions in the formulation of the model. But model input uncertainty is commonly thought of as aleatoric uncertainty and characterized as a random variable. For model predictions outside of the validation domain, the distinction between the two types of uncertainty becomes much more important because the information content in different characterizations of uncertainty can be dramatic. A critical source of uncertainty outside the validation domain is epistemic because the basis of the prediction is the fidelity of the physical principles in the mathematical model used to extrapolate to conditions that have not been already observed. Statistical extrapolation outside of the validation domain has *no* basis, but it is used routinely in computational simulation.

While there are different ways to classify uncertainty, this chapter uses the taxonomy prevalent in the risk assessment community which categorizes uncertainties according to their essence (Cullen and Frey 1999; Haimes 2009; Morgan and Henrion 1990; Vose 2008):

- *Aleatoric*—uncertainty due to inherent variation in a quantity that, given sufficient samples of the quantity, can be characterized via a probability density function; and
- *Epistemic*—uncertainty due to lack of knowledge by the individuals involved, be they modelers, analysts, or experimentalists.

Aleatoric uncertainty (also called irreducible uncertainty, stochastic uncertainty, or variability) is uncertainty due to inherent randomness and can occur among members of a population or due to spatial or temporal variations. Aleatoric uncertainty is generally characterized by either a probability density function (PDF) or a cumulative distribution function (CDF). A CDF is simply the integral of the PDF from minus infinity up to the value of interest. Examples of aleatoric uncertainty are: (a) unit-to-unit variability in the geometric and mass properties of a manufactured product, (b) spatial and temporal variability in the wind speed and direction near a nuclear power plant, (c) spatial variability of the roughness of a highway surface, and (d) spatial variability in the porosity and permeability of underground material. With a sufficiently large number of samples of each of these quantities, both the form of the CDF and the probability parameters describing the distribution of the population can be accurately determined, at least in concept.

Epistemic uncertainty (also called reducible uncertainty or ignorance uncertainty) is uncertainty that arises due to our lack of knowledge of the quantity. If knowledge is added through experimental measurements or observations, control of a quantity, expert opinion, or improved physical understanding, then the uncertainty can be reduced. If sufficient knowledge, which costs time and resources, is added, then the epistemic uncertainty can be reduced or eliminated, possibly resulting in only aleatoric uncertainty. Epistemic uncertainty is commonly represented as either a PDF representing the degree of personal belief of the analyst (as opposed to frequency of occurrence of an event) or as an interval-valued quantity with *no associated PDF*. If nothing is known about a quantity except its bounds, then an interval representation is the precise characterization of the state of knowledge. Some Bayesians claim that a uniform PDF over the range of the interval is an equivalent characterization. This is clearly *not* the case because a uniform distribution makes the further assertion that all values are equally likely over the range of the interval. Thus, an interval-valued quantity could be characterized as the infinite set of all possible PDFs over the interval. Obviously, the uniform PDF is only one member of the set. Some

Bayesians, who accept the difference in information content between an interval-valued quantity and a uniform PDF, claim that in real-world uncertainty analyses there is *never* a situation where there is such little information (Cooke 2004). This is patently false. A more accurate description of the Bayesian perspective is that a Bayesian statistician *only* allows lack of knowledge to be a PDF. The reason for this rigid stance is that an interval-valued quantity wrecks the Bayesian mathematical machinery, i.e., the usual Bayesian machinery cannot function.

Examples of epistemic uncertainty are model form uncertainty, poor understanding of complex physical processes or model input parameters, numerical solution error, a complex sequence of subsystem/component/human failures, and unintentional or intentional misuse of a system. As information is gathered concerning an epistemically uncertain quantity, one of the following situations typically occurs. First, the magnitude of the initial possibility interval typically decreases. In fact, if sufficient information is obtained, or control of the quantity is possible, then the quantity may become a single value. Second, as samples of the quantity are obtained, it may be found that the quantity is actually a random quantity, but the PDF is not known. For example, sufficient information may be obtained to characterize the quantity as random variable that can be described by a specific family of probability distributions, but the distribution parameters themselves are interval-valued quantities. This type of characterization of a quantity is referred to as an *imprecise probability*, i.e., a mixture of aleatoric and epistemic uncertainty. (In this volume, see the chapter by Bradley, but also Augustin, Coolen, de Cooman, and Troffaes 2014; Bernardini and Tonon 2010.) A particularly important application of these types of uncertainties is risk assessment of rare events, commonly called *black swans* (Taleb 2007), because they occur in the tails of distributions. Tails are particularly sensitive to epistemic uncertainty because they commonly have the feature of exponential decay, where epistemic uncertainty in the exponent produces a very large impact on likelihood. Taleb (2008) refers to these situations as the *fourth quadrant* because traditional statistics tends to be notoriously poor at predicting these events.

As a final observation of the important difference between aleatoric and epistemic uncertainty, note that the characterization of all uncertainties as precise probabilities versus imprecise probabilities serves two very different goals of an uncertainty analysis. That is, the use of precise probabilities tells the customer of the analysis the most likely outcome given the present state of incomplete knowledge. A recent example of this type of analysis is the search for the missing Malaysia Airlines Flight 370 (Davey et al. 2016). Specifically, this was an analysis that required an estimate of the most likely location to focus expensive search operations to find the lost aircraft. The use of imprecise probabilities tells the customer the *range* of likelihoods of outcomes, given the present state of incomplete knowledge. An example when this type of analysis is appropriate is the likelihood of failure of a high-consequence system.

3.5 Concluding Remarks

Most simulation analysts probably believe that my stress on the importance of code verification is erroneous. Most would argue that they are careful programmers and computational analysts and that all of the important coding bugs and algorithm deficiencies have been found and eliminated. There is rigorous code testing experience that shows that this assertion is unfounded. Hatton (1997) conducted extensive testing of 100 scientific codes over a period of 7 years. These were *production* codes in 40 application areas ranging from nuclear engineering to chemical engineering to medical software. To the disbelief of many, he found a dismal picture of software unreliability. The CFD community has found the same miserable picture from extensive code verification testing (Abanto et al. 2005; Rumsey et al. 2006), as well as workshops focusing on solving the same challenge problem by many different participants (Eca and Hoekstra 2002; Morrison 2014). Because of this recognition by the CFD community, additional research and practice to improve the reliability of software by way of more rigorous code verification testing have been done. It is encouraging that researchers from other fields are also beginning to call attention to the lack of reproducibility of simulation results (Donoho et al. 2009; Fomel and Claerbout 2009; LeVeque et al. 2012; Stodden 2012).

Solution verification is a well-understood issue, but it is almost universally ignored in computational simulation. This state of affairs exists for two reasons. First, most computational analysts do not feel that numerical solution error is important in their analysis. They feel that the numerical error is small based on their experiences with similar analyses. When numerical error estimators have been used, however, many researchers have shown that the analyst's experience is consistently *unreliable*. Second, existing numerical error estimators are not dependable outside of the asymptotic convergence region. Since the computational costs needed to achieve the asymptotic region for some system response quantities of interest, particularly local quantities, is so expensive on complex simulations, the vast majority of simulations do not use the existing estimation techniques. As a result, numerical solution error contributes to the lack of simulation reliability and reproducibility mentioned above.

The restricted concept of model validation discussed in this chapter is well founded in the philosophy of science. The common criticism of the concept is that it is too restrictive in the sense that it is not useful in most computational simulations. My counterargument to this criticism is that the restrictive view forces the conversation away from the vague and personally expedient concepts of validation, specifically, "My experience says the model is good (or good enough)." The restricted concept of validation compels the conversation into "Show me how your simulation performs for conditions where experimental data are available, then we will discuss how you estimate predictive uncertainty where no data are available." This should be the essence of predictive science.

The concept of predictive capability of a model as discussed here, particularly how it is related to model validation and model calibration, stresses the notion of foretelling system responses based on the physical principles in the mathematical model. Physical-based models have explanatory capability because they represent the systems' detailed, coupled physical structure, and its interaction of components such that they can predict future system behavior, even those never seen before. Descriptive models, such as those built on machine learning, data mining, and artificial neural networks, do not have the capability to predict physical phenomena for conditions that have never been seen before. Given the high dimensionality of the model input space for complex simulations, it should be recognized that nearly all simulation results of both descriptive and explanatory models are outside the convex hull of the observed data.

This chapter is focused on physical-based models, but several chapters in this volume describe models for the social sciences. Most social scientists agree that their theories and models are fundamentally different from physical-based models. Social science behavior models are descriptive, as opposed to explanatory, models that are based on the model builder's interpretation of the behavior of another living organism or the interaction of a group of organisms. Two brief observations contrasting the concepts of modeling and model validation discussed in this chapter, versus modeling human behavior must be noted. First, physical-based models *must obey* a set of fundamental principles, such as conservation of mass, momentum, and energy. In modeling human behavior as a system, there are *no such* mandatory or general principles. For example, the assumption that human behavior must be directed toward survival is not always accurate. Second, Fig. 3.1 in this chapter depicts the concept that model input data for the system and the surroundings drive the system response. As part of model validation, system responses can be measured in an experiment and compared with model predictions. However, in human behavior modeling, the depiction of input→model→response *must include* a feedback loop from the response to both the input as well as the model itself. That is, humans anticipate an expected outcome and commonly adapt both the input data and their internal model of the decision-making process to achieve some type of personally optimized result. As stated in certain social science research, humans are anticipatory systems with real-time feedback control. Furthermore, humans will change their input data and internal decision-making model simply if they know they are being observed. These types of complexities in modeling human behavior and in validating these models must be incorporated at the conceptual level.

Acknowledgements I thank Drs. Timothy Trucano, Patrick Roache, and Theodore Kneupper for carefully reviewing an earlier version of this chapter and providing many helpful suggestions for improvements and clarifications.

References

Abanto, J., Pelletier, D., Garon, A., Trepanier, J.-Y., & Reggio, M. (2005). *Verification of some commercial CFD codes on atypical CFD problems.* Paper presented at the 43rd AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV.

Aeschliman, D. P., & Oberkampf, W. L. (1998). Experimental methodology for computational fluid dynamics code validation. *AIAA Journal, 36*(5), 733–741.

AIAA. (1998). *Guide for the verification and validation of computational fluid dynamics simulations* (AIAA-G-077-1998). Retrieved from Reston, VA.

Ainsworth, M., & Oden, J. T. (2000). *A posteriori error estimation in finite element analysis*. New York: Wiley.

Anderson, M. G., & Bates, P. D. (Eds.). (2001). *Model validation: Perspectives in hydrological science*. New York, NY: Wiley.

ASME. (2006). *Guide for verification and validation in computational solid mechanics* (ASME Standard V&V 10-2006). Retrieved from New York, NY.

ASME. (2009). *Standard for verification and validation in computational fluid dynamics and heat transfer* (ASME Standard V&V 20-2009). Retrieved from New York, NY.

ASME. (2012). *An illustration of the concepts of verification and validation in computational solid mechanics* (ASME Standard V&V 10.1-2012). Retrieved from New York, NY.

Augustin, T., Coolen, F. P. A., de Cooman, G., & Troffaes, M. C. M. (Eds.). (2014). *Introduction to imprecise probabilities*. Chichester, UK: Wiley.

Babuska, I., & Strouboulis, T. (2001). *The finite element method and its reliability*. Oxford, U.K.: Oxford University Press.

Babuska, I., Nobile, F., & Tempone, R. (2008). A systematic approach to model validation based on bayesian updates and prediction related rejection criteria. *Computer Methods in Applied Mechanics and Engineering, 197*(29–32), 2517–2539.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., et al. (2007). A framework for validation of computer models. *Technometrics, 49*(2), 138–154.

Bernardini, A., & Tonon, F. (2010). *Bounding uncertainty in civil engineering*. Berlin: Springer.

Beven, K. (2002). Towards a coherent philosophy of modelling the environment. *Proceedings of the Royal Society of London Series A, 458*(2026), 2465–2484.

Bossel, H. (1994). *Modeling and simulation* (1st ed.). Wellesley, MA: A. K. Peters.

Chen, W., Xiong, Y., Tsui, K.-L., & Wang, S. (2008). A design-driven validation approach using bayesian prediction models. *Journal of Mechanical Design, 130*(2), 021101–021112.

Chiles, J.-P., & Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty*. New York: Wiley.

Cooke, R. (2004). The antomy of the squizzel: The role of operational definitions in representing uncertainty. *Reliability Engineering and System Safety, 85*(1–3), 313–319.

Cullen, A. C., & Frey, H. C. (1999). *Probabilistic techniques in exposure assessment: A handbook for dealing with variability and uncertainty in models and inputs*. New York: Plenum Press.

Davey, S., Gordon, N., Holland, I., Rutten, M., & Williams, J. (2016). *Bayesian methods in the search for MH370*. Springer Nature (Open Access).

Donoho, D. L., Maleki, A., Shahram, M., Rahman, I. U., & Stodden, V. (2009). Reproducible research in computational harmonic analysis. *Computing in Science & Engineering, 11*(1), 8–18.

Duggirala, R. K., Roy, C. J., Saeidi, S. M., Khodadadi, J. M., Cahela, D. R., & Tatarchuk, B. J. (2008). Pressure drop predictions in microfibrous materials using computational fluid dynamics. *Journal of Fluids Engineering, 130*(7), 071302–071313.

Eca, L., & Hoekstra, M. (2002). *An evaluation of verification procedures for CFD applications.* Paper presented at the Proceedings of the 24th Symposium on Naval Hydrodynamics, Fukuoka, Japan.

Ferson, S., & Oberkampf, W. L. (2009). Validation of imprecise probability models. *International Journal of Reliability and Safety, 3*(1–3), 3–22.

Ferson, S., Oberkampf, W. L., & Ginzburg, L. (2008). Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering, 197*(29–32), 2408–2430.

Ferziger, J. H., & Peric, M. (2002). *Computational methods for fluid dynamics* (3rd ed.). New York: Springer.

Fomel, S., & Claerbout, J. F. (2009). Guest editors' introduction: Reproducible research. *Computing in Science & Engineering, 11*(1), 5–7.

Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations* (4th ed.). Baltimore, MD: The Johns Hopkins University Press.

Haimes, Y. Y. (2009). *Risk modeling, assessment, and management* (3rd ed.). New York: Wiley.

Hatton, L. (1997). The T experiments: Errors in scientific software. *IEEE Computational Science & Engineering, 4*(2), 27–38.

Higdon, D., Nakhleh, C., Gattiker, J., & Williams, B. (2008). A bayesian calibration approach to the thermal problem. *Computer Methods in Applied Mechanics and Engineering, 197*(29–32), 2431–2441.

Jolliffe, I. T., & Stephenson, D. B. (Eds.). (2011). *Forecast verification: A practitioner's guide in atmospheric science* (2nd ed.). Hoboken, NJ: Wiley.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 63*(3), 425–450.

Knupp, P., & Salari, K. (2002). *Verification of computer codes in computational science and engineering*. Boca Raton, FL: Chapman & Hall/CRC.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). Berlin: Springer.

Leijnse, A., & Hassanizadeh, S. M. (1994). Model definition and model validation. *Advances in Water Resources, 17,* 197–200.

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Modern Physics, 41,* 253–262.

LeVeque, R. J., Mitchell, I. M., & Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science & Engineering, 14*(4), 13–17.

Li, W., Chen, W., Jiang, Z., Lu, Z., & Liu, Y. (2014). New validation metrics for models with multiple correlated responses. *Reliability Engineering and System Safety, 127,* 1–11.

Li, W., Chen, S., Jiang, Z., Apley, D. W., Lu, Z., & Chen, W. (2016). Integrating bayesian calibration, bias correction, and machine learning for the 2014 sandia verification and validation challent problem. *Journal of Verification, Validation and Uncertainty Quantification, 1*(1), 011004–011012.

Liu, F., Bayarri, M. J., Berger, J. O., Paulo, R., & Sacks, J. (2008). A bayesian analysis of the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering, 197*(29–32), 2457–2466.

Liu, Y., Chen, W., Arendt, P., & Huang, H.-Z. (2011). Toward a better understanding of model validation metrics. *Journal of Mechanical Design, 133*(13), 071001–071013.

Marvin, J. G. (1995). Perspective on computational fluid dynamics validation. *AIAA Journal, 33*(10), 1778–1787.

McFarland, J., & Mahadevan, S. (2008). Multivariate significance testing and model calibration under uncertainty. *Computer Methods in Applied Mechanics and Engineering, 197*(29–32), 2467–2479.

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis* (1st ed.). Cambridge, UK: Cambridge University Press.

Morrison, J. H. (2014). Statistical analysis of the fifth drag prediction workshop computational fluid dynamics solutions. *Journal of Aircraft, 51*(4), 1214–1222.

Neumann, P. G. (1995). *Computer-related risks*. New York: ACM Press, Addison-Wesley Publishing Company.

Oberkampf, W. L., & Aeschliman, D. P. (1992). Joint computational/experimental aerodynamics research on a hypersonic vehicle: Part 1, experimental results. *AIAA Journal, 30*(8), 2000–2009.

Oberkampf, W. L., & Barone, M. F. (2006). Measures of agreement between computation and experiment: Validation metrics. *Journal of Computational Physics, 217*(1), 5–36.

Oberkampf, W. L., & Trucano, T. G. (2008). Verification and validation benchmarks. *Nuclear Engineering and Design, 238*(3), 716–743.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge, UK: Cambridge University Press.

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety, 91*(10–11), 1290–1300.

Reason, J. (1997). *Managing the risks of organizational accidents*. Burlington, VT: Ashgate Publishing Limited.

Reason, J. (2008). *The human contribution: Unsafe acts, accidents and heroic recoveries*. Burlington, VT: Ashgate Publishing Co.

Refsgaard, J. C., & Henriksen, H. J. (2004). Modelling guidelines-terminology and guiding principles. *Advances in Water Resources, 27*(1), 71–82.

Roache, P. J. (1972). *Computational fluid dynamics*. Albuquerque, NM: Hermosa Publishers.

Roache, P. (2009). *Fundamentals of verification and validation*. Socorro, New Mexico: Hermosa Publishers.

Rougier, J. (2007). Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climate Change, 81*(3–4), 247–264.

Roy, C. J. (2010). *Review of discretization error estimators in scientific computing*. Paper presented at the 48th AIAA Aerospace Sciences Meeting, Orlando, FL.

Rumsey, C. L., Reif, B. A. P., & Gatski, T. B. (2006). Arbitrary steady-state solutions with the k–ε model. *AAIAA Journal, 44*(7), 1586–1592.

Rykiel, E. J. (1996). Testing ecological models: The meaning of validation. *Ecological Modelling, 90*(3), 229–244.

Silver, N. (2012). *The signal and the noise*. New York, NY: Penguin Books.

Stodden, V. (2012). Guest editor's introduction: Reproducible research-tools and strategies for scientific computing. *IEEE Computing in Science and Engineering, 14*(4), 11–12.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Taleb, N. N. (2008). The fourth quadrant: A map of the limits of statistics. Retrieved September 14, 2008, from https://www.edge.org/conversation/the-fourth-quadrant-a-map-of-the-limits-of-statistics.

Trucano, T. G., Easterling, R. G., Dowding, K. J., Paez, T. L., Urbina, A., Romero, V. J., … Hills, R. G. (2001). *Description of the sandia validation metrics project* (SAND2001-1339). Retrieved from Albuquerque, NM.

Trucano, T. G., Pilch, M., & Oberkampf, W. L. (2002). *General concepts for experimental validation of asci code applications* (SAND2002-0341). Retrieved from Albuquerque, NM.

Trucano, T. G., Post, D. E., Pilch, M., & Oberkampf, W. L. (2005). *Software engineering intersection with verification and validation of higher performance computational science software: Some observations* (SAND2005-3662P). Retrieved from Albuquerque, NM.

Trucano, T. G., Swiler, L. P., Igusa, T., Oberkampf, W. L., & Pilch, M. (2006). Calibration, validation, and sensitivity analysis: What's what. *Reliability Engineering and System Safety, 91*(10–11), 1331–1357.

Verfurth, R. (2013). *A posteriori error estimation techniques for finite element methods*. Oxford, UK: Oxford University Press.

Vose, D. (2008). *Risk analysis: A quantitative guide* (3rd ed.). New York: Wiley.

Voyles, I. T., & Roy, C. J. (2015). *Evaluation of model validation techniques in the presence of aleatory and epistemic input uncertainties*. Paper presented at the American Institute of Aeronautics and Astronautics SciTech Conference, Kissimmee, FL.

Wang, S., Chen, W., & Tsui, K.-L. (2009). Bayesian validation of computer models. *Technometrics, 51*(4), 439–451.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (3rd ed.). Amsterdam: Elsevier.

Wolpert, R. L. (2004). A conversation with James O. Berger. *Statistical Science, 19*(1), 205–218.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor, Michigan: University of Michigan Press.

Chapter 4
Verification and Validation Principles from a Systems Perspective

David J. Murray-Smith

Abstract This chapter introduces concepts and principles associated with the verification and validation of simulation models, mainly in the context of models of complete systems. The word 'verification' is used here to describe testing processes to establish whether a computer-based representation correctly describes the underlying mathematical, logical and theoretical structure of the model. The word 'validation' is used to describe procedures for establishing whether the model fidelity is adequate for the purposes of the given application. Verification is internal to the model and the computer-based representation while validation processes involve information external to the model, normally using data or observations from the corresponding real system. The goal of the testing process for a simulation model must always be to establish the extent to which a model has the quality and credibility required for the intended application. These model testing processes, involving both verification and validation, are inherently iterative.

Keywords Continuous system simulation models · Testing · Sensitivity · Identifiability · Documentation · Model acceptance · Model upgrading

4.1 Introduction

The purpose of this chapter is to introduce concepts and principles associated with the testing of simulation models. In this context a 'simulation model' is taken to be representation of a real system of some kind (often termed the 'target' system). The simulation model is normally based on a 'conceptual' or mental model which is then translated into mathematical relationships and logical statements (i.e. the underlying 'model'), together with a computer-based representation obtained from that model (as manifest in the 'simulation' program). The complete simulation model, once tested and shown to be an adequate representation for the purposes of the intended

D. J. Murray-Smith (✉)
School of Engineering, University of Glasgow, Rankine Building, Glasgow G12 8QQ, UK
e-mail: David.murray-smith@glasgow.ac.uk

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_4

application, offers a basis for experimentation and analysis that would be difficult or impossible in other ways.

The emphasis within this chapter is on models based on ordinary differential equations (ODEs) or differential algebraic equations (DAEs) involving combinations of ordinary differential and algebraic equations to describe 'lumped-parameter' models (involving discrete entities that approximate the behaviour of separable elements of the target system under certain assumptions). Lumped-parameter descriptions have many applications and, for example, form the basis for many simple models used to describe electrical circuits involving resistors, capacitors and inductors. In a lumped-parameter model each of the elements normally has only one property (e.g. resistance, capacitance or inductance), while the corresponding real element in the circuit would have more than one property. For example, a resistor may have some inductive properties as well as the dominant property that we recognize as resistance. The ordinary differential equations of lumped-parameter models allow a variable, such as an electrical voltage or current, to be found as a function of time at a specific point, whereas a 'distributed parameter' model involving partial differential equations would be necessary if the quantity of interest has to be found not only as a function of time but also as a function of another variable (such as position). Although they may provide a more accurate description of the underlying physical processes within the target system, the use of distributed parameter models involving partial differential equations can introduce significant computational burdens.

Lumped-parameter models are very widely used for the study of systems involving interactions between a number of separate elements including, perhaps, energy conversion and storage elements, actuators, sensors and communication elements. The modelling of systems of this kind may involve many separate elements and may lead to complex non-linear behaviour and may be described as being based on a 'systems perspective'. Examples can be found in many different areas, including engineering and physiology. Regardless of the application area, issues of testing that are raised in this chapter are typical of those encountered with other forms of simulation models too.

Approaches available for modelling and simulation have changed in recent years because of increased computational power at relatively low cost, improved software tools and enhanced user interfaces. Also, simulation software can be moved easily from one computing environment to another, leading to more re-use of programs, to the development of libraries of sub-models and to models that are, to some extent, generic and can be used to represent a range of different target systems. All these developments mean that testing issues are now as important as ever before and possibly even more important because of the rapid growth in the use of simulation techniques in many different application areas. However, detailed consideration of specific methods and their application is left to later chapters and the emphasis in this chapter is rather broader, with a focus on concepts and principles.

Testing issues are especially important in safety-critical areas of engineering (such as the aerospace, defense, marine, off-shore and nuclear application areas) as discussed by several contributors to the volume edited by Cloud and Rainey (1998) and considered further by other authors such as Pace (2004) and Oberkampf and Roy

(2010). In those fields formal approval schemes, are often applied throughout the design and development process, including rigorous testing of simulation models (see, for example Mitre Systems Engineering Guide 2014). However, in some other application areas, the attention paid to model quality issues is often inadequate. Frequently, the use of a model is justified by the fact that it is 'based on well-known physical principles', or involves 'an industry standard', or even that this form of model 'has always been used by us, so must be right'. Also, proper documentation of model development processes has often been neglected, although, in some fields such as biology and medicine, much more emphasis is now being given to model testing as well as to documentation issues. It is vitally important that good practice in those fields, as well as in safety-critical engineering applications, should be extended to simulation modelling for all application areas. One very encouraging sign is the publication of the results of work carried out by the Modeling Good Research Practices Task Force established jointly by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the Society for Medical Decision Making (SMDR). Reports published by both societies (e.g. Caro et al. 2012; Eddy et al. 2012) make recommendations for achieving transparency and validity in simulation modelling activities and strong proposals are made regarding best practices. Although these relate immediately to biomedical and health care system modelling, many of the recommendations apply equally to simulation and modelling activities in other fields.

Simulation models should always be developed for a specific application and then subjected to tests which are considered appropriate for that application. Such tests must include all the processes associated with establishing that a given model is suitable for the intended use. Reaching a conclusion about model acceptability is never straightforward and testing issues should be considered from the outset of the model development process. Rigorous documentation is important and must include information about the model requirements, details of testing procedures, results of tests, names of model developers and testers (if different), dates of the start of model development and of changes in the model, details of re-testing carried out, reasons for acceptance/rejection of the model and any limitations that may apply to its use if accepted. Transparency is vitally important in all aspects of the development, testing and application of simulation models and the application of sound principles of model management, including extensive, clear and easily used documentation is essential.

These testing procedures for simulation models fall into two distinct parts, defined here as 'verification' and 'validation'. There is an unfortunate lack of unanimity about the meaning of these words and they are considered interchangeable by some. However, there is general agreement that the process of simulation model testing involves two issues one of which concerns the correctness, or otherwise, of the process of translating the conceptual, mathematical and logical basis of a model into the description implemented on a computer. The other issue is concerned with potential errors and uncertainties within the structure and logic of the underlying model, along with limitations of that description in terms of accuracy.

If we accept that the first part of the testing procedure involves a process based only on the underlying model and the associated simulation software, it can be described conveniently as an internal process. Not only must the structure and internal logic of the computer program be shown to be correct, but the algorithms within the simulation program must also be shown to be appropriate and implemented properly. It can never be claimed, following verification, that a simulation model is correct. However, it can be stated that the model has been exercised over the full range of relevant conditions and that it has successfully passed the specific tests that were applied. Errors that are detected during verification may be categorized as 'acknowledged' errors which can be estimated (such as errors arising from the use of a specific numerical algorithm) or 'unacknowledged' errors which are simply due to mistakes (in coding, for example).

Successful verification does not tell us that a model adequately represents the corresponding real-world system. That involves the second part of the testing process and is based on information that is external to the model and the associated simulation program and may involve many different comparisons with real-world data. The goal in validation must always be to ensure that the underlying model is sufficiently accurate for the planned application.

The validation of simulation models thus involves establishing the extent to which a model is an accurate representation of the real world from the viewpoint of end-users of that model. Compared with verification, validation is a more open-ended task in which comparisons are made between model behaviour and behaviour of the real system for the same conditions. Some would claim that the emphasis should be on 'invalidation', since any suggestion that a model is correct may be refuted at any time when someone undertakes a different test yielding data that are not fully explained by that model. Thus, although decisive invalidation is always possible, statements involving decisive validation, suggesting that a model or facts derived from it are true, should be avoided. However, the word 'validation' is generally convenient to describe the process of building up trust in a model for a specific application and this view of validation procedures is of great practical importance.

For many modelling applications, such as in climatology or economics, experiments cannot be performed on the real system and the only data sets available for validation are historical. Often only one historical data set is available and may have to be split into two, with one part being used for estimation of unknown or ill-defined model parameters and the second for testing. The same data cannot be used both for model development processes, such as estimation of parameters, and for validation. For applications in engineering, physical sciences and biology, experiments on the real system may be possible and this simplifies the situation, as discussed by the current author in Chap. 15 of this volume. However, a simulation model is often needed for the design of a new engineering system and validation of that model using experiments is clearly impossible until at least a prototype system is available, usually at quite a late stage of the design project for which the model is being developed.

Simulation models often give rise to large numbers of output variable time histories, each with errors that may vary with time. It is vital, therefore, to establish the outputs that are most important for the application. Perfect matching by model variables of corresponding quantities obtained from measured data sets resulting from observations or tests on the real system that is being modelled is never a realistic aim. Model variables should match real system response data only to the level needed for the application and models inevitably involve uncertainties however they are developed and whatever the intended application. These uncertainties may be associated with the modelling assumptions and the boundaries being applied in defining the model. They could also result from environmental unknowns, or the effects of unmodelled quantities considered as external to the model, or may be due to other factors such as inaccuracies in values of model parameters. This suggests a trade-off so that model responses match test data to an acceptable degree, while also showing adequate robustness to all these different sources of uncertainty.

Testing of a simulation model can never be regarded as a simple straight-through type of process. It is an iterative procedure in the sense that any changes made in the underlying model must be followed by further validation tests and then by additional verification checks. In addition, it can be argued that the verification and validation processes are, in themselves, iterative since they involve many steps that are applied many times.

The application of systematic procedures for testing at every stage in the development of a model may lead to three possible outcomes, provided verification requirements are satisfied at each stage. First, there may be no model structure and set of parameters consistent with the measured response data obtained from the real system. One must then start again at the model formulation stage, taking account of what has been learned.

A second type of outcome could involve parameter values that give model behaviour that is in broad agreement with the behaviour of the real system but with some major uncertainties remaining. The model may still be of value if it has a sound theoretical basis and one step could involve investigating parameter sensitivities to enhance understanding of the model. The adjustment of parameters within a non-linear computational model to improve the agreement with real-world information is often termed ‘calibration’. However, model calibration and adjustment of a model using system identification methods should be recognized as being different from validation and re-testing of a model is always necessary after changes have been made in its structure or parameter values.

The third possible outcome is the most desirable and arises in situations where testing shows model predictions that are considered acceptable for the intended application, and where the model structure and parameter values are also plausible. Further analysis, combined with a review of the accuracy requirements defined at the outset of the model development process, may help to establish limits of applicability for the model and, for operation within those limits, the model may be judged to be acceptable for the required application, but only for the tests carried out. It is then possible to use the model, but only until new evidence is found that invalidates the model and leads to further testing. Good documentation is needed at every stage of this iterative process of model development so that there is a clear audit trail showing how every decision is reached, the justification for model acceptance and the range of conditions over which the accepted model can be used.

Although validation is not straightforward, even for models of physical systems, it is at least possible in such cases to establish the credibility of a model through experiments performed on the real system. In contrast, models of social systems and models that contain elements of human decision-making, require validation that more often involves establishing qualitative credibility rather than detailed quantitative comparisons. Conclusions then tend to be based more on statistics rather than point-by-point comparisons of time histories. The task becomes one of showing that the model produces sound insights and sound data based on a wide range of tests and criteria, often without the possibility of experimentation on the real system. The outcome is always a judgement based on available evidence and the previous experience of those involved in the assessment of the model, taking account of the requirements of the intended application.

4.2 Approaches to Verification

As discussed in the Sect. 4.1, the word 'verification' describes the process used to establish that a computer simulation is consistent with the underlying mathematical and logical description. Verification of a simulation model is a process that extends more general and traditional processes that arise in software testing such as those presented by Kit (1995) and by Kaner et al. (1999). However, the situations that arise in verification can be much more demanding than in many other areas of computer software testing and it is recognized, especially, that simulation models based on partial differential equations present difficulties which cannot be handled using traditional software testing processes alone (e.g. Oberkampf and Roy 2010).

Although current research on so-called 'formal methods' in computer science may well lead to future developments that allow firm statements to be made about the correctness or otherwise of a given computer program, currently available methods for routine testing of software do not provide proof that a given computer program is completely free from errors. The use of well-designed test cases is therefore critically important for verification. These must be fully documented and must include information about the computer hardware, operating system and other software used.

There are two distinct phases within the verification process, each involving several distinct steps (see also Chap. 11 by Rider and Chap. 12 by Roache in this volume). The first may be termed the 'code-checking' or 'code-verification' phase:

- The first step of the code-checking part of the verification process is finding and removing simple errors and mistakes in the source code. This is intended to establish the correctness and robustness of the code in the context of the application. It includes checking every line of a program for potential errors and the code review must include checking links to separately tested sub-models or connections between functional blocks if the simulation is developed using a graphical user interface.
- The second step in the code-checking part of the verification process involves assessment of the suitability or otherwise of numerical algorithms used within the simulation. This is important since a simulation model may require the use of many different algorithms, each with its own numerical errors. Whether one

is dealing with a lumped-parameter simulation model based on ordinary differential equations, a distributed parameter simulation involving partial differential equations or a discrete event (e.g. agent based) or hybrid simulation model, many different types of algorithmic errors can arise and investigation of links between error sources can be complex. The overall accuracy of a simulation program in providing an implementation of the underlying model may be very different from the formal order of accuracy for a single algorithm used within that simulation (e.g. for numerical integration). For example, problems of quantization may arise due to the word length being used for certain variables within a simulation model. Truncation errors may also arise in some algorithms because perturbations in variables are not small enough to ensure a solution within the asymptotic convergence region. Also, singularities or discontinuities may lead to problems with integration algorithms and lead to important hard non-linearities being ignored. Such issues must be considered carefully in the context of the model structure and the intended application and establishing the source of any problem often requires a preliminary examination of the behaviour of sub-models. This could involve testing the simulation program for each sub-model separately by running it and examining the results, usually in graphical form, to check for unexpected features, such as very-high-frequency oscillatory behaviour that could not be expected in terms of any reasonable behaviour of the sub-system. Other tests might involve making small changes in parameters associated with each numerical algorithm. If the algorithm is being used in an appropriate way such small changes would not, normally, have major effects on simulation results. Any large changes observed in numerical results following some small change in a parameter associated with a specific algorithm should, therefore, be investigated further.

- The final stage of this code-checking phase of verification, like algorithm checking, also involves testing of the complete simulation program. Although most simulation models are dynamic in form, this step can be applied for specific conditions in which the model is expected to be in a steady state. Prior knowledge about such conditions usually comes from analysis of the underlying dynamic model in mathematical form and is termed ‘static analysis’ since derivative terms in the ordinary differential equations of the model are set to zero. The idea is that the results of the static analysis for the mathematical model should match equivalent steady-state conditions observed from output results obtained from the simulation model. Such tests may also involve the examination of the rate of convergence of numerical solutions to stable values as control parameters within relevant numerical algorithms are adjusted. For example, in fixed-step integration algorithms, this could involve adjustment of the size of the integration step, while in models involving partial differential equations the overall spatial resolution of the discretization grid might be changed.

The second phase of the verification process involves the assessment of numerical errors in responses obtained from a simulation model and is often termed ‘solution verification’. It involves a more direct estimation of errors in numerical solutions to ensure that the simulation model gives outputs that approximate to the true solution of the equations of the underlying mathematical description. This involves a process that is similar in some ways to testing procedures carried out in the first phase but, whereas that is concerned with errors in coding, the focus in this second phase is on the overall performance of the numerical algorithms that form the basis of the simulation model. This could involve separate consideration of specific aspects of the problem such as the following:

- Investigation of the overall accuracy in the context of the requirements of the proposed application, such as accuracy in terms of integration errors and any errors in the detection of discontinuities. In this respect, the model developer must play an important part in devising suitable tests, which could involve comparisons of accurate solutions for specific well-understood cases with solutions obtained by means of the simulation model (such as steady-state conditions or a special set of conditions for which analytical solutions can be found). Testing of this kind for the complete simulation model inevitably involves many runs of the simulation code and raises fundamental issues because simulation techniques are most often used for applications in which analytical solutions are not readily available. The areas of the solution space that are of greatest interest often correspond to situations for which analytical solutions cannot be found. Special cases can be considered and could even involve some modification of the mathematical model and the corresponding simulation to represent a case which has an analytical solution that can then be compared with simulation results. Such testing based on steady-state and other special cases do not eliminate the possibility that other coding errors exist but successful completion of such tests does increase overall confidence. This type of procedure requires considerable experience and insight on the part of the person carrying out the tests and may be difficult to apply for large and complex simulation models.
- In the simulation of continuous dynamic systems, it is usually not appropriate to record simulation variables at every integration step and one very important issue relates to the interval between the time instants at which the outputs are recorded for further analysis and plotting. For example, in a real-time simulator which may include a human operator such as an aircraft pilot, there is also a need to control data transfer through the various communication channels between the simulation model, the human operator and any external hardware coupled to the simulation. Such flows of information are determined by ‘communication intervals’ that are set within the software of the simulation model and their choice is an important issue. An inappropriate value of communication interval could cause transients to disappear from the observed responses of the simulation model even if they were correctly represented within the simulation.
- Uncertainties about the values of the parameters within the model (e.g. quantities that are represented by constants in the underlying equations) and about the structure of the model (e.g. the number and form of the equations describing the system under investigation) must be given due attention when assessments are being made of simulation results. It is pointless to have high levels of numerical accuracy in

a simulation model if basic uncertainties about the model structure and the model parameters remain large.

The practical application of these steps for verification of a specific simulation model has much in common with procedures that are generally accepted and widely used for checking and testing of other types of software for many other applications. However, the following points that relate more specifically to simulation models should also be considered:

- A graphical representation of the simulation model should be included in the documentation for simulation models developed using conventional programming tools. The process of creating this graphical representation for the model may help identify potential problems with a dynamic model (such as the existence of algebraic loops).
- All possible outcomes occurring during the operation of the simulation model should be established and represented within a flow diagram relating to the logical structure of the model and this should form part of the documentation. Again, this may help in identifying potential problems with a computer-based model at an early stage in its development.
- Values of parameters provided as input during testing of a simulation model should be displayed at the end of each test to check that no changes have occurred when running the simulation.
- An animation may be useful for checking that observations from the simulation model do not include features that differ significantly from the known or expected behaviour of the real system. Animations are often easier to understand than a collection of simultaneous output time-history records for different model variables.
- All models and associated simulation programs, together with the documentation, such as graphical representations and flow diagrams, should be subject to independent checks carried out by someone who was not involved in the development of the model.

Although the stages of the verification process discussed above are of general applicability, some special issues arise with specific types of model. For example, distributed parameter models, which are based on partial differential equations (PDEs), are discretized for spatial dimensions as well as for time and involve boundary conditions. The solution quality depends both on truncation errors and discretization errors. Truncation errors depend on the accuracy of the solution of the discretized equations while the latter arise from errors caused by using discretized equations to represent the original PDEs. Errors may also relate to the process of discretizing boundary equations and auxiliary equations (in the form of additional algebraic equations). As discussed by many authors, including Oberkampf and Roy (2010), numerical issues of this kind depend on the size of spatial grid used and verification issues are of great importance. Many approaches are in common use, some leading directly from early work by Steinberg and Roache (1985), who presented a form of verification methodology using analytical solutions based on symbolic manipulation (see Chap. 12 by Roache in this volume).

Verification processes for discrete-event models also divide conveniently into two groups. These involve (a) checks to detect coding errors and (b) checks to establish that the chosen algorithms are appropriate for the application. In the case of discrete-event simulation models, algorithmic checks must establish, for example that routines for generation of specific random variables provide the desired statistical properties. Hybrid models, which are partly continuous and partly discrete, also require algorithmic checks for features involving continuous system simulation elements and special attention must be given to interfaces between continuous elements and discrete elements. For example, for a model representing a digital processor controlling some equipment that operates in a continuous fashion (the plant), careful checks must be made of the sub-models representing analog-to-digital and digital-to-analog converters.

Establishing whether a set of verification tests is sufficient presents obvious difficulties and is a process that may be viewed more as an art than a science. However, some new approaches are being introduced that have an objective basis. One example that provides a measure of the completeness of testing is Modified Condition/Decision Coverage (MC/DC) which is an approach used in avionic system development (Heyhurst et al. 2001). The word 'coverage' is used here to provide an indication of the extent to which the model logic has been exercised during testing. The tools provide a way of analyzing code and detecting reasons for errors when the simulation program is run and thus provide an alternative to other approaches.

4.3 Approaches to Validation

In engineering and the physical sciences, methods of validation are often quantitative and involve direct comparisons of chosen model variables with corresponding measured quantities in the target system. This may be termed 'predictive', 'empirical' or 'pragmatic' validation and is particularly important in applications such as the design of automatic control systems where the quality of the final control system performance depends very much on the quality of the plant model, especially in some specific parts of the relevant range of frequencies. However, it is impossible to consider every case since the number of tests required to demonstrate consistency would be very large and there is never any conclusive 'proof' of validity. Indeed, Oreskes et al. (1994) and others have argued that the use of the word 'validity' may be misleading since validity is a property that applies, on a strict basis, only to logical arguments. For practical and cost reasons testing must always be selective and the number of simulation runs should be based on insight about the behaviour and properties of the real system and the requirements of the application.

Although the level of agreement between selected model variables and the equivalent measured quantities is important in considering model fidelity, we must also consider broader questions in terms of the consistency of models with accepted theoretical laws and principles. It cannot be assumed that a model giving satisfactory predictive accuracy for specific variables is based completely on sound theory or that parameters within the model are physically meaningful. The predictive agreement simply suggests that the model, as a form of input–output 'black box' type of description, has some level of credibility. Examples of this form of description may be found in time series analysis in many different application areas (see, for example Chatfield 2003). For instance, in the chemical process industries and in physiology, highly simplified sub-models (incorporating a pure time delay, for example) are sometimes used to represent much more complex phenomena. This type of approach is discussed in contributions to a special issue (Schlacher and Schöberl 2011) of a journal specializing in dynamic system modelling problems. Approximations of this kind, perhaps used as simplified representations within a larger and more complex simulation model, inevitably have limitations and it is important to ensure that any resulting restrictions in the applicability of the complete model are understood by users. Input–output agreement between model and system is usually only a starting point for analysis aimed at investigating broader issues of 'theoretical' validity which relate to assumptions and simplifying approximations used within the model. For example, in aircraft flight control system design, a linear model may be used for one specific operating point within the flight envelope, but a model of this kind is useful only for small perturbations of variables about that operating point. Such a model may be helpful during the early stages of design (using linear control system design techniques), but a more physically based detailed non-linear model would be essential for investigating system performance for larger changes of model variables and operating conditions.

In other application areas, computer simulation may sometimes be applied when initial knowledge about the real system is very limited. Examples could include situations where experimental investigations may be constrained by practical, safety or ethical issues, as can often apply in physiology and medicine. Clearly, for models involving many uncertainties and limited experimental data, validation processes based on quantitative methods become more difficult to apply and testing may have to include qualitative 'face' validation methods involving peer review and expert opinion, as discussed by many authors including, for example Murray-Smith (2015).

Accuracy requirements for any model must be established before model testing is undertaken and such information should really form part of the model requirements specification established at the start of the model development process. Criteria used to assess the fidelity of the model must remain the same throughout and should be linked closely to the model application. Knowledge and understanding of the system being modelled is of prime importance and is usually enhanced during the model development process.

The capability of a model to predict system outputs for experimental situations that differ from those used during tests performed during model development is often termed 'generalization'. Although interpolation between operating situations explored during the testing process is often appropriate, any extrapolation of model usage beyond the range of conditions for which it was developed and tested must be treated with caution. Models should always be applied with full knowledge of the operating conditions and test inputs considered during the validation process.

4.3.1 Quantitative Approaches to Validation

Quantitative methods of validation may be categorized in several ways. The commonest approach involves predictive methods and simple time-history comparisons which may either involve data from measurements on the real system or comparisons with outputs from other models of the same system that have already been successfully tested and accepted. This latter type of situation can arise, for example, with a simulation model intended for application in real time. Many different deterministic or statistical measures may be useful for assessing the closeness of the model and real system results. However, validation should not only involve comparisons in which the data sets from the real system measurements are assumed to be exact and correct since system observations and measurements also involve errors and uncertainties.

The greater the understanding of all aspects of the real system on the part of all involved with the model development and testing procedures the more straightforward the validation process is likely to be. Parts of the model that include significant uncertainties need to be isolated in some way and one approach to this involves making comparisons between selected pairs of variables of the real system and model, while using other measured quantities from the real system as measured inputs for the model. Plots of residual time histories formed from the differences between model output values and the corresponding measured output values can be particularly helpful in investigating errors in parameters and model structure. For example, for an adequate model, we would expect residuals to be uncorrelated and have properties closely resembling those of white noise. Thus, if the autocorrelation function for a specific residual time history is formed and is found to have properties close to those for white noise, the residuals may be regarded as uncorrelated. Any errors in the model structure would be indicated by correlated residuals.

In applications involving the design of a new engineering system the 'real system' only exists towards the end of the design and development process, when a prototype has been commissioned or when the system itself has been built. Up to that point, only comparisons with similar models from earlier projects are possible. In such cases results from validation work carried out in testing sub-system models during earlier projects can be helpful if these are being re-used but, as pointed out by Hemez (2004) and others, this is only possible when good documentation exists.

Specialized systems-engineering tools involving the techniques of parameter sensitivity analysis can often provide useful insight relating to model credibility. Such analysis of mathematical models can be based on partial differentiation and the resulting mathematical expressions can provide valuable and very direct insight even without numerical evaluation. However, the simplest approach involves varying parameter values one-at-a-time and using numerical differencing of results. As explained in early work on this subject by Tomović (1963) and by Frank (1978), such analysis allows investigation of the effects of variations of model parameters and can highlight possible parameter interactions associated with the fact that different model parameters may offset each other in terms of their effects on model variables. This can introduce difficulties in model validation which can be eliminated if such prior sensitivity information is available. A review by Hamby (1994) in the context of environmental models provides much useful information, including comparisons of different approaches.

Many different researchers, starting from important early work by pioneers such as Bellman and Åström (1970), Grewal and Glover (1976), Beck and Arnold (1977) and Goodwin and Payne (1977), have found that system identification and parameter estimation techniques also provide useful insight if model responses show unexpected features. These techniques, which are well established as experimental modelling tools in fields such as control engineering and pharmacodynamics, provide a further approach to validation. The concepts of identifiability analysis as presented in the work of Bellman and Åström (1970) and subsequently extended by many others, can also provide valuable information about model structures and the effects of parametric interactions and can help in choosing between competing descriptions. More recent accounts by Raue et al. (2009) and by Gàbor et al. (2017) demonstrate the value of identifiability concepts in modelling applications in the bioinformatics area. As discussed further in Chap. 15 in this volume, identifiability analysis and system identification methods also provide insight into the design of experiments to maximize information gathered from tests on the real system.

Although system identification methods usually involve linear models they may also provide insight into the non-linear case. For example, identifying linear models for several test signal amplitudes and different operating points across the system operating envelope allows parameter estimates to be compared with values obtained from linearised theoretical descriptions for the same operating conditions. Trend comparisons for these estimated and theoretical values for several operating points provides an indication of performance of the underlying theoretical non-linear model. This has been successfully used in testing non-linear physically based helicopter flight-mechanics models where estimated parameter values in low-order linear models were compared with theoretical values for several flight conditions (Bradley et al. 1990).

While tools such as parameter sensitivity analysis and system identification and parameter estimation methods are well established, other approaches, such as the 'model distortion' methods of Butterfield and Thomas (1986) and the so-called 'barrier certificate' methodology described by Pranja (2006), have also been developed and used in some specific areas. For example, accounts of barrier certificate methods for model quality assessment for a continuous time biochemical system model and a discrete-time model of population growth may be found in the work of Anderson and Papachristodoulou (2009).

It has often been suggested that formal methods from the computing science field have a role in validation of simulation models, but little evidence can be found that formal methods are being used currently for validation purposes in large-scale modelling applications (Kuhn et al. 2002; Gore and Diallo 2013). There are several difficulties that limit routine adoption of formal methods, but one major problem is that using this approach requires a large investment in time and effort for a new simulation model and this is probably unaffordable except, possibly, in a few safety-critical application areas. However, formal methods are used for design flaw detection in the development of microprocessor chips and other complex electronic hardware (Heitmeyer 2007) and, as usage increases in those areas, the more likely it is that formal methods will also find a role in simulation model development.

As mentioned earlier, the general principles outlined here apply also to distributed parameter models based on partial differential equations. However, some issues arise in the validation of distributed parameter models that merit special consideration. Firstly, testing involving measured data from the corresponding real physical system is undoubtedly more complicated than it is in the case of lumped-parameter models. The quantities to be compared in the system and the model, the number of sensors to be used for measurements in the real system and the positions of those sensors are all important issues. Using more sensors gives improved resolution, but a balance always needs to be found between model quality and costs. Also, in some situations, the measurements may change the system through, for example, the added mass of sensors, wiring and telemetry hardware for signal transmission. Interest in validation issues for distributed parameter models has grown in recent years and methods developed for specific fields, such as computational fluid dynamics, are now being used for other forms of distributed parameter model (e.g. Oberkampf and Roy 2010).

Validation for discrete-event simulation models must include consideration of general issues such as the model structure and parameters, as in the case of continuous system simulation models. However, in discrete-event models, special consideration must also be given to any assumptions that have been made within the model about the probability distributions of events. When a specific distribution is used, the underlying assumptions must be tested using data from observations of the system being modelled. For example, in road traffic modelling, a discrete-event simulation model involves assumptions about the probability of the time intervals between the vehicles arriving at each road junction from different directions and such assumptions must be tested. Although this is particularly important in the testing of discrete-event simulation models, probability distributions may also appear within some continuous system simulation models and the underlying assumptions must again be tested in such cases.

For hybrid system simulation models involving, for example, a system with an embedded digital processor and associated analog-to-digital and digital-to-analog converters as well as continuous dynamic elements, the model validation process is basically the same as for a continuous system model. However, in the hybrid case, additional quantities must be recorded from the real system, such as information about the timing of events within the digital processor and converters to ensure that discrete-time elements within the hybrid system simulation model are properly represented.

4.3.2 Qualitative Methods: Face Validation Approaches

Instead of depending only on quantitative methods, such as those outlined above, a more subjective type of test may also be included within the overall assessment of a simulation model, in some cases. This takes into account opinions of people, who have extensive and detailed knowledge of the operation of the real system in normal and abnormal circumstances (cf. Chap. 17 by Saam in this volume). Such an approach is often termed ‘face’ validation and, although regarded by some as being based too strongly on personal opinions, there is no doubt that these methods have proved beneficial in many simulation applications ranging from aircraft handling qualities research using flight simulators to modelling problems in physiology and medicine. When used appropriately, and carried out with rigour, face validation tests can be regarded as a kind of Turing test in which the human expert is required to distinguish between the behaviour of the real system and the behaviour of the simulation model from observations of the available outputs from both. This type of approach can be especially helpful in establishing the correctness or otherwise of the logic and input–output relationships within the model. Face validation is often most important in the early stages of a project where real system data do not exist, but it can also be helpful in the modelling of cases in which the model parameters and the model structure might be chosen to cover a range of normal (and possibly also some abnormal) conditions, as in physiological models. Validation of such models is then a process which depends very much on interpretation of model behaviour and broad comparisons and previous experience gained from corresponding experimental results or observations.

Examples of face validation can also arise when computer simulations are used together with external hardware to form a ‘hardware-in-the-loop’ simulation. This is especially relevant in control system hardware development where a real-time simulation model of the system to be controlled (the plant) may first be developed. Once the plant model is tested and accepted, the real-time simulation may serve as a test-bed for controller hardware and software development (see, for example, Murray-Smith 2015). Face validation methods may be particularly useful in checking the behaviour of the simulation model for fault situations within the real system where quantitative comparisons may present practical difficulties.

4.3.3 Validation of Library Sub-models and Generic Models

Libraries of sub-models are often used in engineering and are increasingly being developed for other fields. Models which are developed with a structure which allows them to be applied for a range of different applications are often termed ‘generic models’ . They are potentially useful in situations where a broadly based description can be adapted for some new application at a cost that is less than that of developing a new model specifically for that new case. Good examples of generic modelling can be found in some areas of engineering, such as gas turbine engine design.

Testing processes leading to a sub-model being accepted for inclusion in a library are based on the verification and validation procedures that apply for other types of simulation model. Documentation of tests for library sub-models should be as detailed as for any other accepted model. The validation process for a generic model is usually approached through specific applications where other accepted models are already available (see, for example Smith et al. 2007).

4.4 Acceptance or Upgrading of Simulation Models

In considering acceptance of a simulation model for a planned application the computer-based representation must, first and foremost, be consistent with the structure, parameters and logic of the underlying model. This means that all the verification tests must have been completed satisfactorily. Although verification is very important within the overall testing process for a simulation model, eventual acceptance or rejection usually depends more critically on the results of external procedures, considered in the context of the application. Unlike verification processes, which are essentially quantitative in nature, validation can involve qualitative as well as quantitative considerations.

When model upgrades are needed, parametric changes are usually considered before structural changes are looked at. Often, the validity of a model may be improved simply through parameter adjustment, but this is possible only if the parameter value remains within an appropriate range. Adjustments using global optimization methods without consideration of physical limits often produce misleading results. For example, lumped-parameter descriptions are often used to approximate more complex effects and there are limits to the conditions for which these are valid. If parameter values arise that lack physical meaning, this may be because the model has an inappropriate structure. Correction of deficiencies in model structure often requires considerable insight, but once the structural errors are dealt with the upgraded model should have a wider range of applicability and the relevant parameter values can be reconsidered.

It is important to stress, once again, that the complete testing process for simulation models is iterative. After every significant change, further verification and validation tests must be performed and fully documented. To be successful, this process requires good management, and this is something that is too often neglected within organizations that develop and use simulation models. As already discussed in Sect. 4.1, a model is only useful if details are recorded and made available concerning the model requirements, development history, testing processes and acceptance criteria (e.g. Murray-Smith 2015).

Computer-based simulation models that have been fully tested and accepted should capture the complex and often non-linear features of the real system and provide insight that allows the developer and users to gain an understanding of key variables and their causes and effects. Such understanding means that well-supported statements can be made about why events can or cannot occur within the system and may allow the model to be used for decision-making or design.

4.5 Discussion

A careful distinction has been made throughout this chapter between the processes of verification and validation. This distinction is possibly more significant in practice in some application areas than in others and is possibly most obvious in the engineering field where the development of complex simulation models is often a group activity, rather being the responsibility of an individual. In such situations, the procedures for management and documentation of the model development process may well lead to a formal separation of the verification and validation procedures. Another practical issue in many engineering applications is that experiments on a target system to generate data for comparison with simulation model responses can be expensive and this means that the cost of validation activities must be factored carefully into the initial plans for the development of the simulation model. The use of experimental data in simulation model validation is discussed in more detail by this author in Chap. 15 of this volume.

Although verification and validation activities are often considered as separate procedures it must be accepted that there are strong links between these two stages of the overall testing process, as discussed in earlier sections of this chapter. For example, in Sects. 4.2 and 4.3 specific mention is made of the use of analytical solutions for special cases (such as steady-state conditions) within both the verification and validation process. The inherently iterative nature of the testing processes for simulation models has also been discussed in Sect. 4.1, with repeated verification and validation tests being performed when changes are made in the underlying model and in the simulation. Winsberg (2018) has argued that conceptual divisions between verification and validation can be misleading and may lead to difficulties, since such divisions can weaken the significance of links between these processes in the minds of those involved. He argues, for example, that the choice of model structure depends not only on known properties of the target system and the application of established laws and principles but also on computational tractability. The use of lumped-parameter descriptions rather than the more complex distributed parameter form of model, as discussed in Sect. 4.1, is a good example of this. In general, there are many different forms of mathematical description that may be considered in any given modelling situation and the chosen model is usually a compromise. Issues of computational tractability and the overall accuracy necessary for the intended application of the simulation model, therefore, link strongly with other issues considered in selecting the model structure. There is, therefore, a similar link between the associated verification and validation procedures. Although the separation of the verification and validation stages may often be convenient in terms of model management and the documentation of testing procedures, it is important that all involved in the development and testing of a simulation model should be fully familiar with the details of both the verification and the validation processes that are being applied. Further consideration of the issues associated with the separation of verification and validation processes is given in Chap. 42 by Beisbart in this volume.

The main emphasis within this chapter is on lumped-parameter models based, normally, on the use of continuous system simulation software tools. However, the issues that arise in the testing of such simulation models carry over to other types of simulation, including those involving discrete-event and hybrid models (which involve both continuous system and discrete-event simulation methods). The broad concepts of simulation model testing remain broadly similar in all these cases.

Concepts from parameter sensitivity analysis and identifiability analysis can provide valuable insight, as can the use of face validation where people with an understanding of real systems of the type being modelled can propose testing strategies that are based on years of practical experience and can readily pinpoint aspects of simulation model behaviour that are in some ways inadequate. Additional, more quantitative, testing can then be carried out using results from face validation as a starting point.

Issues of accuracy must be looked at very carefully, both in the context of verification and validation. There is little point in having high levels of numerical accuracy in a simulation model if uncertainties in data or in model structure and parameter values remain large. Verification and validation issues must be taken fully into account from the earliest stages of the modelling process and this brings attention back to the definition of requirements in the initial stages of model development where the eventual user of simulation results has an important role. Equally, the end-user also has an important role in decisions regarding the quality and reliability of simulation results.

Model validation is often compared with legal processes where an accused person may be declared innocent or guilty, beyond any reasonable doubt. In modelling, the outcome involves accepting or rejecting a model for a specific application. However, clear differences do exist between legal processes and model validation procedures since, in the latter case, we are also interested in establishing the range of conditions over which the model is useful. In addition, a model must be re-tested whenever new evidence is found that was unavailable when the model was first tested and accepted. This resembles a re-trial in the legal context and emphasizes the fact that validation of a simulation model is an ongoing process and not simply a result. The process gives the user a better understanding of a model's capabilities, limitations, and appropriateness for the intended application but does not prove the correctness of the model itself. Building confidence is an iterative procedure based on repeated testing, both in terms of verification and validation, always with the underlying aim of trying to 'break' the model by demonstrating that its behaviour does not match sufficiently closely the behaviour of the real system that it represents. For this reason, testing of a given simulation model should, ideally, be carried out by people who had no part in the earlier development of that model.

Although many techniques for verification and validation are now available, current practices in many organizations still leave much to be desired. The influence of those involved in educating the next generation of engineers, scientists and applied mathematicians is vitally important if simulation modelling techniques are going to be properly used in the future. The importance of model testing (both in term of verification and validation) must be given much more emphasis at every stage in the education of those who develop simulation models or apply modelling and simulation methods.

Systematic testing of simulation models should also be given greater attention by model developers and users in all fields of application. The proper management of models and of the complete testing process is of central importance and rigorous documentation procedures are essential. As discussed in Sect. 4.1, the outcome of the simulation model testing process is always a judgement that is based on available evidence. When new evidence is obtained about the behaviour of the real system the corresponding simulation model should, ideally, be re-tested. In the author's opinion, development of trust in simulation methods depends more on the general acceptance and widespread use of already-proven and systematic processes of verification and validation than on research that may lead to new simulation methods and software tools. Establishing whether a specific set of verification and validation tests is sufficient for a specific application must still be viewed, in some respects, as much an art as a science.

References

Anderson, J., & Papachristodoulou, A. (2009). On validation and invalidation of biological models. *BMC Bioinfomatics, 10*(1), 132.

Beck, J. V., & Arnold, K. J. (1977). *Parameter estimation in science and engineering*. New York: Wiley.

Bellman, R., & Åström, K. J. (1970). On structural identifiability. *Mathematical Biosciences, 7,* 329–339.

Bradley, R., Padfield, G. D., Murray-Smith, D. J., et al. (1990). Validation of helicopter mathematical models. *Transactions of the Institute of Measurement and Control, 12,* 186–196.

Butterfield, M. H., & Thomas, P. J. (1986). Methods of quantitative validation for dynamic system models—part 1: Theory. *Transactions of the Institute of Measurement & Control, 8,* 182–200.

Caro, J. J., Briggs, A. H., Siebert, U., et al. (2012). Modeling good research practices—overview: A report of the ISPOR-SMDM modeling good research practices task force-1. *Medical Decision Making, 32,* 667–677.

Chatfield, C. (2003). *The analysis of time series: An introduction* (6th ed.). Boca Raton: Chapman and Hall/CRC.

Cloud, D. J., & Rainey, L. B. (Eds.). (1998). *Applied modelling and simulation: An integrated approach to development and operation*. New York: McGraw-Hill.

Eddy, D. M., Hollingsworth, W., Caro, J. J., et al. (2012). Model transparency and validation: A report of the ISPOR-SMDM modeling good research practices task force-7. *Medical Decision Making, 32,* 733–743.

Frank, P. M. (1978). *An introduction to system sensitivity theory*. London: Academic.

Gábor, A., Villaverde, A. F., & Banga, J. R. (2017). Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems. *BMC Systems Biology, 11,* 54. https://doi.org/10.1186/s12918-017-0428-y.

Goodwin, G. C., & Payne, R. L. (1977). *Dynamic system identification: Experiment design and data analysis*. New York: Academic.

Gore, R., & Diallo, S. (2013). The need for usable formal methods in verification and validation. In R. Pasupthy, S.-H. Kim & A. Tolk et al. (Eds.) *Proceedings of the 2013 Winter Simulation Conference* (pp 1257–1268). Washington DC: IEEE. https://doi.org/10.1109/wsc.2013.6721513.

Grewal, M. S., & Glover, K. (1976). Identifiability of linear and nonlinear dynamical systems. *IEEE Transactions on Automatic Control, 21,* 833–837.

Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment, 32,* 135–154.

Heitmeyer, C. L. (2007). Formal methods for specifying, validating and verifying requirements. *Journal of Universal Computer Science, 13*(5), 607–618.

Hemez, F. M. (2004). The myth of science-based predictive modelling. In Proceedings foundtions'04 workshop for verification, validation and accreditation (VV&A) in the 21st century. Arizona State University, Tempe, Arizona, 13–15 October 2004. Report LA-UR-04-6829, Los Alamos National Laboratory, USA.

Heyhurst, K. L., Veerhusen, D. S., Chilenski, J. L., et al. (2001). *A practical tutorial on modified condition/decision coverage, NASA/TM-2001-210876*. Hampton, VA, USA: National Aeronautics and Space Administration, Langley Research Center.

Kaner, C., Falk, J., & Nguyen, H. Q. (1999). *Testing computer software* (2nd ed.). New York: Wiley.

Kit, E. (1995). *Software testing in the real world*. Harlow: Addison Wesley.

Kuhn, D. R., Chandramouli, R., & Butler, R. W. (2002). Cost effective use of formal methods in verification and validation. Invited paper, Presented at Foundations'02 Workshop, US Department of Defense, Laurel, Maryland, October 22–23, 2002. Retrieved from http://csrc.nist.gov/staff/Kuhn/kuhn-chandramouli-butler-02.pdf.

Murray-Smith, D. J. (2015). *Testing and validation of computer simulation models: Principles, methods and applications*. Cham: Springer.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge: Cambridge University Press.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation and confirmation of numerical models in the earth sciences. *Science, 263*(5147), 641–646.

Pace, D. K. (2004). Modeling and simulation verification and validation challenges. *Johns Hopkins APL Technical Digest, 25,* 163–172.

Pranja, S. (2006). Barrier certificates for nonlinear model validation. *Automatica, 42,* 117–126.

Raue, A., Kreutz, C., Maiwald, T., et al. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics, 25,* 1923–1929.

Schlacher, K., & Schöberl, M. (Eds.) (2011). Special issue; Modelling, analysis and control of distributed parameter systems. *Mathematical and Computer Simulation of Dynamical Systems, 17*, 1–121.

Smith, M. I., Murray-Smith, D. J., & Hickman, D. (2007). Verification and validation issues in a generic model of an electro-optic sensor system. *Journal of Defense Modeling & Simulation, 4,* 17–27.

Steinberg, S., & Roache, P. J. (1985). Symbolic manipulation and computational fluid dynamics. *Journal of Computational Physics, 57,* 251–284.

Tomović, R. (1963). *Sensitivity analysis of dynamic systems*. New York: McGraw-Hill.

The Mitre Corporation. (2014). Verification and validation of simulation models. In *Mitre systems engineering guide* (pp 461–469). Bedford: The Mitre Corporation. www.mitre.org/publications/technical-papers/the-mitre-systems-engineering-guide.

Winsberg, E. (2018). Computer simulations in science. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2018 Edition), forthcoming. https://plato.stanford.edu/archives/sum2018/entries/simulations-science/.

Chapter 5
Errors and Uncertainties: Their Sources and Treatment

Christopher J. Roy

Abstract There are numerous sources of error and uncertainty in modeling and simulation. Some of these sources arise because of inherent randomness existing in the system of interest, while others arise due to incomplete knowledge on the part of the person conducting the modeling and simulation activity. Other sources arise due to the fact that all models are imperfect reflections of reality. Finally, when models are sufficiently complex to require approximate numerical solutions (for example, when they take the form of partial differential equations), then the numerical approximations provide an additional source of error and uncertainty. This chapter discusses these different sources of error and uncertainty as well as methods to characterize and treat them. Techniques for rolling up these different uncertainty sources into a total prediction uncertainty are briefly discussed.

Keywords Model · Simulation · Error · Uncertainty · Validation · Calibration · Prediction

5.1 Introduction

This chapter is concerned with identifying the sources of error and uncertainty in modeling and simulation. Mathematical *models* are used to describe the behavior of systems in the natural and social sciences as well as in engineering. In some cases, the model may be sufficiently simple that its solution may be obtained directly (i.e., without any numerical approximation), as in the case of a regression fit of experimental data. However, in most cases, these models are more complex and take the form of partial differential or integral equations, which cannot be solved directly. In such cases, approximate numerical solutions are needed. The focus of this chapter is on these simulations which involve discrete approximations of partial

C. J. Roy (✉)
Crofton Department of Aerospace and Ocean Engineering, Virginia Tech,
215 Randolph Hall, Blacksburg, VA 24061, USA
e-mail: cjroy@vt.edu

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_5

differential or integral equations and will thus give rise to numerical approximation errors. There are other kinds of simulations, e.g., agent-based simulations, in which no numerical approximation errors arise. The general points about modeling errors and uncertainties made in this chapter apply to them as well.

Before proceeding, it is prudent to define just what is meant by "errors and uncertainties." An *error* is defined as the difference between the value obtained and the true value, and thus has both a magnitude and a sign (Oberkampf and Roy 2010); the true value depends on the context of use. An *uncertainty* is an imprecision in a value (Oberkampf and Roy 2010). Modeling and simulation in the presence of uncertainty is considered to be *nondeterministic* in nature, where the output System Response Quantities (SRQs, i.e., the quantities the analyst in interested in predicting) no longer take the form of a single, *deterministic* value, but instead may take the form of a probability distribution, an interval, or some more general form.

In the systems under consideration, some model inputs (e.g., model parameters, boundary conditions, initial conditions, external excitations) may be uncertain. These uncertain quantities may be random in nature (i.e., *aleatory*) or due to a lack of knowledge on the part of the analyst performing the simulation (i.e., *epistemic*) (Oberkampf and Roy 2010; Roy and Oberkampf 2011). Uncertain model inputs must be propagated through the model or simulation (an activity known as *uncertainty propagation*) to determine their effects on the SRQs of interest.

Uncertainties are also present due to model imperfections and, when simulations are used, due to numerical approximations. *Validation* is the assessment of the accuracy of the mathematical model relative to observations of nature which come in the form of experimental measurements (i.e., data) (Oberkampf and Roy 2010; Roache 2009). *Verification*, on the other hand, deals with assessing the numerical accuracy of a simulation relative to the true solution to the mathematical model (Oberkampf and Roy 2010; Roache 2009; Roy 2005). Thus, verification and validation provide a means for assessing the credibility and accuracy of mathematical models and their subsequent simulations (Oberkampf and Roy 2010; Roache 2009; Roy and Oberkampf 2016; AIAA 1998; ASME 2006, 2009).

The verification and validation processes are illustrated from a deterministic point of view in Fig. 5.1. Consider that you wish to know the error in a simulation relative to the true value in nature, which is not known. To estimate the solution error, you first make some observation of nature (i.e., obtain experimental data), which will contain experimental measurement error. Choosing a mathematical model, you could compare model results to the data in order to estimate the modeling error. These first two activities would be considered part of the model validation process. If the model is not directly solvable, then if you could (hypothetically) perform the simulations on a perfect computer (with infinite speed, memory, digits of precision, etc.), any differences between the simulation and the exact solution to the mathematical model would be the due to algorithm deficiencies and software programming (i.e., coding) errors. Finally, the difference between a simulation performed on an actual computer and one performed on the hypothetical "perfect" computer would be due to the numerical approximation errors. These final two activities, which involve

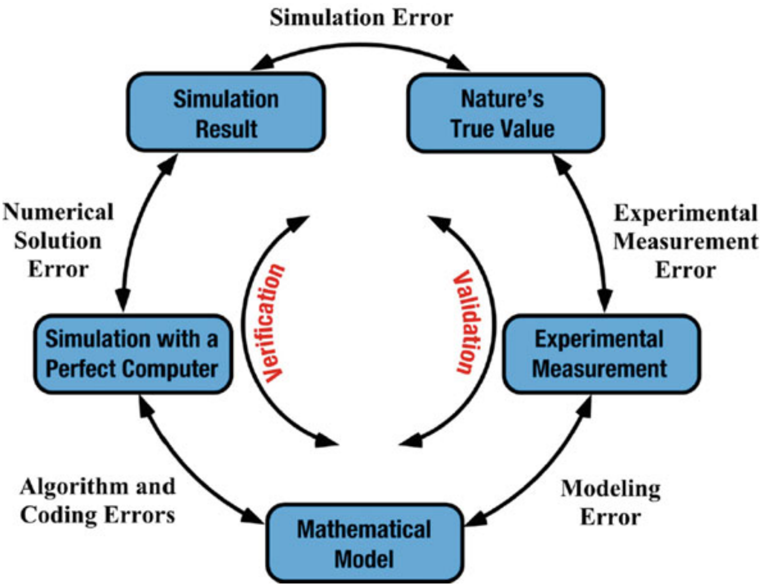


Fig. 5.1 Overview of verification and validation processes in modeling and simulation (adapted from Choudhary and Roy 2018)

the instantiation of an algorithm into software and the application of the software to produce simulation results, are different aspects of verification.

The remainder of this chapter is organized as follows. Verification-related errors are discussed in Sect. 5.2 including those associated with the discrete algorithm choice and software programming Sect. 5.2.1 and numerical approximation errors Sect. 5.2.2. The conversion of numerical errors into uncertainties is addressed in Sect. 5.2.3, while the estimation of total numerical uncertainty is discussed in Sect. 5.2.4. Section 5.3 contains a discussion of validation-related errors and uncertainties including those due to experimental measurement Sect. 5.3.1, model form uncertainty Sect. 5.3.2, and the extrapolation process Sect. 5.3.4. The issue of model calibration is addressed in Sect. 5.3.3. Section 5.4 discusses uncertainties related to the uncertainty propagation process including both model inputs Sect. 5.4.1 and model parameters Sect. 5.4.2. A technique for rolling up all uncertainty sources into a total prediction uncertainty is presented in Sect. 5.5. Finally, some points for additional discussion are presented in Sect. 5.6 and conclusions drawn in Sect. 5.7.

5.2 Verification-Related Errors/Uncertainties

For cases where the mathematical model cannot be directly solved, approximate numerical solutions (i.e., simulations) must be performed. As shown in Fig. 5.2, the

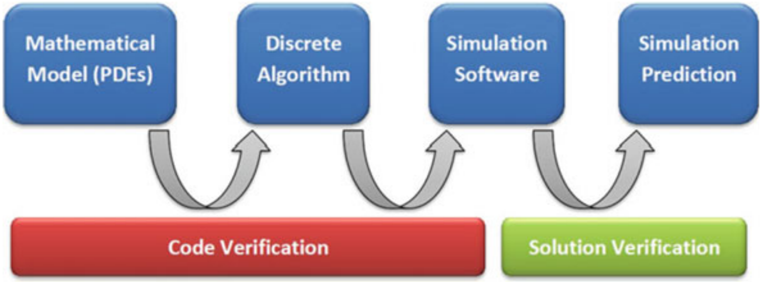


Fig. 5.2 Summary of the simulation process and associated verification activities (reproduced from Roy 2015)

steps to go from a mathematical model to a simulation result involve: (1) the choice of the discrete algorithm, (2) the programming of the chosen algorithm into software, and (3) performing the numerical simulations and estimating the numerical error. The correctness of the first two activities is assessed through *code verification*, which ensures that the simulation software (i.e., the computational model) is an accurate representation of the underlying mathematical model and its solution (Oberkampf and Roy 2010; Roache 2009; Roy 2005; Knupp and Salari 2003; Roy et al. 2004). The last activity comprises *solution verification* and deals with the estimation of the numerical errors that occur when mathematical models are discretized and solved numerically (Oberkampf and Roy 2010; Roache 2009; Roy 2005).

5.2.1 Discrete Algorithm Choice and Software Programming

The discretization process involves both the discretization of the mathematical model as well as discretization of the domain of interest. The latter process results in a mesh (usually a set of nonoverlapping cells or elements) in space and/or time, depending on whether the system is spatial in nature, temporal in nature, or both. The discrete algorithm must be selected which will ensure that (1) the discretized equations approach the mathematical model as the mesh is refined (i.e., *consistency*), (2) the numerical solutions obey the stability conditions, and (3) the numerical solution approaches the exact solution to the mathematical model with mesh refinement (i.e., *convergence*). Lax's equivalence theorem states that convergence can be guaranteed when a consistent discretization is used and the stability conditions are met (Strang 1986; Despres 2004); however, along with other caveats, this theorem only applies to linear problems.

For complicated modeling and simulation applications, the software used to instantiate the discrete algorithm may consist of thousands or even millions of lines of source code. While traditional software engineering techniques such as version control, static analysis, unit testing, regression testing, etc., are critical, they are not sufficient for ensuring the correctness of the programming of the computational model. The main difficulty is that the “correct” values for the code output SRQs are never known; they depend on the discretization scheme, the mesh, the iterative tolerance, the digits of precision, etc.

The most rigorous test of code and algorithm correctness for simulation-based codes is the *order of accuracy* test (Oberkampf and Roy 2010; Roache 2009; Roy 2005; Roy and Oberkampf 2016; Knupp and Salari 2003). This test determines whether the discrete solution produced by the code approaches an exact solution to the mathematical model at the theoretical rate (i.e., at the formal order of accuracy) as the mesh is refined. The *formal order of accuracy* is usually determined by a truncation error analysis. The *observed order of accuracy* is the actual rate at which the numerical solutions converge to the exact solution to the mathematical model with systematic refinement of the mesh and/or time step (Oberkampf and Roy 2010; Roache 2009) and can be computed as

$$p = \frac{ln\left(\frac{f_2 - \widetilde{f}}{f_1 - \widetilde{f}}\right)}{\ln(r)} \tag{5.1}$$

where f_2 and f_1 are the coarse and fine mesh SRQs, respectively, $\widetilde{f}$ is the exact solution to the mathematical model, and r is the ratio between coarse and fine grid spacing in each direction (i.e., the grid refinement factor). See Refs. (Oberkampf and Roy 2010; Roache 2009; Roy 2005; Roy and Oberkampf 2016; Knupp and Salari 2003) for more details.

Code order of accuracy testing is usually applied on simpler problems than the actual application since it (1) is testing the correctness of the discrete algorithm and the software programming and (2) requires an exact solution to the mathematical model. This exact solution can be found from traditional means for simple mathematical models or, for more complex models, it can be obtained using the method of manufactured solutions (see Chap. 10 and Refs. Oberkampf and Roy 2010; Roache 2009; Roy 2005; Roy and Oberkampf 2016; Knupp and Salari 2003; Roy et al. 2004; Roache and Steinberg 1984). The main concept behind the method of manufactured solutions is to take an original mathematical model, made up of partial differential or integral equations, and modify it by appending an analytic source term so that it satisfies a chosen (usually nonphysical) solution.

The effects of discrete algorithm and computational model programming errors on the simulation SRQs are essentially impossible to estimate. When such errors are present and affect the computed solution, then they are typically identified (via code or algorithm debugging) and removed. The code order of accuracy testing procedure described above thus provides a means of ensuring that there are no algorithm or programming mistakes that affect the numerical solutions.

5.2.2 *Numerical Approximation Errors*

This section describes the various sources of numerical approximation errors that can occur in a simulation and suggests methods for quantifying these errors.

5.2.2.1 Round Off

Round-off errors occur due to the fact that only a finite number of significant figures can be used to store floating-point numbers in a digital computer. While round-off errors are usually small, it is their accumulated effects on the SRQs after repeated arithmetic operations that is of interest. They can be reduced if necessary by increasing the number of significant figures used in floating-point computations (e.g., by changing from single to double precision arithmetic). Round-off error can be estimated by repeating the simulation with higher precision arithmetic (Oberkampf and Roy 2010; Roy and Oberkampf 2016). For example, if the underlying simulation is performed with single precision floating-point arithmetic, then round-off error can be estimated as

$$\varepsilon_{\text{Round - Off}} \cong f_{Single} - f_{Double} \tag{5.2}$$

where f_{Single} is the single precision SRQ and f_{Double} is the double precision SRQ. Note that both simulations must use the same mesh, and iterative error (if present) must be reduced well below the round-off error.

5.2.2.2 Iteration

Although not all simulations have iterative convergence error, it can be present when discretization of the mathematical model results in a simultaneous set of algebraic equations that are solved approximately or when relaxation techniques are used. The *iterative error* can be defined as the difference between the current approximate solution to the discretized equations and the exact solution to the discretized equations (Oberkampf and Roy 2010; Roache 2009; Roy and Oberkampf 2016). For a SRQ f, we can thus define the iterative error at iteration k as

$$\varepsilon_h^k = f_h^k - f_h \tag{5.3}$$

where h refers to the discrete solution on a mesh with discretization parameters (Δx, Δy, Δt, etc.) represented collectively by h, f_h^k is the current iterative solution, and f_h is the exact solution to the discrete equations (not to be confused with the exact solution to the mathematical model $\tilde{f}$).

The iterative convergence of a simulation is generally assessed by examining the iterative residuals. The *iterative residual* is found by substituting the current iterative solution into the discrete form of the governing equations and taking a norm of the resulting nonzero remainder. Although monitoring the iterative residuals often serves as an adequate indication as to whether iterative convergence of the solution has been achieved, it does not by itself provide any guidance as to the magnitude of the iterative error in the SRQ of interest.

The iterative residual norms have been shown to follow closely with actual iterative errors for many problems, usually differing only by a scaling constant (Oberkampf and Roy 2010; Roy 2005; Roy et al. 2004). Thus, a small number of cases should be sufficient to determine how the iterative errors in the SRQ scale with the iterative residuals for the range of cases of interest. During this process, the iterative error is usually estimated by iteratively converging down to "machine zero", i.e., the point where the iterative error can no longer be reduced due to the presence of round-off error. See Refs. (Oberkampf and Roy 2010; Roy and Oberkampf 2016) for more details.

5.2.2.3 Discretization

The *discretization error* is the difference between the exact solution to the discretized equations and the exact solution to the mathematical model (Oberkampf and Roy 2010; Roache 2009; Roy 2005; Roy and Oberkampf 2016). It arises due to the fact that the spatial domain is decomposed into a finite number of nodes, volumes, or elements and, for time-dependent systems, time is advanced with a finite time step. The discretization error is difficult to estimate for complex simulations and is often the largest of the numerical error sources. As shown in Fig. 5.3, methods for estimating discretization error can be broadly categorized as either recovery methods or residual-based estimators (Oberkampf and Roy 2010). Recovery methods involve post-processing of the solution(s) and include Richardson extrapolation (Oberkampf and Roy 2010; Roache 2009; Roy 2005), order extrapolation (Oberkampf and Roy 2010), and recovery methods from finite elements (Zienkiewicz and Zhu 1992; Ainsworth and Oden 2000). Residual-based methods employ additional information about the problem being solved and include discretization error transport equations (Oberkampf and Roy 2010; Zhang et al. 2000; Shih and Williams 2009; Roy 2009; Phillips and Roy 2011), defect correction methods (Skeel 1986; Stetter 1978), implicit/explicit residual methods in finite elements (Oberkampf and Roy 2010; Ainsworth and Oden 2000; Stewart and Hughes 1998; Cao 2005), and adjoint methods for estimating the error in solution functionals (i.e., SRQs) (Ainsworth and Oden 2000; Pierce and Giles 2000; Venditti and Darmofal 2000, 2003). The recovery methods have the drawback of requiring multiple mesh levels (Richardson extrapolation), requiring multiple solutions with different order of accuracy (order extrapolation), or providing accurate error estimates for only a limited class of problems (finite element recovery methods). The residual methods all require an additional solution to be computed (generally on the same mesh) and may provide more accurate error estimates since they use additional information about the problem being solved.

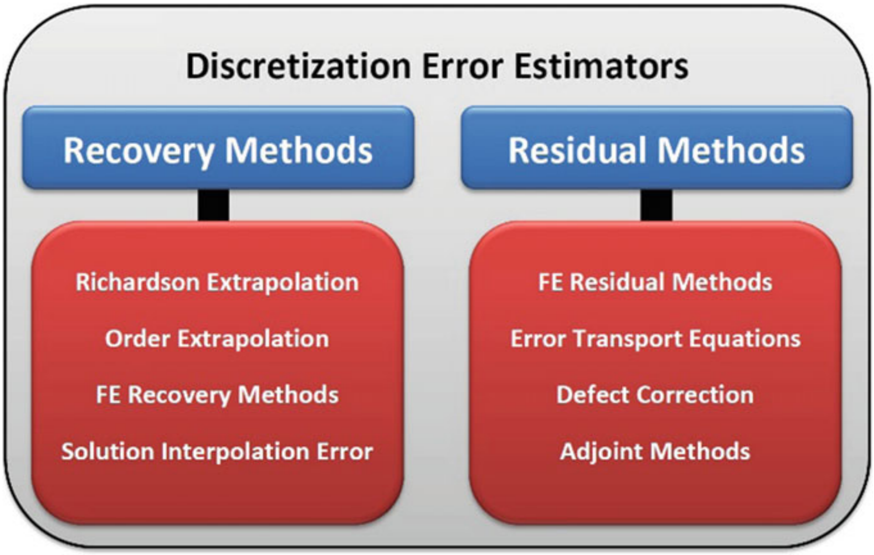


Fig. 5.3 Overview of discretization error estimation approaches (reproduced from Roy and Oberkampf 2016)

For the purposes of this chapter, we will provide high level details of the most broadly applicable recovery method, *Richardson extrapolation*, which uses solutions on two (or more) systematically refined meshes to estimate the exact solution to the mathematical model. This estimate of the exact solution to the mathematical model can, in turn, be used to provide an error estimate for the numerical solutions. Consider two systematically refined meshes with spacing h and rh, respectively. Assuming that the solutions are in the asymptotic range (i.e., that the observed order of accuracy is near the formal order), one may obtain for an estimate $\overline{f}$ of the exact solution to the mathematical model $\tilde{f}$ to be

$$\overline{f} = f_h + \frac{f_h - f_{rh}}{r^p - 1} \tag{5.4}$$

which is generally a $(p + 1)$-order accurate estimate of the exact solution to the mathematical model $\tilde{f}$. This equation can be used to estimate the discretization error in the fine grid solution, i.e., $\overline{\varepsilon}_h = f_h - \overline{f}$, resulting in the error estimate:

$$\overline{\varepsilon}_h = \frac{f_{rh} - f_h}{r^p - 1}. \tag{5.5}$$

Note that in addition to the assumption that both solutions are in the asymptotic range, this error estimate will be accurate only when the other numerical error sources (e.g., due to iteration and round off) are much smaller than the fine grid discretization error (a factor of 100 smaller is recommended Oberkampf and Roy 2010).

5.2.2.4 Surrogate Models

When simulations are expensive and involve the propagation of uncertainty from the model inputs to the SRQs, surrogate (or response surface) models are often used to approximate the simulation behavior, but at a much lower cost (Queipo et al. 2005). Surrogate models thus provide a low-cost method for mapping the uncertain input parameters to the SRQs, but are usually only feasible for a modest number of input dimensions (on the order of 10). There are numerous approaches to surrogate modeling including multidimensional polynomial curve fitting, least squares, kriging, and machine learning. A sampling strategy is needed to obtain the training data from the simulation. Some surrogate modeling approaches satisfy the training data points exactly (e.g., curve fitting), while other approaches (e.g., least squares) will only match the training data in an approximate sense. Cross validation is often used to partition the available simulation results into training data and "validation" data in order to estimate the accuracy of the surrogate model (Geisser 1993); however, such procedures can be risky unless one can establish independence between the training data and the validation data. In the current context, the assessment of the accuracy of the surrogate model is in fact a verification (i.e., mathematics) activity since it seeks to characterize the error between the simulation output and the surrogate model.

5.2.3 Conversion of Numerical Errors into Uncertainties

In some cases, when numerical errors can be estimated (both sign and magnitude) with a high degree of confidence, they can be removed from the numerical solution, a process similar to that used for well-characterized bias errors in an experiment (Oberkampf and Roy 2010; Roy and Oberkampf 2011, 2016). More often, however, the numerical errors are estimated with significantly less certainty, for example, only a rough estimate of the absolute value of the error may be available. As a result, these error estimates should be treated as numerical uncertainties, with the uncertainty coming from the error estimation process itself. One of the simplest methods for converting an error estimate to an uncertainty is to use the magnitude of the error estimate to apply uncertainty bounds about the simulation prediction, possibly with an additional factor of safety included. For example, the Richardson extrapolation estimate of discretization error $\bar{\varepsilon}_h$ discussed above can be represented as a numerical uncertainty U_{DE} as,

$$U_{\mathrm{DE}} = F_s |\bar{\varepsilon}_h| \tag{5.6}$$

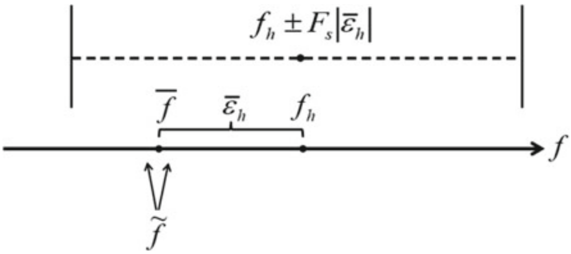


Fig. 5.4 Example of converting a discretization error estimate into a numerical uncertainty centered about the numerical solution (adapted from Roy and Balch 2012)

where $F_s \geq 1$ is the factor of safety (this is in fact a generalization of Roache's grid convergence index, see Roache 2009, 1994). The resulting interval for the numerical solution, accounting for numerical uncertainties, can be approximated by applying this uncertainty symmetrically about the fine grid solution

$$f_h \pm U_{\mathrm{DE}} = f_h \pm F_s|\bar{\varepsilon}_h|. \tag{5.7}$$

These concepts are shown graphically in Fig. 5.4 with a factor of safety of approximately $F_s = 1.5$ (Roy and Balch 2012). The numerical solution f_h has a signed error estimate $\bar{\varepsilon}_h$ as well as an uncertainty band created by taking plus/minus the absolute value of $\bar{\varepsilon}_h$ centered on the numerical solution. The factor of safety is needed because $\bar{f}$ is only an approximation of $\tilde{f}$. In other words, even when the error estimate is reasonably accurate, the true exact solution $\tilde{f}$ could still be slightly larger or slightly smaller than the estimated exact solution $\bar{f}$. When the error estimate $\bar{\varepsilon}_h$ is poor, this heuristic approach is designed to still potentially provide conservative numerical uncertainty estimates, depending of course on the chosen factor of safety. It is recommended that this uncertainty be centered about the numerical solution f_h rather than the estimated exact solution $\bar{f}$ since the latter can lead to erroneous (and possibly physically non-realizable) values.

5.2.4 Estimating Total Numerical Uncertainty

When multiple sources of numerical error are present, then a conservative approach is to simply add the numerical uncertainties together (Oberkampf and Roy 2010; Roy and Oberkampf 2011), e.g.,

$$U_{NUM} = U_{RO} + U_{IT} + U_{DE} + U_{SURR}. \tag{5.8}$$

where U_{RO}, U_{IT}, U_{DE}, and U_{SURR} refer to uncertainty due to round off, iteration, discretization, and surrogate modeling, respectively. While this method assumes that the uncertainty sources are independent, it is guaranteed to be a conservative bound on the numerical uncertainty if the uncertainty estimate for each of the components is conservative (i.e., the uncertainty bounds the true error). Numerical uncertainties are epistemic (i.e., due to a lack of knowledge rather than inherent randomness) in nature since they can be reduced by adding additional information (e.g., more digits of precision, more iteration, finer meshes, more training points).

5.3 Validation-Related Errors/Uncertainties

There are many ways to use experimental data for uncertainty quantification and reduction in computational modeling. The most common way is through calibration of model parameters (including possibly a model error/discrepancy term). While calibration may result in an improved model, the errors and/or uncertainties associated with the newly calibrated model are generally not known. At the other end of the spectrum, one could use the data to estimate the uncertainty in the original (un-calibrated) model. In either case, the uncertainty of the experimental data should be taken into account. Finally, techniques for inferring the modeling uncertainty at conditions where no data are available (i.e., extrapolation) must also be considered (see Chap. 3 by Oberkampf and Chap. 15 by Murray-Smith in this volume). This section describes these various validation-related errors and uncertainties and their treatment.

5.3.1 Experimental Measurement

There are two types of measurement errors: random measurement errors and systematic (or bias) errors. The experimental uncertainty due to random error sources can be reduced by adding additional replicate measurements, with the uncertainty scaling as $1\Big/\sqrt{N}$, where N is the number of experimental replicates. Bias errors, when estimated accurately, can be removed from the measurement via experimental calibration procedures. Unknown or estimated bias errors are usually converted to random errors via design of experiments (Montgomery 2017) or other blocking techniques (Oberkampf and Smith 2017).

The uncertainty in an experimental measurement is the root sum square of the standard systematic uncertainty and the random uncertainty. See Refs. (ASME PTC 2005; ISO 1995; Coleman and Steele 2009) for details. Reported experimental uncertainty usually refers to some confidence level on the mean value. For example, a measurement reported with 10% uncertainty generally means that the true value (i.e., the actual value found in nature) is within the stated interval (reported value ±10%) with 95% confidence.

5.3.2 Model Validation

Model error arises due to all of the assumptions and approximation that occur during the model development process. Model error is often one of the largest contributors to the overall uncertainty in modeling and simulation (in the words of statistician George Box "All models are wrong but some are useful" (Box 1979)). If the modeling error can be accurately characterized over the entire application domain, then it should be removed via calibration procedures (see Sect. 5.3.3 below); however, this is rarely the case. It is much more common that the modeling error can only be estimated in a rough sense, and only in a small portion of the application domain. In such cases, it is more appropriate to try to estimate the *Model Form Uncertainty* (MFU), not to be confused with the uncertainty due to the value of the model parameters.

One approach for treating MFU is to use all experimental data to quantify the MFU and make no attempt to improve (or calibrate) the model. This process is called *model validation* (Oberkampf and Roy 2010; Roache 2009), model accuracy assessment, or the estimation of MFU and is discussed in more detail below. Another approach for dealing with MFU was developed by Kennedy and O'Hagan (Kennedy and O'Hagan 2000, 2001) and has the advantage of fitting within a Bayesian framework. In their approach, the MFU is quantified by parameterizing the difference between the outputs of the computational model and experimental observations as a stationary Gaussian process, whose hyperparameters can either be assumed a priori or inferred from the data. The Gaussian process model provides the estimated model discrepancy (i.e., model error) along with a Gaussian uncertainty, which is generally small where data are available (depending on the uncertainty of the data themselves) and grows in regions of the parameter space where data are lacking. Thus their approach inherently incorporates elements of validation, calibration (see Sect. 5.3.3), as well as extrapolation (see Sect. 5.3.4). Note that the Kennedy and O'Hagan approach does not result in a true validation metric (as described below) since it does not provide a true distance measure (see Chap. 7 by Beisbart in this volume).

Validation metrics provide a means by which the accuracy of a model can be assessed relative to data (Oberkampf and Roy 2010; Roy and Oberkampf 2011, 2016; Ferson et al. 2008). Liu et al. (2011) proposed a classification system for validation metrics based on whether or not (1) the metric incorporates uncertainty sources in the simulation predictions and the experimental measurements (i.e., the metric is classified as either deterministic or stochastic), (2) the comparison is made for a single SRQ or multiple SRQs, and (3) the metric provides a quantitative distance-based measure that can be used to quantify modeling error/uncertainty. Note that the latter criterion is related to the general requirements for a mathematical metric (see Chap. 13 by Marks in this volume).

While there are many possible validation metrics, we will discuss one approach called the area validation metric (Ferson et al. 2008) which is a true mathematical metric that provides quantitative assessment of disagreement between a stochastic model SRQ and experimental measurements. When probabilistic uncertainties are present in the model inputs, propagating these uncertainties through the model allows the evaluation of a *Cumulative Distribution Function* (CDF) of the SRQ. Experimental measurements are then used to construct an empirical CDF of the SRQ. The absolute value area between these two CDFs is referred to as the area validation metric d (also called the Minkowski L_1 norm) and is given by

$$d(F, S_n) = \int_{-\infty}^{\infty} |F(x) - S_n(x)| \, dx \tag{5.9}$$

where $F(x)$ is the CDF from the simulation, $S_n(x)$ is the empirical CDF from the experiment, and x is the SRQ. The area validation metric d has the same units as the SRQ and goes to zero when the experimental and model CDFs are identical, thus providing a measure of the *evidence for disagreement* between the two (Ferson et al. 2008). This metric represents an epistemic uncertainty because it embodies the bias effect of all of the assumptions and approximations in the formulation of the mathematical model compared to measurements of the SRQ in nature. The area validation represents the MFU and is usually applied as an interval symmetrically about the simulation outcome $F(\mathrm{x})$ as $F(x \pm d)$. Note that the area validation metric can also contain sampling (epistemic) uncertainty due to a finite number of experimental measurements, or a finite number of computational samples.

An example of the area validation metric for a case with aleatory uncertainties occurring in the model input parameters is given in Fig. 5.5. In this figure, the aleatory uncertainties have been propagated through the model (e.g., with a large number of Monte Carlo samples), but only four experimental replicate measurements are available. The stair-steps in the experimental CDF are due to the different values observed in each of the four experimental measurements and are separated by cumulate probabilities of 0.25. The stochastic nature of the measurements can be due to variability of the experimental conditions and/or random measurement uncertainty. This metric can also be computed for cases involving both aleatory and epistemic (including interval-characterized) uncertainty in the model inputs, as well as situations where only a small number of simulation samples are available (e.g., see Oberkampf and Roy 2010; Ferson et al. 2008).

5.3.3 Model Calibration

It is important to draw a clear distinction between the concepts of validation and calibration (see Chap. 41 by Frisch in this volume). While validation involves the quantitative assessment of a model relative to experimental data, *calibration* (a.k.a., parameter estimation, parameter optimization, or model updating) involves the adjustment of model input parameters to improve agreement with experimental data. For example, if all uncertain model inputs are probabilistic, then Bayesian updating can be used to update the probability distributions of the model inputs. While calibration

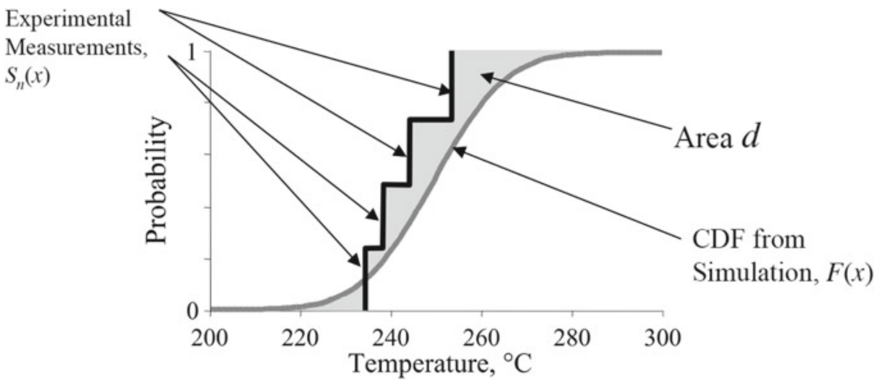


Fig. 5.5 Area validation metric example (reproduced from Ferson et al. 2008)

may be an important part of the model building and improvement process, it does not in itself provide quantitative estimates of MFU. The key difference is that model calibration results in a modified model, because model parameters or their distributions are updated, that must still be assessed for accuracy when new experimental data become available.

5.3.4 Extrapolation

Extrapolation is the use of a model to make predictions beyond conditions where experimental data are available. Extrapolation is appropriate when: (1) it is performed using a physics-based (or first principles) model and (2) it is believed that the physics-based model includes all of the necessary physics involved at the prediction conditions. Extrapolation should not be performed with regression-based or empirical models, just as one should not apply a polynomial curve fit far outside of the conditions where data are available. Thus, extrapolation should be a physics-based process as opposed to a statistical process.

The conditions where the model will be applied is considered the application domain. The conditions where experimental data are available are called the validation domain. Figure 5.6 presents a simple example of one possible relationship between the validation domain (represented by the letters "V") and the application domain. Here, the application domain is determined by two (ideally nondimensional) input parameters, and the validation domain is contained entirely within the application domain, but does not fully cover it. The validity of the model may be directly assessed, using validation metrics along with an interpolation strategy, within the validation domain; however, the validity of the model outside the validation domain must be inferred from model assessments made within the validation domain. The validation domain is generally not coincident with the application domain, thus extrap-

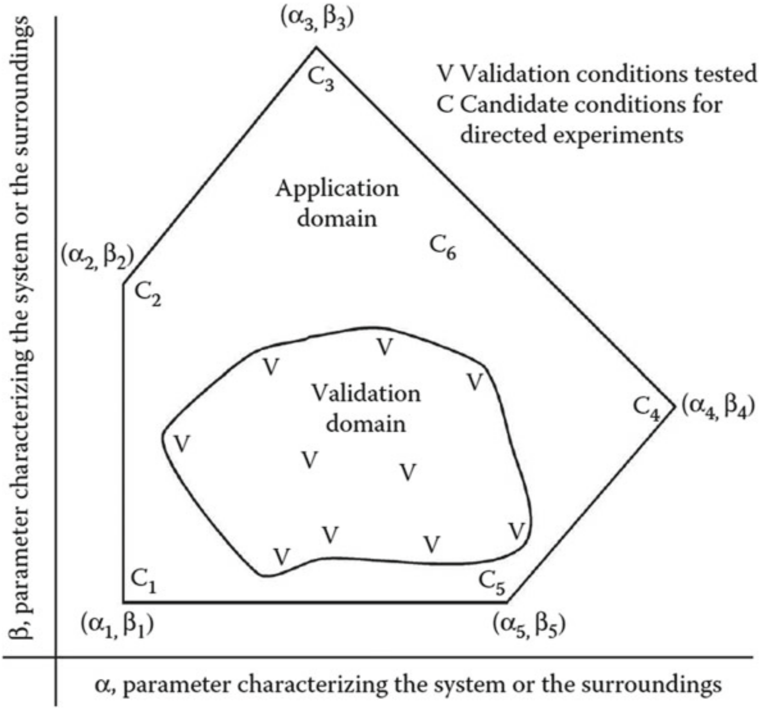

(α3, β3)
C3
V Validation conditions tested
C Candidate conditions for directed experiments
Application domain
C6
(α2, β2)
C2
V
V
V
Validation domain
C4
(α4, β4)
V
V
V
V
V
V
V
C1
C5
(α1, β1)
(α5, β5)
β, parameter characterizing the system or the surroundings
α, parameter characterizing the system or the surroundings

Fig. 5.6 Schematic showing one possible relationship between the validation domain and the application domain; many other set relationships are of course possible (reproduced from Oberkampf and Roy 2010)

olation of the MFU to the conditions of interest is needed. The key point is that the statistical extrapolation occurs in the estimation of the MFU, not in the simulation prediction itself. When the application domain is defined by a large number of parameters (i.e., a high-dimensional space), even copious amounts of experimental data generally only cover a small portion of the application domain. The Kennedy and O'Hagan model discrepancy approach (Kennedy and O'Hagan 2001) provides a natural means of extrapolation of the model discrepancy from the validation domain to the application domain. In other cases, extrapolation of the MFU can be achieved statistically using prediction intervals (Roy and Oberkampf 2011; Roy and Balch 2012).

5.4 Uncertainty Propagation-Related Uncertainties

For nondeterministic simulations, when model inputs and model parameters are uncertain, then they must be propagated through the model to assess their effects on

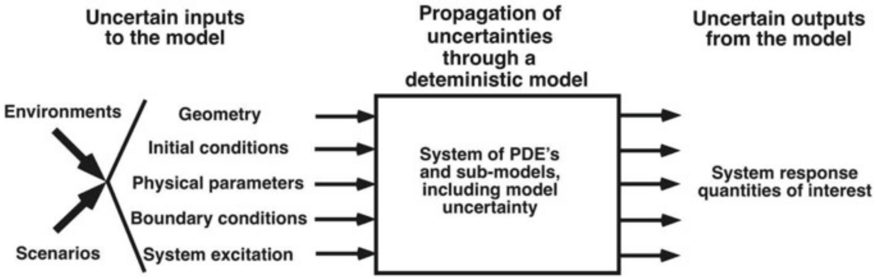

Uncertain inputs to the model
Propagation of uncertainties through a deteministic model
Uncertain outputs from the model
Environments
Scenarios
Geometry
Initial conditions
Physical parameters
Boundary conditions
System excitation
System of PDE's and sub-models, including model uncertainty
System response quantities of interest

Fig. 5.7 Schematic showing the general nondeterministic simulation process (reproduced from Oberkampf and Roy 2010)

the model outputs (i.e., the SRQs). This *uncertainty propagation* process is shown graphically in Fig. 5.7, where the model itself is assumed to be deterministic, but model inputs and parameters may be uncertain. As discussed in the previous sections, additional uncertainty may occur in the SRQs due to MFU as well as numerical uncertainty (when simulations are involved). This section describes the sources of uncertainty and their treatment related to the uncertainty propagation process.

5.4.1 Model Inputs

Model inputs and model parameters may either be deterministic (i.e., have a single, known value) or uncertain. When they are uncertain, they can be classified as (1) *aleatory*—the inherent variation in a quantity, (2) *epistemic*—uncertainty due to lack of knowledge, or (3) a mixture of the two (Roy and Oberkampf 2011). Aleatory uncertainty is generally characterized probabilistically by either a probability density function or a CDF, the latter being simply the integral of the probability density function from minus infinity up to the value of interest. A purely epistemic uncertainty is usually characterized either probabilistically as a uniform distribution (note that although this is the least informative distribution, it is still a precise probability distribution) or as an interval with no associated probability distribution. The interval characterization is a weaker statement about the value of a quantity than a uniform probability because any value in the interval is considered possible and no likelihood is ascribed to any one value over another. Mixed aleatory and epistemic uncertainty can be characterized by set-theoretical approaches such as probability bounds analysis, evidence theory, and fuzzy probabilities. These approaches are part of imprecise probability theory which characterizes the uncertainty as a set of possible probability distributions that could exist. See Chap. 21 by Bradley in this volume and Refs. (Ferson and Ginzburg 1996; Beer et al. 2013; Ferson and Hajagos 2004) for more details.

When uncertain model inputs are characterized probabilistically, there are a number of different approaches for propagating input uncertainty through the model. The simplest approach is sampling (e.g., Monte Carlo and Latin hypercube) where inputs are sampled from their probability distribution, propagated through the model, and then used to generate a sequence of SRQs. However, sampling methods tend to converge slowly: Monte Carlo methods converge at a rate proportional to $1\big/\sqrt{N}$ and Latin hypercube sampling converges as $1\big/N^{3/2}$ for small sample size N. Other approaches that can be used to propagate probabilistic uncertainty include perturbation methods and stochastic spectral methods (e.g., polynomial chaos), the latter of which comes in both intrusive (i.e., requiring modifications to the computational model software) and nonintrusive (i.e., employing the code as a black box) formulations (Smith 2013). Furthermore, when a surrogate model of an SRQ as a function of the uncertain model inputs is available (see Sect. 5.2.2.4), then any nonintrusive method discussed above, including sampling, can be computed efficiently.

When *all* uncertain inputs are characterized by intervals, i.e., they are purely epistemic, there are two popular approaches for propagating these uncertainties through the model to the SRQs. The simplest is sampling over the input intervals in order to estimate the interval bounds of the SRQs. However, the propagation of interval uncertainty can also be formulated as a constrained optimization problem: given the possible interval range of the inputs, determine the resulting minimum and maximum values of the SRQs. Thus, standard approaches for constrained optimization such as local gradient-based searches and global search techniques can be used. For mixed probabilistic and interval-characterized uncertainties, a segregated approach to uncertainty propagation is recommended. See Refs. (Roy and Oberkampf 2011; Ferson and Ginzburg 1996) for details.

When nonintrusive approaches are used, the errors associated with using a finite number of samples should be estimated. In some cases, confidence intervals can be used to assess the accuracy of the output distributions. Although sampling uncertainty is epistemic in nature (as more samples can be added to reduce it), it is a special case where the epistemic uncertainty can be appropriately characterized probabilistically.

5.4.2 Model Parameters (i.e.., Parametric Uncertainty)

Uncertainty in model parameters can be propagated in the same way as uncertainties in model inputs. However, when experimental data are available, model parameters may also be treated by calibration processes (see Sect. 5.3.3). Distinctions can be made between different types of model parameters.

1. Measurable properties of the system can be estimated/calibrated independently from the system.
2. Physical modeling parameters are those not measurable outside of the context of the model.

3. Ad hoc model parameters are corrections applied to the SRQs directly to “correct” the model relative to data.

Calibration is recommended for measurable properties, but becomes increasingly hard to justify as one moves towards the ad hoc model parameters. As discussed earlier, calibration may be for deterministic values (i.e., parameter estimation) or probability distributions (via Bayesian updating).

5.5 Total Prediction Uncertainty

When performing modeling and simulation analyses, the analyst is interested in estimating the total uncertainty in their predictions. As described in detail above, the total prediction uncertainty has contributions from the uncertain model inputs (uncertainty propagation), from the modeling process (validation), and, in some cases, from the simulation process (verification). These three contributors to the total prediction uncertainty are shown schematically in Fig. 5.8. Model inputs may be deterministic, aleatory, epistemic, or mixed aleatory/epistemic, and are propagated through the model (or simulation) to determine their effects on the SRQs. Code and solution verification activities provide estimates of the (epistemic) numerical uncertainties associated with the simulations. Finally, validation activities are conducted by computing validation metrics which compare simulation and experimental outcomes to estimate the (epistemic) MFU at the validation conditions. The MFU is then extrapolated to the prediction conditions of interest.

Consider an example (Roy and Balch 2012) where there is both aleatory and epistemic uncertainty in the model inputs. The aleatory uncertainty is characterized probabilistically and the epistemic uncertainty is characterized as an interval. These uncertainties are propagated through the model using segregated uncertainty propagation (Roy and Oberkampf 2011; Roy and Balch 2012), resulting in a p-box for the SRQ (in this case, thrust produced by a rocket nozzle) as indicated by the blue shaded region in Fig. 5.9. This p-box represents the family of all possible CDFs that can exist within its bounds. The outer bounding shape of the p-box is due to the probabilistically characterized input uncertainty and the width of the p-box is due to the interval-characterized input uncertainty.

If the p-box of the SRQ resulting from propagating both the random and the interval-characterized model input uncertainties through the model is denoted by $F(x)$, then accounting for MFU (U_{MODEL}) and numerical uncertainty (U_{NUM}) would result in an extended p-box $F(x \pm U_{TOTAL})$ where $U_{TOTAL} = U_{MODEL} + U_{NUM}$. That is, the left side of the initial p-box resulting from the propagation of input uncertainties is displaced negatively by U_{TOTAL}, and the right side of the p-box is displaced positively by U_{TOTAL}, to obtain the extended p-box for the SRQ. In Fig. 5.9, the contribution from MFU is shown in green and that from numerical uncertainty is shown in red.

The extended p-box can be interpreted as follows. Consider a requirement that the thrust produced by the rocket nozzle must be at least 2,600 N. After accounting for

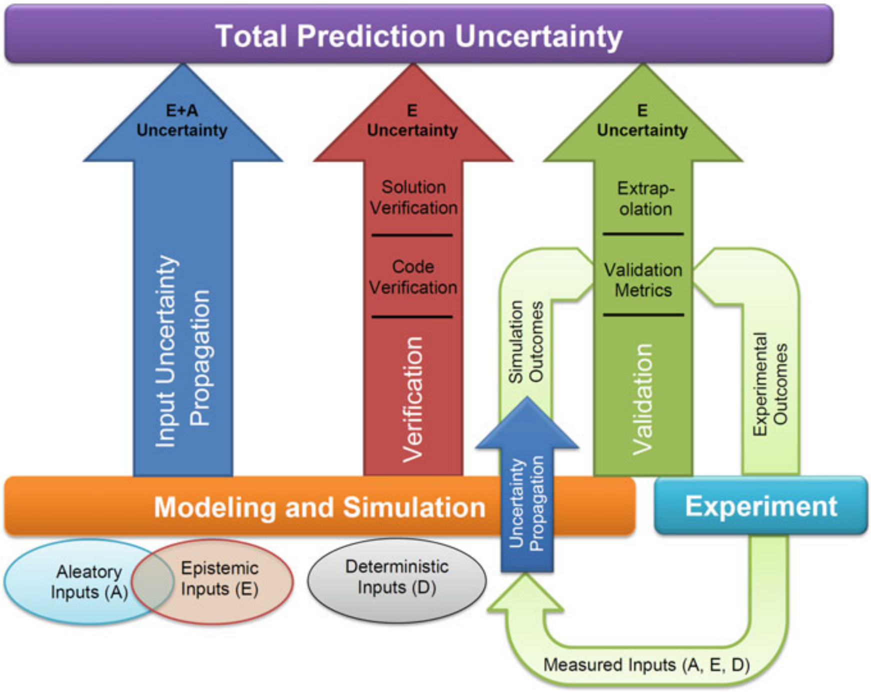


Fig. 5.8 Overview of the sources contributing to the total prediction uncertainty (adapted from Roy and Balch 2012)

both MFU and numerical uncertainties, the probability that this requirement could be violated is in the interval range [0%, 22%]. That is, it could be as low as a 0% or as high as a 22% chance that the thrust will be below the required value. This interval range represents the lack of knowledge (i.e., ignorance) that the analyst has about the system, its inputs, and the modeling and simulation process. The prudent decision maker would realize that the probability of violating the requirement could be as high as 22% and would look for ways the epistemic uncertainties could be reduced. It should be noted that if numerical uncertainty and MFU were ignored, then the blue p-box of Fig. 5.9 indicates that the thrust requirement will be met with nearly 100% probability.

In this notional example, there are three ways to reduce the epistemic uncertainty in the prediction. First, additional information could be gathered to reduce the epistemic uncertainties in the model inputs, thus reducing the width of the initial p-box. Second, the numerical uncertainty could be reduced by performing the simulations on a finer mesh, doing more iterations, etc. Finally, the MFU could be reduced by gathering additional experimental data near the prediction conditions (reducing uncertainty due to extrapolation) or by improving/calibrating the model at the conditions where data are available. The choice of which approach to take to

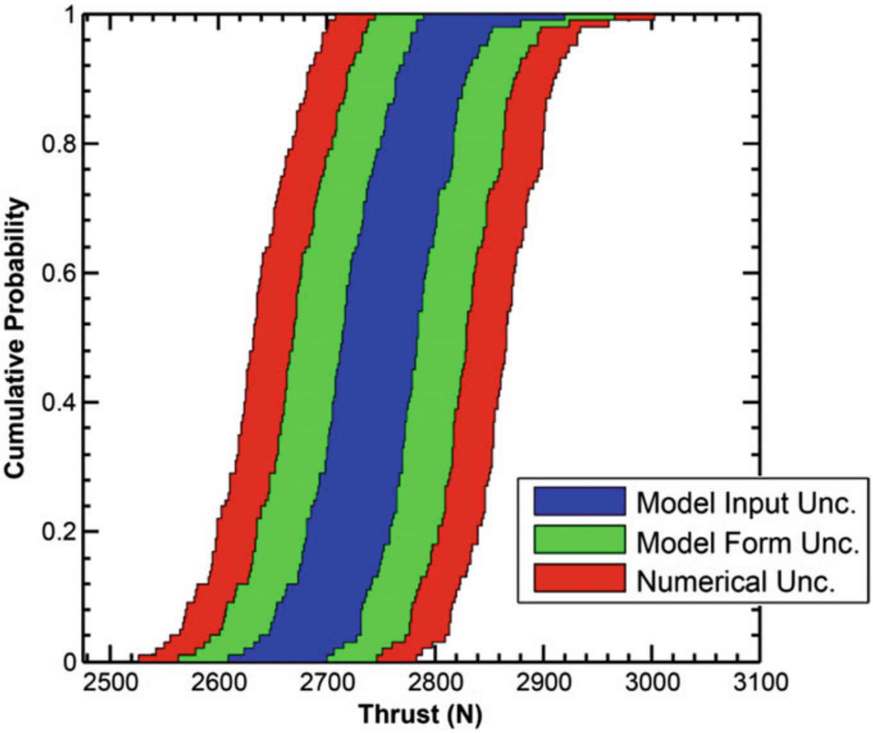


Fig. 5.9 Example of total prediction uncertainty represented as an extended p-box (reproduced from Roy and Balch 2012)

reduce the epistemic uncertainty would depend on the cost and time associated with each of these uncertainty reduction strategies.

5.6 Discussion

There are two primary issues that merit additional discussion. The first has to do with how uncertainties from the various sources are aggregated to obtain the total prediction uncertainty. The second issue revolves around how experimental data are used: for improving the model (calibration), estimating the uncertainty in the model (validation), or a mixture of the two.

When numerical errors are converted to uncertainties as discussed in Sect. 5.2.3, they are generally centered about the simulation value, even when they arise from signed error estimates (e.g., see Roache 1994). To fit more naturally in a probabilistic framework, one possibility would be to correct the solution with the estimated bias error, then place the uncertainty about the corrected solution; however, this should only be done when the error estimates are deemed reliable. Another approach would be to treat the uncertainty as an interval about either the simulation value (as discussed above in Sect. 5.5) or the corrected solution. In general, it is an open question as to whether the numerical uncertainty sources should be characterized probabilistically (Stern et al. 2001; Phillips and Roy 2013) or in some other fashion (e.g., as intervals) (Roy and Oberkampf 2011; Roy and Balch 2012).

There is still a great deal of debate on how to account for model form uncertainty (MFU) in modeling and simulation. One extreme would be to use all available experimental data to improve (i.e., calibrate) the model and thus reduce the MFU. At the other end of the spectrum, one may choose use all of the data to estimate the MFU. Rigorous assessments of the best approach for treating MFU have not appeared extensively in the literature to date. The choice of treatment for MFU likely will depend on the risk associated with how the modeling and simulation predictions will be used as well as the quantity and quality of the experimental data. For example, in preliminary system design, model calibration may be appropriate; however, for the final simulations to support the regulatory filing for a new type of nuclear power plant, estimates of MFU without calibration would carry less risk. In addition, when reliable and extensive experimental data are available over the entire application domain, then a calibration approach is feasible. However, when data are sparse and/or deemed unreliable, then a more conservative approach is recommended where the baseline model is not calibrated, but instead the MFU is estimated.

5.7 Conclusions

The various sources of error and uncertainty in modeling and simulation were described in this chapter. Verification, validation, and uncertainty propagation are the three processes used to estimate these error and uncertainty sources. It is only after each of these sources has been addressed that one can gain confidence in the predictions made using modeling and simulation. While no single approach has emerged to aggregate all of these error and uncertainty sources into a total prediction uncertainty, some possible approaches were discussed.

References

Ainsworth, M., & Oden, J. T. (2000). *A posteriori error estimation in finite element analysis*. New York: Wiley Interscience.

AIAA. (1998). Guide for the verification and validation of computational fluid dynamics simulations, American Institute of Aeronautics and Astronautics, AIAA-G-077–1998, Reston, VA.

ASME PTC 19.1-2005. (2005). Test Uncertainty.

ASME. (2006). Guide for verification and validation in computational solid mechanics, American Society of Mechanical Engineers, ASME Standard V&V 10-2006, New York, NY.

ASME. (2009). Standard for verification and validation in computational fluid dynamics and heat transfer, American Society of Mechanical Engineers, ASME Standard V&V 20-2009, New York, NY.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press.

Beer, M., Ferson, S., & Kreinovich, V. (2013). Imprecise probabilities in engineering analysis. *Mechanical Systems and Signal Processing, 37*(1–2), 4–29.

Cao, J. (2005). Application of a posteriori error estimation to finite element simulation of incompressible Navier-Stokes flow. *Computers & Fluids, 34*(8), 972–990.

Choudhary, A., & Roy, C. J. (2018). Verification and validation for multiphase flows. In G. H. Yeoh (Ed.), *Handbook of multiphase flow science and technology*. Springer. (to appear).

Coleman, H. W., & Steele, W. G. (2009). *Experimentation, validation, and uncertainty analysis for engineers* (3rd ed.). New York: Wiley.

Despres, B. (2004). Lax theorem and finite-volume schemes. *Mathematics of Computation, 73*(247), 1203–1234.

Ferson, S., & Ginzburg, L. R. (1996). Different methods are needed to propagate ignorance and variability. *Reliability Engineering and System Safety, 54,* 133–144.

Ferson, S., & Hajagos, J. G. (2004). Arithmetic with uncertain numbers: Rigorous and (often) best possible answers. *Reliability Engineering and System Safety, 85,* 135–152.

Ferson, S., Oberkampf, W. L., & Ginzburg, L. (2008). Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering, 197,* 2408–2430.

Geisser, S. (1993). *Predictive inference*. New York: Chapman and Hall. ISBN 0-412-03471-9.

ISO Guide to the Expression of Uncertainty in Measurement. (1995). ISO, Geneva, Switzerland.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B—Statistical Methodology, 63*(3), 425–450.

Kennedy, M. C., & O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika, 87*(1), 1–13.

Knupp, P. M., & Salari, K. (2003). *Verification of computer codes in computational science and engineering*. Boca Raton: Chapman and Hall/CRC. K. H. Rosen (ed.).

Liu, Y., Chen, W., Arendt, P., & Huang, H.-Z. (2011). Toward a better understanding of model validation metrics. *Journal of Mechanical Design, 133,* 1–13.

Montgomery, D. C. (2017). *Design and analysis of experiments* (9th ed.). New Jersey: Wiley.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge: Cambridge University Press.

Oberkampf, W. L., & Smith, B. L. (2017) Assessment criteria for computational fluid dynamics model validation experiments. *Journal of Verification, Validation, and Uncertainty Quantification*, *2*(3).

Pierce, N. A., & Giles, M. B. (2000). Adjoint recovery of superconvergent functionals from PDE approximations. *SIAM Review, 42*(2), 247–264.

Phillips, T. S., & Roy, C. J. (2011). Residual methods for discretization error estimation. AIAA Paper 2011–3870.

Phillips, T. S., & Roy, C. J. (2013). A new extrapolation-based uncertainty estimator for computational fluid dynamics. AIAA Paper 2013–0260.

Queipo, N. V., Haftka, R. T., Shyy, W., Goel, T., Vaidyanathan, R., & Tucker, P. K. (2005). Surrogate-based analysis and optimization. *Progress in Aerospace Sciences, 41,* 1–28.

Roache, P. J., & Steinberg, S. (1984). Symbolic manipulation and computational fluid dynamics. *AIAA Journal, 22*(10), 1390–1394.

Roache, P. J. (1994). Perspective: A method for uniform reporting of grid refinement studies. *Journal of Fluids Engineering, 116,* 405–413.

Roache, P. J. (2009). *Fundamentals of verification and validation*. Socorro, New Mexico: Hermosa Publishers.

Roy, C. J., Nelson, C. C., Smith, T. M., & Ober, C. C. (2004). Verification of Euler/Navier–stokes codes using the method of manufactured solutions. *International Journal for Numerical Methods in Fluids, 44*(6), 599–620.

Roy, C. J. (2005). Review of code and solution verification procedures for computational simulation. *Journal of Computational Physics, 205,* 131–156.

Roy, C. J. (2009). Strategies for driving mesh adaptation in CFD. AIAA Paper 2009–1302.

Roy, C. J., & Oberkampf, W. L. (2011). A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering, 200*(25–28), 2131–2144.

Roy, C. J., & Balch, M. S. (2012). A holistic approach to uncertainty quantification with application to supersonic nozzle thrust. *International Journal for Uncertainty Quantification, 2*(4), 363–381.

Roy, C. J. (2015). Verification. In B. Engquist (Ed.), *Encyclopedia of applied and computational mathematics* (pp. 1530–1537). Heidelberg: Springer.

Roy, C. J., & Oberkampf, W. L. (2016). Verification and Validation in computational fluid dynamics. In R. W. Johnson (Ed.), *Handbook of fluid dynamics* (2nd Ed). Boca Raton: CRC Press.

Shih, T. I.-P., & Williams, B. R. (2009). Development and evaluation of an a posteriori method for estimating and correcting grid-induced errors in solutions of the navier-stokes equations. AIAA Paper 2009–1499.

Smith, R. C. (2013). Uncertainty quantification: Theory, implementation, and applications. *SIAM*.

Skeel, R. D. (1986). Thirteen ways to estimate global error. *Numerische Mathematik, 48,* 1–20.

Stetter, H. J. (1978). The defect correction principle and discretization methods. *Numerische Mathematik, 29,* 425–443.

Stern, F., Wilson, R. V., Coleman, H. W., & Paterson, E. G. (2001). Comprehensive approach to verification and validation of cfd simulations—part 1: Methodology and procedures. *Journal of Fluids Engineering, 123,* 793–802.

Stewart, J. R., & Hughes, T. J. R. (1998). A tutorial in elementary finite element error analysis: A systematic presentation of a priori and a posteriori error estimates. *Computer Methods in Applied Mechanics and Engineering, 158*(1–2), 1–22.

Strang, G. (1986). *Introduction to applied mathematics.* Wellesley-Cambridge Press.

Venditti, D. A., & Darmofal, D. L. (2000). Adjoint error estimation and grid adaptation for functional outputs: Application to quasi-one dimensional flow. *Journal of Computational Physics, 164,* 204–227.

Venditti, D. A., & Darmofal, D. L. (2003). Anisotropic grid adaptation for functional outputs: Application to two-dimensional viscous flows. *Journal of Computational Physics, 187,* 22–46.

Zhang, X. D., Trepanier, J.-Y., & Camarero, R. (2000). A posteriori error estimation for finite-volume solutions of hyperbolic conservation laws. *Computer Methods in Applied Mechanics and Engineering, 185*(1), 1–19.

Zienkiewicz, O. C., & Zhu, J. Z. (1992). The Superconvergent patch recovery and a posteriori error estimates, Part 2: Error estimates and adaptivity. *International Journal for Numerical Methods in Engineering, 33,* 1365–1382.

Part II
Foundations—Validation as a Scientific Method: Philosophical Frameworks for Thinking about Validation

Chapter 6
Invalidation of Models and Fitness-for-Purpose: A Rejectionist Approach

Keith Beven and Stuart Lane

> I am….an almost orthodox adherent of unorthodoxy: *I hold that orthodoxy is the death of knowledge since the growth of knowledge depends entirely on the existence of disagreement.*
>
> (Karl Popper 1994, p. 34)

Abstract This chapter discusses the issues associated with the invalidation of computer simulation models, taking environmental science as an example. We argue that invalidation is concerned with labelling a model as not fit-for-purpose for a particular application, drawing an analogy with the Popperian idea of falsification of hypotheses and theories. Model invalidation is a good thing in that it implies that some improvements are required, either to the data, to the auxiliary relations or to the model structures being used. It is argued that as soon as epistemic uncertainties in observational data and boundary conditions are acknowledged, invalidation loses some objectivity. Some principles for model evaluation are suggested, and a number of potential techniques for model comparison and rejection are considered, including Bayesian likelihoods, implausibility and the GLUE limits of acceptability approaches. Some problems remain in applying these techniques, particularly in assessing the role of input uncertainties on fitness-for-purpose, but the approach allows for a more thoughtful and reflective consideration of model invalidation as a positive way of making progress in science.

Keywords Epistemic uncertainty · Model equifinality · Bayes · GLUE · Limits of acceptability · Behavioural models

K. Beven (✉)
Lancaster Environment Centre, Lancaster University, Lancaster, UK
e-mail: k.beven@lancaster.ac.uk

S. Lane
Institute of Earth Surface Dynamics, Université de Lausanne, Lausanne, Switzerland

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_6

6.1 Setting the Scene for Model Evaluation

In this chapter, we discuss the problems in applying a scientific methodology to computer models and, in particular, to the issue of rejection or invalidation of computer simulation models. We use invalidation of a simulation model structure and falsification of any of its component hypotheses here equivalently, to indicate that a simulation model has been shown to fail in some important respect and should consequently not be considered fit-for-purpose in making predictions. We do so from the point of view of practical environmental modellers in the domains of hydrology and hydraulics, who have an interest in the philosophical underpinnings of the modelling process and its role in the development of the associated science. Computer simulation is widely used in this domain, with (often complex) models being constructed to represent environmental systems with elements that sometimes have a good theoretical basis (e.g. mass and energy balance principles); that sometimes are derived from empirical studies (e.g. roughness relationships to represent bulk energy losses); and that sometimes have a purely conceptual basis (e.g. canopy resistance for transpiration from a vegetated surface). In such models, many of the functional relationships involve parameters that need to be identified for particular applications. These are often considered to be constant at a particular location and through time (especially when calibrating parameter values to past data) but have often been shown to change with the state of the system, or over time. Uncertainties in the input or forcing data as well as data used in model calibration, and also competing model structures, are intrinsic to the modelling process (see e.g. Beven 2009, 2012a). The dominant sources of uncertainty are often epistemic (i.e. the result of a lack of knowledge) rather than aleatory (i.e. statistical or resulting from random natural variability) in nature.

This chapter essentially addresses the question of how to do science when using models in the face of such epistemic uncertainties. This is discussed in the context of Popper's falsificationist approach in Sect. 6.2 and how this might then be applied as a rejectionist methodology in model evaluation when all models are known to be false to some extent. Sections 6.3 and 6.4 discuss the concepts of verisimilitude and fitness-for-purpose in the context of how models that are false might be useful. In Sect. 6.5, some principles of model invalidation as a positive methodology for advancing the science are discussed and it is shown how rejection can be considered within a modified Bayesian framework that either allows a choice between model structures or applies some limits of acceptability. Section 6.6 discusses how epistemic uncertainties impact on a rejectionist framework, and Sect. 6.7 how to resolve the advocacy of models that might not be fit-for-purpose with making scientific progress.

There is a long and continuing debate in both science and philosophy about what constitutes, or what should constitute, a scientific method in different domains of science (e.g. Chalmers 1976; Howson 2000; Hackett 2013). The recent rise of simulation models as a methodology for doing science has been much discussed in this respect (e.g. Cartwright 1999; Winsberg 2003). Simulation models combine elements of deductive inference in arguing from premises based on established concepts and theories, and inductive inference as a way of deriving functional relationships and parameterisations (see Young 2013 for a recent discussion in the field of hydrology), but also as a way of inferring values for the parameters of models by calibration against observational data. Perhaps reflecting the extent to which current scientific method has given primacy to observation, the parameter values required to make the model reproduce those data may either have no physical equivalent or vary from those that have been measured (see Lane et al. 2011; Lane 2012). They are "effective" in that the calibrated values are needed to make the model perform against observational data (Beven 1989, 2016). Nearly all, if not all, environmental simulation models incorporate "conceptual" or inductive elements of this type.

However, as Hume (1748) first pointed out, there is a problem with induction. It demands that the future will be the same as the past. There is then an implication that any theory or model of nature can never be verified on the basis of past observations because there is always a possibility that "*the course of nature may change*" (Hume 1748, Sect. IV.2). It has proven difficult to provide any philosophical resolution to Hume's problem of induction even though it challenges the fundamental belief that environmental simulation models are a means of getting a handle on events that have yet to happen. Indeed, Howson (2000) argues that Hume is correct, but that does not mean that we cannot use reasoned argument based on observations, as well as deductive argument, to modify our scientific understanding and beliefs about the future. Widespread use of the term "physically-based" in environmental modelling is the implicit manifestation of a faith in this form of deductive argument. The term "physically-based" suggests a simulation model is based upon assumed-to-be time-invariant "laws of nature" and so capable of better getting at the future than other kinds of approaches (e.g. belief systems; expert judgement). Of course, however, physically based a model might be, Hume's proposition means that we will sometimes get surprises as the future unfolds.

It would appear evident that the use of simulation models that involve inductive elements as either parameterisations or calibrated/effective parameter values, might be most susceptible to future surprise when the observational data used in the inference comes from the past. As modellers, we expect the future to be uncertain and past experience suggests that we should expect some element of surprise, if only because future boundary conditions cannot be known (see, for example the post-audit analysis of groundwater model simulations of Konikow and Bredehoeft 1992; Anderson and Woessner 1992, discussed later). One aim of simulation modelling is then to minimise the element of surprise by ensuring that any model used for predicting the future is fit-for-purpose, in so far as its current and past performance has been evaluated. This is the process of model evaluation or validation or, from another perspective, model invalidation or falsification of the theoretical or conceptual components of a simulation model.

6.2 The Falsification Framework of Karl Popper

We hold that the invalidation view of model evaluation is a useful alternative to model validation because of the critical role of falsification in the development of science. Popper (1959) argued that either inferring universal statements from singular or particular ones, or confirming a universal statement with particular statements could not be justified because no matter how many instances something was observed and used to justify a particular universal statement, there was always the possibility that one observation may falsify that statement. In this context, for a hypothesis or theory to be considered scientific it must be advanced a priori, be testable and have the capacity to be falsified in some way. Hypotheses or theories that cannot be falsified in this way consequently can be considered as only pseudo-scientific. Scientific method is then the process of developing hypotheses and confronting them with the available evidence. Successful hypotheses in this context are not more probably true, because obtaining more evidence does not necessarily change the probability that a hypothesis might be falsified. For this reason, Popper argued that it is better to talk of the corroboration of hypotheses, where a better corroborated hypothesis is one that has been tested more rigorously individually, widespread or for a longer period of time. Equally, the most rapid of scientific progress may be made when a long-established or well-corroborated hypothesis is shown to no longer hold.

There are distinct parallels with the notion of multiple working hypotheses (Chamberlain 1895) and the idea that it may be necessary to work with a set of potentially contradictory hypotheses. In a Popperian framework, in which hypotheses are subject to testing and potential rejection, a hypothesis is defined as admissible if it is testable. This concept can be applied to simulation models, in that any particular realisation of a model of a process or system can be considered as a working hypothesis of how that system functions (e.g. Herskowitz 1991; Beven 2002). While recognising that all models are idealisations and consequently necessarily false in some respects, models that are successful in making useful predictions over a period of time can be considered corroborated; models that not successful should be considered as invalid or not fit-for-purpose and revised by changing beliefs (Klein and Herskowitz 2007). Similar considerations apply to simulation models used for different purposes, either for testing scientific concepts or for practical applications. The criteria of invalidation might, however, be different for different types of purpose.

Popper's falsification approach to scientific inference has not been without its dissenters. Indeed, it has been suggested that falsification itself cannot be falsified and that, in many celebrated examples, theories have not been falsified, despite contradictory observational evidence being available, because of some other intrinsically attractive features (e.g. Chalmers 1976; Ladyman 2002). Certain theories cannot be rejected because it would be too costly to do so (Latour and Woolgar 1979). It has also been pointed out that experimental observations are often conditioned by the theoretical framework within which they are developed, allowing free parameters to be derived from the observations and leaving no possibility of falsification (the Duhem–Quine thesis, see Quine 1975; Chalmers 1976). It may take a change of paradigm to evaluate a theory in a different way, causing it to be replaced (though in the past, this has sometimes happened even though the new paradigm has been initially less supported by the available observations, see Kuhn 1970; Feyerabend 1975; Lakatos 1978).

More recently a statistical version of the falsification method has been promoted by Mayo and her co-workers (see, for example, Mayo 1996 and the discussions in Mayo and Spanos 2010, cf. also Chap. 19 by Robinson in this volume). This approach recognises that all experimental methods are subject to observational and sampling uncertainties, so hypotheses and theories should be exposed to strong statistical testing in validation. Failure of such tests would then constitute falsification. An example is the "5σ" test used in particle physics, where σ represents the standard deviation of the observations, such as in the identification of the Higgs Boson in the Large Hadron Collider at CERN (e.g. CMS Collaboration 2013). This requires that the variability of the data are well described by a Gaussian distribution, but if this assumption is accepted, then 5σ represents a 1 in 3.5 m chance ($p = 0.0000003$) of making a Type I error, where a false theory is accepted as correct. If this test is passed, then the hypothesis or theory is not rejected and can be considered as corroborated by the evidence. Similarly, tests on discrepancies between the data and theoretical predictions can be used to suggest when a theory should be rejected, though interestingly there do not seem to be any equivalent accepted standards in such cases for the probability at which falsification is confirmed. This is almost certainly an effect of the general bias against the publication of failures (see, for example, Masicampo and Lalande 2012), even though the statistics of negative results might be an important consideration in risk management (e.g. Mayo 1991). More often, hypotheses and theories that (to a more or less extent) conflict with observations are modified or replaced rather than simply being discredited or falsified in the literature. We revise our beliefs and hence our theories (Quine 1969; Morton 1993; Klein and Herskowitz 2007) through the addition of auxiliary information (e.g. empirical parameterisations of momentum loss and secondary circulation in rivers) even though we know that the reason that makes this auxiliary information needed (depth-averaging of the full 3D Navier–Stokes equations) fundamentally invalidates the capacity of depth-averaged models to represent the nature of river flow.

An incorrect rejection would be a Type II or false negative error (rejecting a model as a hypothesis that should not be rejected). In any statistical test there is a trade-off between Type I and Type II errors so the lower the required probability of avoiding a Type I error (as in the 5σ case), the higher the probability of a Type II error. This probability can be reduced by adding more informative observations, when this is feasible. Mayo's response to the Duhem–Quine thesis is to suggest that strong statistical testing implies the testing of any auxiliary conditions related to the theory. "*A claim can only be said to be supported by experiment if the various ways in which the claim could be at fault have been investigated and eliminated*" (Mayo 1996, p. 199). This represents severe testing but is not always possible, particularly when we wish to test the implementation of theories, and complex, multi-component models based on theories, to situations where controlled experiments are impossible or difficult to justify economically. This is the case for the very many models of environmental systems currently being used, where knowledge of parameter values and boundary conditions may be subject to significant epistemic uncertainties. However, it also emphasises the need to test not only the model per se, but the constituent hypotheses, theories, models or auxiliary relations that are contained within it. Testing model outputs may not be sufficient.

6.3 Simulation Models, Invalidation and Falsification

These difficulties become particularly apparent where theories about some aspect of reality are combined and implemented as a computer simulation model, and it is the outputs from the model that are compared with observations. In many cases, for applications of environmental models to real-world open systems, the models are based on theories that are not expected to represent fully the complexity of the real world. This may be because full knowledge of the processes relevant to that complexity is lacking; because the processes have had to be simplified, or even ignored, to make the model tractable; because knowledge about the boundary conditions, initial states and characteristics of the system is insufficient; or it may simply be because the currently available computational resource does not allow a closer degree of approximation. These are all sources of epistemic uncertainty that might result in complex and nonstationary structures in model residuals when simulation outputs are compared against observations.

In such cases, auxiliary rules are often introduced to represent the consequences of simplification of the system being modelled, whether of the hypotheses being used in the model, the boundary or initial conditions needed to apply the model or the spatio-temporal scale at which the model is applied. Such rules commonly invoke free parameters, difficult to estimate a priori given limited information about the complex system and thus they are often calibrated against available observations (e.g. Morton 1993; Beven 2002). For the modeller, such parameters may not simply be a consequence of model implementation (e.g. simplification, approximation) but a necessary element of being able to make a model perform through the process of model calibration (Lane 2012). For example, in river flood modelling, modellers have typically used a single empirical parameter to represent friction losses due to a range of different processes (e.g. dispersion effects due to secondary circulation, turbulence, friction at the stream bed and energy losses at the water surface). Lane (2014) reports that an attempt to improve the determination of one of these parameters (the Manning roughness coefficient) was largely rejected in practice, because it was needed as an adjustable effective parameter that allowed modellers to make their model perform against observations. The improved parameterisation was not and could not be adopted. The notion that a model is made to perform reminds us that this performance might achieve the right results but this is not necessarily for the right reasons (Beven 1989): a model can be forced to be empirically adequate (Oreskes et al. 1994) and in some sense acceptable by the calibration of effective values of its parameters; even if it might be falsified in terms of the validity of the auxiliary relations that are used to make it acceptable. The question is then whether it will be equally fit-for-purpose in predicting future changed conditions.

There can also be issues about the commensurability of observables and model variables, due to differences in scale or meaning, even when both are given equivalent names in the theoretical context used. In environmental systems, for example it can often be the case that observations are made at a "point" in space and time, while a model predicts a variable of the same name at some larger space–time discretisation. When there is little information about the sub-discretisation heterogeneity of the observable, it can then be difficult to relate one to the other. In many circumstances, it can also be difficult to assess that heterogeneity. For example, Hills and Reynolds (1969) examined the variability of point soil moisture measurements in a field and concluded that more than 150 measurements were necessary to estimate the mean value to within ±5%. Even in research projects such a sampling density is rarely affordable and such a field might represent just a single model grid element. In this case, recent advances in measurement technology can help overcome this problem by sampling surface soil moisture at larger scales (e.g. the COSMOS method, Zreda et al. 2012). However, hydrologists are not only interested in the surface soil moisture, but also in the water stored in the full soil profile, which is even more difficult to observe experimentally (but see the recent study of Güntner et al. 2017, using microgravity as an indication of how new measurement techniques might help constrain uncertainties). Similar issues arise at larger scales for variables within global or earth system science models. Such commensurability issues represent a fundamental limitation for the validation or falsification of such models.

These issues underlie George Box's aphorism that "*all models are wrong but some are useful*" (Box 1979), or as expressed by Morton (1993, p. 662): "the modelling assumptions are generally false, **and known to be false**, relative to a standard governing theory" (emphasis added). There is thus an **expectation** that our models could be falsified, especially if we look at what they predict in close detail (even if this is not reflected in how those models are presented in the literature). In this situation, therefore, there is an issue of what degree of approximation to the observational data we are prepared to accept before we allow that our modelling assumptions are wrong, knowing that there are uncertainties associated with the boundary conditions and evaluation data for any model application. Effectively, this requires a definition of the point at which we accept that a model might be invalidated as not fit-for-purpose in making the predictions required of it, while making proper allowance for the epistemic and aleatory uncertainties in the modelling process. We can, therefore, differentiate between invalidation of a simulation model structure based on the outputs relevant to a particular purpose, and the falsification of any of the individual hypotheses or theoretical constructs that might be involved as components of that model based on more controlled experimental testing (see also the frameworks suggested by Bennett et al. 2013; Augusiak et al. 2014).

6.4 Fitness-for-Purpose, Verisimilitude and Likelihood

The question of fitness-for-purpose is analogous to, but somewhat different from, Popper's original discussion of the evaluation of the verisimilitude or truthlikeness of a theory about reality. Popper suggested that we should accept that ultimately we could never be sure to have found a correct theory; even if it has survived all tests to date, the next inference it makes might prove to be wrong. However, the very process of testing and rejecting in this way and consequently building new theories should, over time, increase the degree of verisimilitude of the theory being applied. Reasoned argument suggests that we should, in principle, prefer theories or models as hypotheses with a greater degree of verisimilitude than others. This requires a scale of verisimilitude in order to determine a ranking of the multiple working hypotheses under consideration. Popper made some specific suggestions about the nature of that scale: that for a hypothesis to have greater verisimilitude than some competing hypothesis, the truth content of the first should include that of the second; while the false content of the first should be a sub-set of that of the second (Popper 1976). This proposal was shown to be logically untenable by Miller (1974). Subsequently, a variety of other technical definitions of verisimilitude have been proposed to try and overcome this limitation (see the recent discussion of Niiniluoto 2017). It also led to Popper to suggest later that the concept of verisimilitude need not be considered an essential part of his theory (Introduction 1982, p. xxxvi, in Popper 1983).

However, as scientists we still tend to think that it is possible to move from hypotheses that are known to be false in some sense, towards hypotheses that are closer to a correct description of the real system, even if still false in some lesser sense, i.e. from a lower to a higher degree of verisimilitude. Watkins (1985) expresses this in the sense of trying to assess the relative merits of hypotheses when one might be more readily corroborated than another, even if both might be far from the truth. In his later writings Popper accepted that, even if corroboration could not be used as a scale of verisimilitude, it could be used as an indicator of verisimilitude. Thus: "*If two competing theories have been criticized and tested as thoroughly as we could manage, with the result that the degree of corroboration of one of them is greater than that of the other, we will, in general, have* **reason to believe** *that the first is a better approximation to the truth than the second*" (Popper 1983, p. 58). In this context, the aim of the method is to justify a *preference* for one hypothesis over another, as a closer approximation to the truth, based on the evidence available, and using reasoned argument (Deutsch 1997; Klein and Herskovitz 2007). This does not now imply, however, that such a preference will necessarily be equivalent to a greater degree of verisimilitude.

But such corroboration with the evidence can be considered as a form of induction (e.g. O'Hear 1975), at odds with Popper's aim of providing a hypothetico-deductive scientific method. This will be even more the case when the hypotheses are implemented as computer simulation models with free parameters that need to be calibrated for some specific application, especially in the case of models that become over-parameterised with respect to the information content of the available observations.

This inevitably invokes induction from the (uncertain) empirical observations used in calibration when making inferences about the future behaviour of the system under study. It also makes falsification and the assessment of degrees of verisimilitude more difficult. Many potential models might fit the available observations to some acceptable degree of error (e.g. Beven 2006; Chap. 33 in this volume); some might be more truth-like or fit-for-purpose than others, but how do we make such an assessment?

Howson (2000) has suggested that one solution to the problem of induction is to work within a Bayesian framework (see also Chap. 7 by Beisbart and Chap. 20 by Jiang et al. in this volume). When we may not be able to assess a degree of verisimilitude of a hypothesis, we might be able to assess how the evidence could change our degree of belief in that hypothesis (see Howson and Urbach 1993 and this volume, Chap. 19). In modern applications of Bayes, the degrees of belief are most commonly expressed as terms of probability and the degree of explanation is called the likelihood. As new evidence becomes available Bayes theorem can be applied recursively so that hypotheses that are successful in the sense of having higher likelihoods will gradually develop higher posterior probabilities or degrees of belief. At no point, however, is it necessary to invoke any measure of truthfulness or verisimilitude, which makes the framework evidently suitable for application to hypotheses implemented as models while accepting that all models are idealisations of reality (or to some greater or lesser extent false).

This Bayesian framework, however, has been criticised for its subjectivity in both the prior assessments of degree of belief and in the choice of likelihood measure. The latter subjectivity has been addressed by statisticians in developing formal likelihood measures (or objective functions) that follow from specific assumptions about model errors (see, for example, Box and Taio 1992; Bernado and Smith 2000; Fernandez and Steele 1998; Beven 2009; Schoups and Vrugt 2010; Rougier 2007) but in applications to complex open systems it may be difficult to justify those assumptions. In such cases, the use of a formal statistical likelihood can lead to overconfidence in model evaluation when a large number of observations are available, for example, when time series are used in model evaluation (e.g. Beven 2012b, 2016; Beven and Smith 2015). This is because of the way in which the contributions of individual model residuals are combined multiplicatively, which may lead to models that have nearly equal error variance simultaneously having orders of magnitude differences in likelihood (even when bias and autocorrelation of model residuals are included in the likelihood function, see Beven 2016). Alternative subjective definitions of likelihood, that allow for the fact that model errors may not be simply stochastic, can avoid this stretching of the likelihood surface but do not have the same formal theoretical foundation.

There are other aspects of the formal Bayesian framework as based on probabilities that are relevant to the current discussion. The first results from the fact that the probability and statistical likelihood distribution functions that are commonly used have infinite tails (e.g. Bernado and Smith 2000). This means that no hypothesis that has a finite prior probability will be given a posterior probability of zero. The posterior probability might become very small for those models that do not perform well relative to the observations, but never zero. Consequently there is no falsification within this framework, unless some other, more subjective, threshold of incompatibility with the evidence is imposed such that the likelihood can be set to zero. Falsification is then a limiting case of updating, but is outside the framework of formal statistical likelihood theory.

Another aspect of the formal Bayesian framework is that the hypothesis or model with the highest posterior likelihood will not necessarily be good enough to be useful for its intended purpose (let alone approach a truth-like representation of the real system). A further, related, point is that the approach normally takes no account of the fact that the probabilities might be incomplete: the approach is normally applied without taking any account of the fact that there might be other competing hypotheses (and consequent model structures) that have not been included.

6.5 If All Models May Be False, When Can They Be Considered Useful?

In some sense, we are all Bayesians because we have an expectation that additional evidence should lead to a refinement in our hypotheses and models about how the real-world system works. The question, therefore, is whether we have sufficient information to differentiate between hypotheses given the uncertainties associated with the modelling process. This in the Bayesian context equates to how best to define a likelihood to condition our degree of belief in a particular hypothesis, and to determine when the likelihood should be set to zero in cases where we infer that not only is the model false, but we have no belief that it will be useful for the purpose for which it is intended to be used.

This represents a challenge for four reasons, that apply to all models in the environmental and ecological sciences, including those that claim to be based on physical principles (Cartwright 1999; Beven 2002, 2012a, 2016). First, repeated runs of the computer simulation program using Monte Carlo techniques to make many different realisations using the same model structure, but different parameter sets and (sometimes) boundary and initial conditions, will often reveal a spectrum of responses from the best models found to those that clearly do not represent the observed behaviour well at all. Very different values of the same parameter (or even models with very different structures) may lead to equally "good" evaluation (or likelihood) measures; this is the equifinality thesis of von Bertalanffy (1968) and Beven (1993, 2006, Chap. 33 in this volume). Second, the evaluation or likelihood measures may reveal different things about what constitutes a good model performance. There may be Pareto trade-offs between the rankings of different models when evaluated against different criteria. Different periods of evaluation data can also change the rank ordering. Third, some of the data available to drive a model and to evaluate the outcomes of a model run might be disinformative in respect of whether a model performs well or not (Beven and Smith 2015; Beven 2016). Fourth, fitness-for-purpose implies more than just an epistemological concern as to when a model cannot be rejected against certain statistical criteria but also a series of wider concerns that relate to the way in which the model sits within both wider scientific communities and decision-making processes.

In terms of scientific communities, models may continue to be used, even when it can be shown that alternative model structures can give better performance or even when the fundamental bases of the model (e.g. an auxiliary relation, as in the case of effective roughness parameters noted above) are not correct. For example simple empirical models of climate seem to provide better predictions to recent periods of historical data than general circulation models of climate (GCMs) climate models, even for global mean temperature (Fildes and Kourentzes 2011; Suckling and Smith 2013; Young 2018; for validation of climate simulations see Chap. 30 by Rood in this volume). However, GCMs continue to be used on the basis of the argument that their theoretical physical basis allows a greater degree of belief in their projections for the future (Shackley et al. 1998; Knutti 2018), whereas we cannot be sure that data-based models developed from historical observations will continue to be valid into the future. This is an argument for fitness-for-purpose based on the physical bases of process representations (Knutti 2018). Yet, GCMs involve empirical or conceptual elements in many process representations, and may be just as "empirical" as simpler models in terms of their dependence upon observational data to parameterise them (Shackley et al. 1998; Parker 2018). GCMs may also contain significant epistemic uncertainties, notably because of unknown boundary conditions (e.g. future decisions on fossil fuel use), which is why GCMs are run with different scenarios of future emissions. Yet, the number of such runs into the far future is often small because of the computational expense involved in resolving finer and finer detail in the atmospheric and oceanic circulations with each generation of model. Given these issues, Shackley et al. (1998) argue that GCMs remain dominant because they have mutually reinforced relations between GCM scientists, policy communities, climate impact communities and surrounding scientists, such that they have developed "*a wider symbolic significance than implied by their scientific credentials alone*" (Shackley et al. 1998, p. 188; see Winsberg 2003, for a wider discussion). The resilience of these relations to being challenged may explain why the question of fitness-for-purpose has rather rarely been questioned within the climate modelling community (though see Collins et al. 2012; Hargreaves and Annan 2014; and comments in Parker 2009, 2018 on the adequacy for purpose of climate models).

The above points emphasise that it is necessary to decide on what constitutes fitness-for-purpose and that such a decision may not be one that is only defined by scientific communities and past performance. What constitutes being fit-for-purpose, in general, will be highly context dependent (e.g. Barraque 2002; Wimsatt 2007; Knutti 2018) even where models are not developed with a pragmatic purpose in mind, but more because "*we are intrigued by the possibility of assembling our knowledge into a neat package to show that we do, after all, understand our science and its complex interrelated phenomena*" (Kohler 1969). For this purpose it is sufficient to be able to justify giving a likelihood of greater than zero in model evaluation, i.e. to have some degree of belief that the model mimics the functioning of the real system in some measurable sense. As Beven (2002) suggests, most modellers are pragmatic realists in this context. They would like to be able to equate the variables in their computer models with quantities and fluxes in the real system, but they are pragmatic in recognising that there are real limitations as to how far that is possible. As with GCMs, however, past performance may not be the only factor in deciding on that degree of belief: there may be strong prior beliefs about the nature of the assumptions that underlie a model, beliefs that might vary between research groupings as well as being subject to strong influence by those who wish to use model results. For this purpose model evaluations are made not only with respect to demonstrable performance, but also in terms of what is considered acceptable within a research programme in terms of assumptions and degrees of uncertainty or error in the predictions, as well as the suitability of the predictions for the purpose to which they are to be put; the "*antecedently established credentials of the model building techniques developed over an extended tradition of employment*" (Winsberg 2003, p. 122). Thus, the evaluation could be against the *opinions* of experts or users, as conditioned on expectations about sources of uncertainty in the modelling process, as much as against any kinds of observable variables used to test a model. Similar considerations will apply to experts as referees on scientific papers and research reports, with their own experiences and impressions of what might be considered as acceptable.

Given the subjectivity implicit to the above argument, it might be expected that the faith in models as a contribution to decision-making might be undermined by the eventual realisation that those models were not fit-for-purpose when viewed after the fact. But, modellers are protected to some extent from being judged as to whether past predictions were fit-for-purpose because model predictions are generally constructed as scenarios or projections. With GCMs, for instance, the most recent Intergovernmental Panel for Climate Change report (IPCC 2013, p. 21) estimates a range in globally averaged warming by 2100 (as compared to 1986–2005) of between +0.4 °C and +5.5 °C according to the combination of scenario and aleatory uncertainty chosen. It is not generally expected that any of the assumed scenarios regarding future boundary conditions will actually prove to be correct. The simulations are projections not predictions. These projections are intended as the best available simulations **conditional on** the assumed emissions scenarios and other assumptions (and therefore not expected to occur in the future). In this way they are deemed to be useful, despite the better performance of data-based models on decadal time scales noted earlier.

There has been an interesting discussion in the simulation modelling community about the value of such projections in terms of the robustness of simulating future outcomes (e.g. Weisberg 2006; Lloyd 2010, 2018). This debate has recognised that individual models might be deficient in their predictions in the past, but that across an ensemble of models, some features of the projections might be robust to the specification of parameterisations and auxiliary conditions in individual models (Oldenbaugh 2018). The general trends in global warming in response to specific emission scenarios in the CMIP5 ensemble of GCMs is an example. It has also been pointed out, however, that in the climate model case the different model projections are not independent, but share common histories of development and prioritisation of added components over time (Oreskes 2018). It is also the case that robustness of projections across the ensemble, in terms of a commonality of outcomes, does not imply that any of the models is fit-for-purpose, but is simply corroboration of one model by another (Parker 2018). To get round this, it has been suggested that robustness should only be inferred when all the models in the ensemble have been empirically validated against past observations (Lloyd 2010), but clearly GCMs have limitations in reproducing past observations in detail (Parker 2009, 2018). The CMIP5 ensemble is currently the set of best available models; it remains unclear as to whether they are fit-for-purpose when they require bias corrections and flux corrections when used for evaluating the impacts of future climate changes on societies.

In other areas, where (rarely) post hoc assessments of modelled futures have been carried out more formally, the results have not been good. Examples are provided in the post hoc assessments of groundwater models reported in Konikow and Bredehoeft (1992) and Anderson and Woessner (1992). In some of the cases considered the conceptual model of the groundwater system proved to be inadequate; in others, the conceptual model was adequate but the estimation of future boundary conditions proved to be totally inadequate. Groundwater modelling is an example of where modelling technology has developed rapidly in the more than two decades since those papers were published and the four decades since the original modelling studies have been done. But, in most groundwater modelling applications, we still have limited knowledge of the subsurface geological characteristics and parameters, particularly in fractured rock systems, and future boundary conditions (climate, recharge, well development, pumping rates, etc.) are necessarily uncertain. Similar issues will arise in all areas of the inexact natural sciences. It is likely to be an even greater impediment for the social sciences even though that has not stopped attempts to model the joint development of natural and social systems into the future (e.g. in sociohydrology, see Viglione et al. 2014; Elshafei et al. 2014; Jeong and Adamowski 2016; Pande and Savenije 2016).

6.6 Defining Fitness-for-Purpose and Model Invalidation

The above argument is predicated upon the idea that a simulation model should be shown to be fit-for-purpose, that is corroborated against some kind of observation or judgment, even if there are few rules about precisely what constitutes “fit” and “purpose”, such that its use can be justified. For both the purpose of understanding our science and informing decisions, the question that arises is how good is good enough to be useful, given the uncertainties in the modelling process. This can be posed as a problem of showing that a simulation model is invalid for the purpose intended, while taking proper account of those uncertainties. No modeller wants to present a model that is invalid of course: within research programmes considerable efforts are put into ensuring that the assumptions on which the model is based are justifiable; that the equations derived from those assumptions are correctly formulated; that the coded version of those equations is debugged and numerically accurate; that the parameter values used within the model are suitable; and that the model produces presentable results against some evaluation observations. However, we wish to argue here for the importance of seeing model *invalidation* as a good thing, perhaps the ultimate goal of model use in science, in contrast with the simple use of the best models available in applications to society, when the best models (or ensemble of best models) might not be fit-for-purpose.

From a scientific perspective, model rejection is a positive outcome; it implies that we need to do better, either in defining better model structures or in generating better observations to drive and evaluate models. Of course, when modelling is used in practice, and uncertainties in the modelling process are recognised, there can be substantial constraints upon the capacity for a model to be shown to be false or invalid. The limited research that has traced the transition of model development into model adoption has revealed how social and economic constraints determine the extent to which a scientifically rejected model leads to the evolution of modelling practice (e.g. see Lane et al. 2013, for the case of flood inundation models and the discussion of GCMs above). Such constraints emphasise the difficulty that can exist in rejecting a model formulation as false. The philosopher of science, Isabelle Stengers (2013)[1] argues for a resistance to the constraints upon scientific practice related to both socio-economic limits as well as scientists' own institutional and community settings. She argues that being "scientific" requires us to recover our own capacity to be wrong and, in so doing, to raise different questions to those which we are being forced to ask. In 2005 she wrote: "*How can we present a proposal intended not to say what is, or what ought to be, but to provoke thought, a proposal that requires no other verification than the way in which it is able to 'slow down' reasoning and create an opportunity to arouse a slightly different awareness of the problems and situations mobilising us?*" (Stengers 2005, p. 994). Stengers' position here is interesting because it is in marked contrast to one of the traditional raison d'être of models which is to speed up time, to allow the future to become present today, such that society can invest now to make the future that becomes manifest more palatable. We develop Stengers' ideas more specifically below.

There is a very strong parallel here between the notion of model rejection or invalidation and the Popperian concept of falsification. By allowing for models to be invalidated, we may be able to move towards truer theories and models in an evolutionary way (e.g. Popper 1969; Dolby 1996; Deutsch 1997; Wimsatt 2007). Popper also made this point in saying that a falsificationist would "*prefer to solve an interesting problem by a bold conjecture, even (and especially) if it turns out to be false, to any recital of a sequence of irrelevant truisms*" (1969, p. 231). Learning from our mistakes should bring us further to a realistic representation of a system of interest, even if only an approximation to reality is attainable. The **nature** of the rejection can then provide valuable information about the assumptions on which a model is based, or the data needed to apply and evaluate the model, provided we allow it to do so. The question that then arises is twofold. First, how do we define criteria to invalidate a model as fit for its intended purpose? This is a problem analogous to defining a measure of verisimilitude in the Popperian framework, albeit that fitness-

[1] This is written in French. See Lane (2017) for an English interpretation.

for-purpose is a lesser requirement than truthlikeness. The second question, addressed in part below, is how to reconcile the self-interest of model advocates who want to present predictions as acceptable and useful, with the fundamental scientific progress that comes from accumulating our (posterior) beliefs that a model is no longer fit-for-purpose.

Wimsatt (2007, pp. 100–106) provides an analysis of 7 ways in which models might be wrong, and 12 ways of learning from models that are wrong (and sometimes designed to be wrong as a way of illuminating system processes). He suggests that the ways in which models are modified over time as a result of testing and thoughtful reasoning is the way in which much of science is normally practiced (similarly arguments are made by Koen (2003) in a discussion of engineering practice, and Klein and Herkowitz 2007, from a simulation philosophy perspective). This is a rather instrumentalist view of scientific method, in that all the time that theoretical tools and models provide some utility, they will not be rejected; and when they appear to be wrong, we learn from how they appear to be wrong. However, it is very similar to Quine's (1969) notion of "belief revision". Mayo (1996, Chap. 1) also considers learning from mistakes, but firmly within a falsificationist approach, with a heavy use of error statistics within a statistical theoretical approach. Such an approach depends, of course, on making strong aleatory assumptions in statistical testing, which may be difficult in the applications of models to open systems with epistemic uncertainties that are characteristic of the environmental sciences.

The discussion of the previous sections and past experience suggests some principles on which to base any assessment of model invalidation.

a. Within the feasible model space (of model structures and parameter sets) it should be accepted that model outputs often show a wide spectrum of goodness-of-fit from the best models found to those that are far from any evaluation data or evidence.
b. Fitness-for-purpose is concerned with the best simulation models found, but these may be localised in a high dimensional model space and may not be easy to find.
c. The best simulation models found will depend on the criteria of evaluation used, and also on the set of forcing and evaluation data used. The criteria used should therefore, as far as possible, reflect the framing of the purpose intended.
d. Uncertainty in the input or forcing data is important—by analogy with statistical hypothesis testing we do not want to accept a "false" model or reject a "useful" model just because of uncertainties or disinformation in the forcing and boundary condition data (or other auxiliary conditions).
e. The structure of a simulation model should add value; we should not accept a simulation model that is not significantly better than a parsimonious non-parametric data-based model for the variable of interest. The data-based model might be overfit, but so could the simulation model when used with the same forcing data.
f. Fitness-for-purpose should be defined prior to running any model simulations, taking account of understanding of uncertainties in the modelling process; we do not want to compensate poor performance simply by an error model with large variance.

g. A simulation model that is deemed fit-for-purpose should not be expected to necessarily remain fit-for-purpose if the assimilation of further evidence suggests the model fails in some important respect.

There are a variety of methods for model choice available. These include methods based on Bayesian inference, statistical implausibility measures and methods based on tolerance thresholds or limits of acceptability. As discussed earlier Bayesian inference is based on defining a measure of likelihood together with any estimates of prior probability for model formulations that might be based on past applications or testing. The definition of a likelihood measure is now commonly based on more or less complex statistical assumptions about the nature of the model residuals (e.g. Bernado and Smith 2000).

6.6.1 Using Bayes Ratios to Differentiate Between Models

Model comparisons can be made in terms of the posterior marginal probability distributions for different model structures, expressed as Bayes factors or ratios. The Bayes ratio can be defined as

$$K_B = \frac{\int[P_o(M_1\{\theta_1\})L(O \vee M_1\{\theta_1\})]d\theta_1}{\int[P_o(M_2\{\theta_2\})L(O \vee M_2\{\theta_2\})]d\theta_2} \tag{6.1}$$

where M_1 and M_2, with parameter vectors θ_1 and θ_2, are two different model structures under consideration; P_o is the prior probability for each model and L is the likelihood when model predictions are evaluated against the observations O. Since the ratio is defined in terms of probability integrals, it will not give a crisp differentiation between valid and invalid models. Some rules of thumb have been suggested for model choice using the Bayes ratio. Thus, for ratios of >20 we should have a strong preference for M_1 over M_2; and for ratios >150 we should have a very strong preference for M_1 over M_2. (e.g. Kass and Rafferty 1995). Note, however, that to be directly comparable the likelihood definition used in evaluation of each model should be directly comparable. Where this is based on statistical assumptions about the nature of the model residuals it requires the same structural assumptions. This may, or may not, be appropriate for the different error model structures and is an assumption that should be checked in good practice. Experience suggests that such ratios can be sensitive to such assumptions and can vary dramatically (by tens of orders of magnitude) depending on what periods of data are used in the evaluation (see the discussion in Beven 2016).

For cases where it is difficult to define an explicit likelihood measure, the Bayes ratio can be approximated using Approximate Bayesian Computation (ABC e.g. Robert et al. 2011). Interestingly the ABC methodology depends on defining some tolerance level for model acceptance. This is sometimes refined as the search within the model space (or spaces in the case of multiple model structures) proceeds. We know of no cases, however, where it has been defined on the basis of fitness-for-purpose, rather than ensuring a sufficient sample of acceptable models.

Note also that the integral for each model in Eq. 6.1 integrates over all plausible model parameter sets; it does not focus on the best performance for each model structure. In evaluating fitness-for-purpose it might therefore be better to consider only the maximum likelihood associated with each model in which case [3] reduces to a likelihood ratio test that involves only a single parameter set in each model structure. Again under the proviso that a similar error model assumption is appropriate for each of the models considered, the likelihood ratio can be used to evaluate whether one model is more acceptable than another, but not necessarily whether either is fit-for-purpose.

6.6.2 *Use of Implausibility Measures to Differentiate Between Models*

A somewhat different statistical approach has been suggested by Vernon et al. (2010) for cases where it is difficult to specify a likelihood measure based on residual error characteristics. Rather than use a likelihood measure, they propose the use of an implausibility scaling of the following form:

$$I^2(x_i) = \frac{\{O_i - M(x_i;\theta)\}^2}{\left\{Var\left(e_{M,i}\right) + Var\left(e_{O,i}\right)\right\}} \tag{6.2}$$

where x_i is the ith model output variable, $M(x_i; \theta)$ is the model prediction of x_i given a parameter set θ; O_i is the equivalent observed variable, $e_{M,i}$ is an estimate of model uncertainty (arising from allowable model discrepancy or from stochastic forcing) and $e_{O,i}$ is an estimate of the observation uncertainty for the ith variable. Separate implausibility measures can be calculated for all available observation–prediction matching couples, and combined into a total measure of implausibility. The measure can be updated as new information becomes available. Implausibility, as defined in this way, is similar to the Bayes ratio, in that it represents a continuous relative scale with no sharp cut-off. Again some rule of thumb is required to decide where the limit of plausibility lies on that scale. In Vernon et al. (2010) and Woodhouse et al. (2015) the plausible model space is defined by a threshold of $I < 3$, based on the 3σ rule, implying that the plausible region contains the most plausible model, allowing for both model and observational uncertainty, with probability greater than 95%. Other forms of plausibility measure are discussed in Halpern (2005).

6.6.3 *Use of Limits of Acceptability to Define Behavioural Models*

Both the Bayesian and implausibility measure approaches depend on the magnitude of the model residuals evaluated after each model run. They do not require any decision to be made about some threshold of acceptability before making a model run. An alternative method that does require a prior definition of acceptability is the Limits of Acceptability implementation of the Generalised Likelihood Uncertainty Estimation (GLUE) methodology as outlined by Beven (2006). GLUE is based on Monte Carlo sampling of the model space to identify an ensemble of acceptable or "behavioural" models that will be used in prediction. Simulations that do not pass the limits of acceptability test are rejected as non-behavioural or invalidated, i.e. they are not considered to be fit-for-purpose. The approach is general in that it can be applied to parameter sets and uncertain boundary conditions for one or more model structures, with likelihood measures defined and combined in different ways (Beven and Binley 1992, 2014). Statistical likelihood functions and combining likelihoods using Bayes equation represent a special case within GLUE, where the necessary assumptions can be justified. Different search algorithms can be used to explore the model space (e.g. Beven and Binley 1992, 2014; Blasone et al. 2008; Vrugt 2016; Vrugt and Beven 2018).

Within this framework, the ensemble of behavioural models can be used to produce likelihood weighted predictions, but it also allows for the possibility that none of the sampled models reach the level of performance required for a particular purpose. Thus, in GLUE, the choice of a behavioural threshold assumes a particular importance, but allows the consideration of fitness-for-purpose for a given application in doing so. In the past GLUE has been criticized for the subjectivity in making such a choice so Beven (2006) suggested that the choice should be made more objective by considering what is known about the data that is used to drive and evaluate the model, as well as what level of performance is needed for the predictions to be considered useful. The use of limits of acceptability in this way is analogous to the tolerance limits used in ABC (e.g. Nott et al. 2012; Sadegh and Vrugt 2013), or applying a limit to an implausibility measure, except in that the limits should be defined before making any model runs.

In doing so, limits of acceptability can be applied to predictions of either individual observations (e.g. Liu et al. 2009), or of summary statistics relevant to the purpose (e.g. Westerberg et al. 2011; Westerberg and McMillan 2015). It is, therefore, possible that (harking back to Popperian falsification) a model could be rejected on the basis of the failure to simulate a single observation within the limits of acceptability, if that observation is considered sufficiently important. Popper notes, however, that "*a few stray basic statements contradicting a theory will hardly induce us to reject it as falsified. We shall take it as falsified only if we discover a* **reproducible effect** *which refutes the theory. In other words, we only accept the falsification if a low-level empirical hypothesis which describes such an effect is proposed and corroborated. This kind of hypothesis may be called a* **falsifying hypothesis**" (1959,

p. 86). On the other hand we should, perhaps, be rather wary of generalising this idea of reproducibility to a form of simple statistical 3σ/95% threshold, since it is quite possible that the remaining 5% might be those observations that are of most interest to the purpose for which the model is being used (e.g. the hydrograph peaks in a hydrological model application, see Beven 2016). However, in considering single observations the limits of acceptability should reflect the impact of input errors on how well a model might be expected to perform. This is important so as to avoid the Type II error of rejecting a good model that would be fit-for-purpose just because of errors in the inputs or forcing data (models are very much subject to the “garbage in garbage out” phenomenon).

The advantage of an approach based on model invalidation is that it encourages honesty in the modelling process, including about just how well we might expect a model to perform given the understanding of how a particular system works, and the data available with which to drive and evaluate the model performance (see also Smith and Stern 2011). It also allows for the possibility that all the models tried might be invalidated as not fit-for-purpose (see for example Brazier et al. 2000; Choi and Beven 2007; Dean et al. 2009; Mitchell et al. 2011; Hollaway et al. 2017). Where this happens, the model structure has been effectively invalidated, at least for that application. Commonly, however, it will survive in other applications, perhaps with less constrictive evaluation measures or limits of acceptability, rather than being reconsidered and modified. We would surely learn more from trying to understand the reasons for such rejections (Beven 2018).

6.7 Epistemic Uncertainties and Model Invalidation

The types of open system models that have been discussed in the last section are commonly subject to epistemic uncertainties or knowledge gaps. In such cases, the use of strong statistical assumptions about the sources of uncertainty might lead to overconfidence in inference because they result in a stretching of the likelihood surface, such that model and parameter uncertainty tends to be underestimated, and the residual error variance will expand to compensate. Where there are time series of data, with large numbers of observations, this stretching can be extreme and unrealistic (see Beven 2016).

Clearly, other forms of likelihood measure can be used (as, for example, in the GLUE methodology), but at the expense of losing the formal probabilistic interpretation embodied in a formal statistical likelihood function that follows from specific distributional assumptions about the model residuals. However, for good epistemic reasons, it will remain difficult to capture the nature of perceived epistemic uncertainties in the form of a statistical likelihood measure. This is particularly true for input data that might be subject to epistemic uncertainties because such uncertainties will be propagated through the (generally nonlinear) dynamic structure of the system model, interacting with any model structural error to produce complex output error structures. Even if input errors could be defined simply (e.g. as Gaussian distributions with homoscedastic variance) the output errors would then be nonstationary in bias, variance and autocorrelation, depending on the sequence of events. But the input errors are more likely to be epistemically nonstationary in complex ways, compounding the problem of how to represent the uncertainty in model evaluation. In extreme cases, the available input and output data might, at least in part, be physically inconsistent and therefore not informative about whether a model is fit-for-purpose. Where this can be identified, it can also be taken into account in setting limits of acceptability and making predictions (e.g. Beven and Smith 2015).

That is one reason why such limits should be defined a priori, before running a model, to avoid rejecting periods of data just because they are not well fitted by the model. The question is then how to do so, if we expect that there will be a significant impact of epistemic input errors on model predictions and consequently the appropriate limits of acceptability in assessing fitness-for-purpose. This is analogous to the problem of defining the term $e_{M,I}$ in the implausibility framework, but without knowing how to define the stochastic input variation. This remains a problem to be resolved, including for cases where interaction with stakeholders and decision makers might introduce more qualitative evaluation of models (see, for example Landström et al. 2011; Haasnoot et al. 2014).

6.8 The Model Advocacy Problem

We want to finish this Chapter with some thoughts on what we call the "model advocacy problem": how is it that we can move from advocating our models as somehow useful to seeing scientific progress as arising when we realise from our accumulated (posterior) beliefs that a model is no longer fit-for-purpose? The relevance of this question has been touched upon at a number of points throughout this Chapter, in relation to Global Climate Models and flood inundation models, for instance. It is an important concern because it has been shown (e.g. Landström et al. 2011) that "*[A]ccustomed to living in their entrenched fields, researchers end up with eyes only for the problems which are born in their laboratories*" (Callon et al. 2009, pp. 94–95). Research that has followed the evolution of modelling as a practice has shown that models can become bound into an assemblage that resists attempts (e.g. new knowledge) that might break it apart. In relation to flood inundation modelling, the Manning's *n* roughness parameter was too valuable as a model parameterisation tool that attempts to improve its measurement and representation failed (Lane 2014). If models can develop resistance to their own invalidation through the assemblage of people (scientists, consultants, policy-makers), technologies and places of which they come a part, what are the conditions that may break down that resistance, that make model invalidation possible?

One response is a fundamentally scientific one, to be empirical in the very broadest sense of the term. How is it that we can establish practices that allow the world "to speak back" to the modeller, to challenge the way the world is being represented (Baker 2017; Lane 2017; Beven and Alcock 2012; Beven 2018) by the model. This is not always straightforward because of the assembled network of constraints that serve to protect the model's (and modeller's) status as it has become (e.g. Lane et al. 2013). Stengers (2013) argues that one way of doing this is through finding ways that make a scientist turn away from their normal communities of practice (as scientists) and the abstraction of their investigation out of the milieu of which it is normally a part (see also Baker 2017, in relation to hydrology; Landström et al. 2011, in relation to flood modelling). For Stengers, this should be done through the "*enrolment of phenomena*" (trans. p. 127) that don't dictate how they should be described but rather are given the "*capacity to evaluate the relevance of the way they are being described*" (trans. p. 68). Stengers' argument points to the need to focus less on a model's goodness-of-fit and more on those points that don't fit the model and, as a result, cause us to slow down our reasoning to the point at which other kinds of hypotheses and simulation models might be deemed suitable or other, quite different, kinds of approaches meaningful (Lane 2017).

It is right, then, to admit that our models can be wrong (see Beven 2016, 2018), in that this implies that further improvements to either input data or modelling hypotheses need to be made. How this might be done in practice is not, however, evident. We can perhaps distinguish between model use in relation to applied questions, where a model might be a tool that assists with decision-making, and model use in scientific research where progress will be made when a model is found to be invalid. When the latter is the case, it implies that the model might not be fit-for-purpose for applied uses, but it is clearly evident that for applied use there is so much investment and vested interests in the development of modelling packages that any invalidation will tend to be hidden within the improvements associated with new version releases. A new version will be developed when it is found that modifying parameters or auxiliary conditions within a modelling framework is not sufficient to match the observational data to a degree acceptable to the client (or a critical bug in the code is found), but there may still be significant resistance to the invalidation of the fundamental concepts on which a modelling package is based. The question of when to use model invalidation is then intrinsically embedded in the communities of practice within which model applications are situated, and dependent on critical feedback from those communities.

Stengers suggests that model advocacy works against thoughtful scientific progress. There is also the issue as to whether models that can be considered invalidated with respect to the science can be considered useful when providing predictions for applied decision-making. We suggest therefore that a new way of appreciating a problem is required that allows invalidation to be pursued more widely and more thoughtfully. One way of doing so might be to give the concept of fitness-for-purpose more prominence in both the scientific and applied use of models.

6.9 Conclusions

This paper has discussed a number of aspects of invalidation of models as not fit-for-purpose for a particular application, drawing an analogy with the Popperian idea of falsification of hypotheses and theories. It has been shown that as soon as epistemic uncertainties in observational data and boundary conditions are acknowledged invalidation loses some objectivity. The original Popperian concept of falsification as a way of resolving Hume's problem of induction then becomes less tenable, in favour of a Bayesian framework of corroboration that contains elements of induction, particularly when evaluation allows the modification of prior estimates of boundary or auxiliary conditions and parameter values in model calibration.

This is particularly the case of models of open systems that are subject to epistemic uncertainties such that there is an expectation of models being (more or less) false when examined in detail and where it can be difficult to represent model error in terms of well-defined probabilistic structures. This means that it can be difficult to justify the strong assumptions of formal definitions of likelihood within a Bayesian conditioning framework. In addition, a Bayesian framework based on statistical likelihood functions does not explicitly allow for model invalidation, only evaluation of relative likelihoods of different model formulations (and that only under the assumption that the same statistical error structure is appropriate). Some rules of thumb for Bayes ratios have been proposed in comparing different model representations, but where the integral likelihoods are used to define the ratio, the approach does not explicitly evaluate whether the maximum likelihood models are fit-for-purpose. Other approaches based on implausibility measures and the prior definition of limits of acceptability are discussed, both of which can be applied to the evaluation of simulated individual observations for different variables and which attempt to allow for input and observational error, either as variances or in terms of support for the limits of acceptability. The limits of acceptability approach also focuses attention on how good a performance is required for a model to be fit-for-purpose in a particular application, whether that is to demonstrate scientific understanding or to inform a decision-making process. Some problems remain in applying these techniques, particularly in assessing the role of input uncertainties on fitness-for-purpose, but the approach allows for a more thoughtful and reflective consideration of model invalidation as a positive way of making progress in the science.

Acknowledgements The discussions on which this paper is based were initiated while KB was supported by the Fondation Herbette as visiting professor at the University of Lausanne. We thank Claus Beisbart, Nicole Saam and an anonymous referee for their comments on an earlier draft of this chapter.

References

Anderson, M. P., & Woessner, W. W. (1992). The role of the postaudit in model validation. *Advances in Water Resources, 15*(3), 167–173.

Augusiak, J., van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to'evaludation': A review of terminology and a practical approach. *Ecological Modelling, 280,* 117–128.

Baker, V. R. (2017). Debates— Hypothesis testing in hydrology: Pursuing certainty versus pursuing uberty. *Water Resources Research, 53,* 1770–1778.

Barraque, B. (2002). Modélisation et gestion de l'environnement. In P. Nouvel (Ed.), *Enquète sur le concept de modèle* (pp. 121–141). Paris: Presses Universitaires de France.

Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., et al. (2013). Characterising performance of environmental models. *Environmental Modelling and Software, 40,* 1–20.

Bernado, J. M., & Smith, A. F. M. (2000). *Bayesian theory*. Chichester: Wiley. ISBN 978-0-471-49464-5.

Beven, K. J. (1989). Changing ideas in hydrology: The case of physically-based models. *Journal of Hydrology, 105,* 157–172.

Beven, K. J. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources, 16,* 41–51.

Beven, K. J. (2002). Towards a coherent philosophy for environmental modelling. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 458,* 2465–2484.

Beven, K. J. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology, 320,* 18–36.

Beven, K. J. (2009). *Environmental modelling: An uncertain future?* Routledge: London.

Beven, K. J. (2012a). *Rainfall-runoff modelling: The primer* (2nd ed.). Chichester: Wiley-Blackwell.

Beven, K. J. (2012b). Causal models as multiple working hypotheses about environmental processes. *Comptes Rendus Geoscience, Académie de Sciences, Paris, 344,* 77–88. https://doi.org/10.1016/j.crte.2012.01.005.

Beven, K. J. (2016). EGU Leonardo Lecture: Facets of hydrology—epistemic error, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal, 61*(9), 1652–1665. https://doi.org/10.1080/02626667.2015.1031761.

Beven, K. J. (2018). On hypothesis testing in hydrology: Why falsification of models is still a really good idea. *WIRES Water*. https://doi.org/10.1002/wat2.1278.

Beven, K. J., & Alcock, R. (2012). Modelling everything everywhere: A new approach to decision making for water management under uncertainty. *Freshwater Biology, 56,* 124–132. https://doi.org/10.1111/j.1365-2427.2011.02592.x.

Beven, K. J., & Binley, A. M. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes, 6,* 279–298.

Beven, K., & Binley, A. (2014). GLUE: 20 years on. *Hydrological Processes, 28*(24), 5897–5918.

Beven, K. J., & Smith, P. J. (2015). Concepts of Information content and likelihood in parameter calibration for hydrological simulation models. *ASCE Journal of Hydrologic Engineering*. https://doi.org/10.1061/(asce)he.1943-5584.0000991.

Blasone, R. S., Vrugt, J. A., Madsen, H., Rosbjerg, D., Robinson, B. A., & Zyvoloski, G. A. (2008). Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. *Advances in Water Resources, 31*(4), 630–648.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press.

Box, G. E. P., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. New York: Wiley.

Brazier, R. E., Beven, K. J., Freer, J., & Rowan, J. S. (2000). Equifinality and uncertainty in physically-based soil erosion models: Application of the GLUE methodology to WEPP, the Water Erosion Prediction Project–for sites in the UK and USA. *Earth Surface Processes and Landforms, 25,* 825–845.

Callon, M., Lascoumes, P., & Barthe, Y. (2009). *Acting in an uncertain world. An essay on technical democracy*. Cambridge, MA: MIT Press.

Cartwright, N. (1999). *The dappled world. A study of the boundaries of science*. Cambridge: Cambridge University Press.

Chalmers, A. (1976). *What is this thing called science?* St Lucia, Queensland: University of Queensland Press.

Chamberlin, T. C. (1895). The method of multiple working hypotheses. *Science*, *15*(old series), 92–96.

Choi, H. T., & Beven, K. J. (2007). Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in distributed rainfall-runoff modelling within GLUE framework. *Journal of Hydrology, 332*(3–4), 316–336.

CMS Collaboration. (2013). Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *Journal of High Energy Physics*, *6,* 81.

Collins, M., Chandler, R. E., Cox, P. M., Huthnance, J. M., Rougier, J. C., & Stephenson, D. B. (2012). Quantifying future climate change. *Nature Climate Change, 2,* 403–409.

Dean, S., Freer, J. E., Beven, K. J., Wade, A. J., & Butterfield, D. (2009). Uncertainty assessment of a process-based integrated catchment model of phosphorus (INCA-P). *Stochastic Environmental Research and Risk Assessment, 2009*(23), 991–1010. https://doi.org/10.1007/s00477-008-0273-z.

Deutsch, D. (1997). *The fabric of reality*. London: Allen Lane.

Dolby, R. G. H. (1996). *Uncertain knowledge*. Cambridge: Cambridge University Press.

Elshafei, Y., Sivapalan, M., Tonts, M., & Hipsey, M. R. (2014). A prototype framework for models of socio-hydrology: Identification of key feedback loops and parameterisation approach. *Hydrology and Earth System Sciences, 18*(6), 2141–2166.

Fernandez, C., & Steel, M. J. F. (1998). On Bayesian modeling of fat tails and skewness. *Journal of American Statistical Association, 93,* 359–371.

Feyerabend, P. (1975). *Against method*. New York: Verso Books.

Fildes, R., & Kourentzes, N. (2011). Validation and forecasting accuracy in models of climate change. *International Journal of Forecasting, 27*(4), 968–995.

Güntner, A., Reich, M., Mikolaj, M., Creutzfeldt, B., Schroeder, S., & Wziontek, H. (2017). Landscape-scale water balance monitoring with an iGrav superconducting gravimeter in a field enclosure. *Hydrology and Earth System Sciences*, *21*, 3167–3182. https://doi.org/10.5194/hess-21-3167-2017.

Haasnoot, M., Van Deursen, W. P. A., Guillaume, J. H., Kwakkel, J. H., van Beek, E., & Middelkoop, H. (2014). Fit for purpose? Building and evaluating a fast, integrated model for exploring water policy pathways. *Environmental Modelling & Software, 60*, 99–120.

Hackett, J., & Zalta, E. N. (Eds.) (2013). *Roger bacon.* Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/spr2015/entries/roger-bacon/.

Halpern, J. Y. (2005). *Reasoning about uncertainty*. Cambridge, MA: MIT Press.

Hargreaves, J. C., & Annan, J. D. (2014). Can we trust climate models? *WIREs Climate Change, 5,* 435–440. https://doi.org/10.1002/wcc.288.

Herskovitz, P. J. (1991). A theoretical framework for simulation validation: Popper's falsficationism. *International Journal of Modelling and Simulation, 11,* 56–58.

Hills, R. C., & Reynolds, S. G. (1969). Illustrations of soil moisture variability in selected areas and plots of different sizes. *Journal of Hydrology, 8,* 27–47.

Hollaway, M. et al. (2017). The challenges of modelling phosphorus in a headwater catchment: Applying a 'limits of acceptability' uncertainty framework to a water quality model. Under review.

Howson, C. (2000). *Hume's problem: Induction and the justification of belief*. Oxford: Oxford University Press, Clarendon Press.

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach* (2nd ed.). Chicago, IL: Open Court.

Hume, D. (1748). *Philosophical essays concerning human understanding*. London: A. Millar.

IPCC. (2013). Summary for policymakers. In T. F. Stocker, D. Qin, G. -K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex & P. M. Midgley (Eds.), *Climate change 2013: The physical science basis. Contribution of working Group I to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge: Cambridge University Press.

Jeong, H., & Adamowski, J. (2016). A system dynamics based socio-hydrological model for agricultural wastewater reuse at the watershed scale. *Agricultural Water Management, 171,* 89–107.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association., 90*(430), 791. https://doi.org/10.2307/2291091.

Klein, E. E., & Herskovitz, P. J. (2007). Philosophy of science underpinnings of prototype validation: Popper vs Quine. *Information Systems Journal, 17*(1), 111–132.

Knutti, R. (2018). Climate model confirmation: From philosophy to predicting climate in the real world. In E. A. Lloyd & E. Winsberg (Eds.), Climate modelling: Philosophical and conceptual issues. Palgrave Macmillan (Chap. 11).

Koen, B. V. (2003). *Discussion of the method: Conducting the engineer's approach to problem solving*. New York: Oxford University Press.

Kohler, M. A. (1969). Keynote address, in *Hydrological Forecasting*, WMO Technical Note No. 92, pp. X1–XV1, WMO, Geneva.

Konikow, L. F., & Bredehoeft, J. D. (1992). Ground-water models cannot be validated. *Advances in Water Resources, 15*(1), 75–83.

Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.

Ladyman, J. (2002). *Understanding philosophy of science*. London: Routledge.

Lakatos, I. (1978). Philosophical papers. In J. Worrell & G. Curry (Eds.), *The methodology of scientific research programmes* (Vol. 1). Cambridge University Press.

Landström, C., Whatmore, S. J., Lane, S. N., Odoni, N., Ward, N., & Bradley, S. (2011). Coproducing flood risk knowledge: Redistributing expertise in critical 'participatory modelling'. *Environment and Planning A, 43*(7), 1617–1633.

Lane, S. N. (2012). Making mathematical models perform in geographical space(s). In J. Agnew & D. Livingstone (Eds.), *Handbook of geographical knowledge*. Sage, London (Chap. 17).

Lane, S. N. (2014). Acting, predicting and intervening in a socio-hydrological world. *Hydrology and Earth System Sciences, 18,* 927–952.

Lane, S. N. (2017). Slow science, the geographical expedition, and critical physical geography. *The Canadian Geographer, 61,* 84–101.

Lane, S. N., Landstrom, C., & Whatmore, S. J. (2011). Imagining flood futures: Risk assessment and management in practice. *Philosophical Transactions of the Royal Society, A, 369,* 1784–1806.

Lane, S. N., November, V., Landström, C., & Whatmore, S. J. (2013). Explaining rapid transitions in the practice of flood risk management. *Annals of the Association of American Geographers, 103,* 330–342.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*.

Liu, Y., Freer, J. E., Beven, K. J., & Matgen, P. (2009). Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *Journal of Hydrology, 367,* 93–103. https://doi.org/10.1016/j.jhydrol.2009.01.016.

Lloyd, E. A. (2010). Confirmation and robustness of climate models. *Philosophy of Science, 77*(5), 971–984.

Lloyd, E. A. (2018). The role of "complex" empiricism in the debates about satellite data and climate models. In E. A. Lloyd & E. Winsberg (Eds.), *Climate modelling: Philosophical and conceptual issues*. Palgrave Macmillan (Chap. 6).

Masicampo, E. J., & Lalande, D. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*.

Mayo, D. (1991). Sociological versus meta-scientific views of risk management. In D. G. Mayo & R. D. Hollander (Eds.), *Acceptable evidence: Science and values in risk management* (pp. 249–279). Oxford: Oxford University Press.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago, IL: University of Chicago Press.

Mayo, D. G., & Spanos, A. (Eds.). (2010). *Error and inference*. Cambridge: Cambridge University Press.

Miller, D. (1974). Popper's qualitative concept of verisimilitude. *The British Journal for the Philosophy of Science, 23,* 166–177.

Mitchell, S., Beven, K. J., Freer, J., & Law, B. (2011). Processes influencing model-data mismatch in drought-stressed, fire-disturbed, eddy flux sites. *JGR-Biosciences, 116.* https://doi.org/10.1029/2009jg001146.

Morton, A. (1993). Mathematical models: Questions of trustworthiness. *British Journal for the Philosophy of Science, 44,* 659–674.

Niiniluoto, I. (2017). Verismilititude: Why and how? In N. Ber-Am & S. Gattei (Eds.), *Encouraging openness: Essays for Joseph Agassi*. Springer. ISBN: 978-3-319-57669-5.

Nott, D. J., Marshall, L., & Brown, J. (2012). Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection? *Water Resources Research, 48*(12), W12602. https://doi.org/10.1029/2011wr011128.

O'Hear, A. (1975). Rationality of action and theory-testing in Popper. *Mind, 84*(334), 273–276.

Oldenbaugh, J. (2018). Building trust, removing doubt? Robustness analysis and climate modeling. In E. A. Lloyd & E. Winsberg (Eds.), *Climate modelling: Philosophical and conceptual issues*. Palgrave Macmillan (Chap. 10).

Oreskes, N. (2018). The scientific consensus on climate change: How do we know we're not wrong? In E. A. Lloyd & E. Winsberg (Eds.), *Climate modelling: Philosophical and conceptual issues*. Palgrave Macmillan (Chap. 2).

Oreskes, N., Shrader-Frechette, K., & Berlitz, K. (1994). Verification, validation and confirmation of numerical models in the earth sciences. *Science, 263,* 641–646.

Pande, S., & Savenije, H. H. (2016). A sociohydrological model for smallholder farmers in Maharashtra, India. *Water Resources Research, 52*(3), 1923–1947.

Parker, W. S. (2009). Confirmation and adequacy-for-purpose in climate modelling. *Aristotelian Society Supplementary Volume., 83,* 233–249.

Parker, W. S. (2018). The significance of robust climate projections. In E. A. Lloyd & E. Winsberg (Eds.), *Climate modelling: Philosophical and conceptual issues*. Palgrave Macmillan (Chap. 9).

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Popper, K. R. (1969). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge.

Popper, K. R. (1976). A note on verisimilitude. *British Journal for the Philosophy of Science, 27,* 147–159.

Popper, K. (1983). *Realism and the aim of science*. London: Hutchinson.

Popper, K. R. (1994). *The myth of framework: In defence of science and rationality*. London: Routledge.

Quine, W. V. (1969). *Ontological relativity and other essays*. New York: Columbia University Press.

Quine, W. V. (1975). On empirically equivalent systems of the world. *Erkenntnis, 9,* 317–328.

Robert, C. P., Cornuet, J., Marin, J., & Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences, 108*(37), 15112–15117.

Rougier, J. C. (2007). Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change, 81,* 247–264.

Sadegh, M., & Vrugt, J. A. (2013). Bridging the gap between GLUE and formal statistical approaches: Approximate Bayesian computation. *Hydrology and Earth System Sciences, 17*(12), 4831–4850.

Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research, 46*(10), W10531. https://doi.org/10.1029/2009wr008933.

Shackley, S., Young, P., Parkinson, S., & Wynne, B. (1998). Uncertainty, complexity and concepts of good science in climate change modelling: Are GCMs the best tools? *Climatic Change, 38,* 159–205.

Smith, L. A., & Stern, N. (2011). Uncertainty in science and its role in climate policy. *Philosophical Transactions of the Royal Society*, *369*(1956), 4818–4841 (Handling Uncertainty in Science).

Stengers, I. (2005). The cosmopolitical proposal. In B. Latour & P. Weibel (Eds.), *Making things public* (pp. 994–1003) Cambridge, MA: MIT Press.

Stengers, I. (2013). *Une autre science est possible*! Paris: La Découverte.

Suckling, E. B., & Smith, L. A. (2013). An evaluation of decadal probability forecasts from state-of-the-art climate models. *Journal of Climate, 26*(23), 9334–9347.

Vernon, I., Goldstein, M., & Bower, R. G. (2010). Galaxy formation: A Bayesian uncertainty analysis. *Bayesian Analysis, 5*(4), 619–669. https://doi.org/10.1214/10-ba524.

Viglione, A., Di Baldassarre, G., Brandimarte, L., Kuil, L., Carr, G., Salinas, J. L., et al. (2014). Insights from socio-hydrology modelling on dealing with flood risk–roles of collective memory, risk-taking attitude and trust. *Journal of Hydrology, 518,* 71–82.

Von Bertalanffy, L. (1968). *General systems theory*. New York: Braziller.

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling and Software, 75,* 273–316.

Vrugt, J. A., & Beven, K. J. (2018). Embracing equifinality with efficiency: Limits of acceptability sampling using the DREAM (LOA) algorithm. *Journal of Hydrology, 559,* 954–971.

Watkins, J. (1985). *Science and scepticism*. Princeton: Princeton University Press.

Weisberg, Michael. (2006). Robustness analysis. *Philosophy of Science, 73*(5), 730–742.

Westerberg, I. K., & McMillan, H. K. (2015). Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences, 19*(9), 3951–3968.

Westerberg, I. K., Guerrero, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., et al. (2011). Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, *15*, 2205–2227. https://doi.org/10.5194/hess-15-2205-2011.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings*. Cambridge: Harvard University Press.

Winsberg, E. (2003). Simulated experiments: Methodology for a virtual world. *Philosophy of Science, 70,* 105–125.

Woodhouse, M. J., Hogg, A. J., Phillips, J. C., & Rougier, J. C. (2015). Uncertainty analysis of a model of wind-blown volcanic plumes. *Bulletin of Volcanology, 77*(10), 83. https://doi.org/10.1007/s00445-015-0959-2.

Young, P. C. (2013). Hypothetico-inductive data-based mechanistic modeling of hydrological systems. *Water Resources Research, 49*(2), 915–935.

Young, P. C. (2018). Data-based mechanistic modelling and forecasting globally averaged surface temperature. *International Journal of Forecasting, 34*(2), 314–335. https://doi.org/10.1016/j.ijforecast.2017.10.002.

Zreda, M., Shuttleworth, W. J., Zeng, X., Zweck, C., Desilets, D., Franz, T., et al. (2012). COSMOS: The cosmic-ray soil moisture observing system. *Hydrology and Earth System Sciences, 16*(11), 4079–4099.

Chapter 7
Simulation Validation from a Bayesian Perspective

Claus Beisbart

Abstract Bayesian epistemology offers a powerful framework for characterizing scientific inference. Its basic idea is that rational belief comes in degrees that can be measured in terms of probabilities. The axioms of the probability calculus and a rule for updating (e.g., Bayesian conditionalization) emerge as constraints on the formation of rational belief. Bayesian epistemology has led to useful explications of notions such as confirmation. It thus is natural to ask whether Bayesian epistemology offers a useful framework for thinking about the inferences implicit in the validation of computer simulations. The aim of this chapter is to answer this question. Bayesian epistemology is briefly summarized and then applied to validation. Updating is shown to form a viable method for data-driven validation. Bayesians can also express how a simulation obtains prior credibility because the underlying conceptual model is credible. But the impact of this prior credibility is indirect since simulations at best provide partial and approximate solutions to the conceptual model. Fortunately, this gap between the simulations and the conceptual model can be addressed using what we call Bayesian verification. The final part of the chapter systematically evaluates the use of Bayesian epistemology in validation, e.g., by comparing it to a falsificationist approach. It is argued that Bayesian epistemology goes beyond mere calibration and that it can provide the foundations for a sound evaluation of computer simulations.

Keywords Degree of belief · Bayesian updating · Problem of old evidence · Error · Uncertainty · Confirmation · Verification

7.1 Introduction

Suppose that Connie is a computer scientist who has recently run a computer simulation of the Earth's climate. One result of this simulation is that the mean global temperature of our planet will increase by 2° in the next 30 years, if the carbon

C. Beisbart (✉)
Institute of Philosophy, University of Bern, Bern, Switzerland
e-mail: claus.beisbart@philo.unibe.ch

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_7

dioxide emissions continue as usual. This at least is what the outputs of running the simulation program say: The computed value of the mean global temperature in 30 years is by 2° higher than the present one. Can Connie trust this result? Certainly not if no case can be made for trusting the result. But how can such a case be made? That is, how can the simulation and its results be validated?

This question concerns *scientific inference*, which is a crucial part of the scientific method. Scientific inference is studied in the epistemology of science, which is part of the philosophical reflection on science. In more detail, epistemology of science investigates how scientists gain knowledge and understanding. The epistemology of science will of course not directly answer the above question, which refers to a specific case on which the answer will depend. But epistemology of science may offer a useful framework for thinking about the question.

A fairly recent approach to epistemology of science is *Bayesian epistemology*. It goes back to ideas by Thomas Bayes (ca. 1701–61) and was elaborated in the twentieth century by Bruno de Finetti, Frank P. Ramsey and Leonard Savage, among others.[1] Bayesians assume that trust, or, more specifically, belief, comes in degrees that are measured in terms of probabilities. They further provide principles of rationality with which the degrees of belief should comply.

The aim of this chapter is to discuss the application of Bayesian epistemology to the validation of computer simulations. Our main question is whether this approach offers a useful framework for thinking about validation. Since, so far, little has been said about this question, I concentrate on the philosophical foundations of a Bayesian view of validation and thus put technicalities aside.[2] My focus is on building up (or destructing) trust in simulations via so-called updating. Chapter 20 by Jiang et al. in this volume is complementary first because it provides an introduction to mathematical techniques and second because it is focused on model acceptance vs. rejection on the basis of Bayesian notions.

This chapter is structured as follows: It first proves useful to summarize the basics of Bayesian epistemology (Sect. 7.2). Since very good introductions to Bayesian epistemology are on the market, I can in principle be brief.[3] Nevertheless, to address a broad audience, the chapter is kept self-contained and does not assume familiarity with Bayesian epistemology. Section 7.3 addresses the application of Bayesian epistemology to the validation of simulations. In Sect. 7.4, I systematically evaluate Bayesianism as a framework for thinking about validation. In particular, I compare it to a falsificationist approach. Conclusions are drawn in Sect. 7.5.

Before we start, two comments are in order. First, the central ideas behind Bayesian epistemology allow for much elaboration and many applications. For instance, they

[1] See e.g., von Plato (1989, 1994), in particular Chap. 8 for the history. Gillies (2000), Chap. 5 contains a brief historical account about the development of Bayesianism in the twentieth century.

[2] See Oakley and O'Hagan (2004), Bayarri et al. (2007), Chen et al. (2006, 2008), Wang et al. (2009) for recent Bayesian approaches to validation. I will return to some of this literature below.

[3] See e.g., Howson and Urbach (2006) for an extensive exposition and defense of Bayesian epistemology; see Strevens (2006), Hartmann and Sprenger (2010), Hájek and Hartmann (2010a), Easwaran (2011a), Easwaran (2011b) and Talbott (2016) for short overview articles of Bayesian epistemology. Bayesian decision theory is surveyed by Beisbart (2011) and Jensen (2011).

suggest an interpretation of probabilities. Bayesian decision theory extends Bayesian epistemology to principles of rational action. In philosophy of science, Bayesian ideas have been used to spell out notions like unification (Myrvold 2003). It is no surprise then that Bayesian ideas are controversial and much discussed in philosophy.[4] For the purposes of this chapter, we can neglect many applications and consequences of Bayesian ideas. We will focus on the core of Bayesian epistemology. The term "Bayesians" is therefore meant to refer to proponents of Bayesian *epistemology*.

Second, as far as validation is concerned, I start from the following assumptions: It is first useful to distinguish between two types of model that are associated with a computer simulation. The *conceptual model* is a model that scientists intend to use for scientific reasons, for instance, because it captures the principles that are supposed to govern the dynamics of the target system. The *computational model*, by contrast, is the model implicit in the computer code such that the simulation output provides exact solutions to the computational model. In many examples of computer simulations, the computational model differs significantly from the conceptual one because the former includes approximations not implicit in the latter. Validation is supposed to be the evaluation of the conceptual or computational model, as far as rational credibility of its results or its assumptions is concerned. The focus of my discussion will be on the credibility of model results, but I will also comment on the credibility of assumptions that enter the computer model. It is not assumed that validation only proceeds in terms of comparing simulation outputs with measured data. Nor is "validation" supposed to be a success term that only refers to a successful evaluation.

7.2 The Fundamentals of Bayesian Epistemology

7.2.1 Basic Tenets of Bayesian Epistemology

Let us explain the basic tenets of Bayesian epistemology by using Connie as an example.

One task that Connie has as a scientist is to gain new *knowledge*. Much scientific knowledge is *propositional* because it has a proposition as its content. The content of such knowledge can be expressed in terms of that-clauses. Connie's propositional knowledge includes, e.g., knowledge that the mean global temperature has already risen during the past century. The conclusion that Connie might want to draw from her simulation would at best constitute propositional knowledge too.

Most epistemologists take it that knowledge is a species of *belief*. According to the traditional definition of knowledge, famously mentioned in Plato's dialog "Theaetetus," knowledge is justified, true belief (Plato 2015, 201c–d). So, Connie's belief that the mean global temperature has already risen in the past century only

[4] See e.g., Howson and Urbach (2006) for a defense of Bayesianism and Earman (1992) for a more critical evaluation.

constitutes knowledge since it is true and Connie is justified to believe it. The latter condition is fulfilled because there is a lot of evidence for this. Challenges to the traditional definition of knowledge due to, e.g., Gettier (1963) are not relevant for our purposes.

Whereas belief has often been taken to be an all-or-nothing affair, Bayesian epistemologists assume that belief comes in *degrees (of belief)*. That this is correct is indicated by everyday talk. For instance, we may say that Connie believes *strongly* that ..., or that she takes this to be *very likely*. Note that justification and evidence come in degrees too, and if justification is a matter of degree, so should be belief, because it seems sensible to require that we believe a proposition to the extent to which we are justified in doing so.

Bayesians think further that degrees of belief are *quantitative* and can be measured using real numbers.[5] One method to determine the precise degree of Connie's belief that proposition q holds true concentrates on what she takes to be fair bets on q. Clearly, if Connie feels 100% sure that q is the case, she will take bets unfair that offer a prize of 1 unit if q is true, but cost either more, or less, than 1 unit. More generally, Bayesians can define Connie's degree of belief in a proposition q as her *fair betting quotient* for a bet on q (see e.g., de Finetti 1937, p. 62 or Gillies 2000, p. 55). Here, the *betting quotient* is the ratio of the costs for accepting the bet (i.e., the money that is lost if the bet is lost), on the one hand, and the stakes (i.e., the prize), on the other hand. The betting quotient is *fair* for an agent if she would accept the bet even if she did not know whether she would have to buy the bet or sell it to another person.

This way of measuring degrees of belief may be motivated as follows. Assume that Connie thinks that q is true with a belief strength of $p(q)$ (whatever this means in detail). Suppose further that Connie is offered a bet on q and that she has to pay a price of l to bet. When she is faced with the decision on whether to bet on q, it seems rational to consider the money she can expect from taking the bet and to check whether it is larger than what she can expect without taking the bet. If Connie wins the bet, she will obtain the stakes s minus the costs for accepting the bet: $s - l$. If she loses the bet, she will just have to pay l. The betting quotient is l/s, and the expected monetary value of accepting the bet, given Connie's belief, is $p(q) \times (s - l) - (1 - p(q)) \times l$, which reduces to $(p(q) \times s - l)$. This expectation value is nonnegative, and thus larger than the expectation value in case Connie does not take the bet, if the betting quotient l/s is no larger than $p(q)$. So Connie will only buy the bet if the betting quotient is no larger than her degree of belief. Likewise, she will only *sell* the bet if the betting quotient is no smaller than her degree of belief. So, her fair betting quotient is her degree of belief assuming that she maximizes the expected value of money.

[5] See Ramsey (1926), de Finetti (1931a, b, 1937), Savage (1972) for important original sources on the measurement of degrees of belief and Gillies (2000), Chap. 5 and Hájek (2012), Sect. 3.3.2 for secondary literature. There are subtle differences between several proposals to measure degrees belief; they can be ignored in what follows.

The idea now is to obtain Connie's degree of belief on q by finding out her fair betting quotient on q, even if she isn't aware of any degree of belief and if she doesn't run through the calculation. Very roughly, Connie is asked which bets she would take part in if she didn't know whether she would be the seller or the buyer. Bayesians thus conceptualize Connie's betting behavior by reading the maximization of expected monetary value into it and obtaining a degree of belief from it. The so-called representation theorems from decision theory state necessary and jointly conditions for this to be possible (see e.g., Gillies 2011). Bayesians think further that there are facts of the matter of what Connie would take to be fair bets even if we wouldn't ask her, and thus matters of fact what precisely her degrees of beliefs are.

Some Bayesians think that the bets used for measuring degrees of belief should not be about money, but rather about what Connie takes to be valuable, where the value is quantified using *utilities* (see e.g., Gillies 2000, p. 56 for some discussion of this issue). As, e.g., Ramsey (1926) and Savage (1972) have shown, it is possible to obtain both the degrees of belief and the utilities of Connie from her behavior, given that the latter respects certain constraints. This is the central result from *Bayesian* decision theory, which characterizes rational choice in terms of maximizing expected utility. For the most part of this chapter, we will not use Bayesian decision theory, so it does not matter for our purposes whether we measure degrees of belief using bets about money or value.

The next step for Bayesians is to argue that degrees of beliefs are *probabilities*. The latter obey the axioms of the probability calculus as , e.g., specified by Kolmogorov.[6] Following the axioms, probabilities take values from the interval [0, 1]; a tautological proposition (e.g., that it rains, if it rains) has a probability of 1; and probabilities of incompatible propositions q_1 and q_2 add up to the probability of the disjunction $q_1 \vee q_2$:

$$p(q_1 \vee q_2) = p(q_1) + p(q_2)\ . \tag{7.1}$$

To justify the claim that degrees of belief are proper probabilities, Bayesians use *Dutch book arguments*. The latter show first that a person will lose money for sure if she bets on beliefs the degrees of which are not probabilities. For instance, if Connie's probability for q is larger than one, she is willing to pay more than the stakes for a bet, and she will lose money for sure because she can maximally get back the stakes. Likewise and conversely, one can show that Connie does not lose money for sure if her degrees of belief are probabilities.[7]

Now, it seems irrational to take bets to be fair, although one can lose money with them for sure. The proper conclusion of the Dutch book argument is thus that the

[6] The Kolmogorov axioms of the probability calculus (Kolmogorov 1956, Chaps. 1 and 2) were originally defined for probabilities of so-called events, which are introduced using a sample space. We have here adapted the axioms for probabilities of propositions, which is more natural for our purposes (see Howson and Urbach (2006), p. 14 for the relationship between events and propositions). There are slightly different sets of axioms for the probability calculus, see e.g., Gillies (2000), pp. 65–69 for an example.

[7] A Dutch book argument can be found in Ramsey (1926). For a short proof of Dutch book arguments see e.g., Mellor (2005), pp. 69–70.

degrees of belief of a rational agent are probabilities. At this point, there is a normative twist to Bayesian epistemology because the tenets of Bayesian epistemology hold only true of rational agents. The principles can thus be understood as telling what agents ought to think rationally. This normative twist is not a problem because epistemology of science is to some extent normative anyway; it is concerned with what scientists should infer, or how knowledge may be justified, which are clearly normative matters.

As is common, we will call the degrees of beliefs of a rational agent *her probabilities*. To stress that we think about a rational agent, we will sometimes talk of rational degrees of belief or credibility too.

The axioms of the probability calculus restrict the degrees of belief of Connie at one time, but do not fix them uniquely. Nor do they specify how degrees of belief ought to be changed. In the latter respect, Bayesians assume that probabilities should be *updated* as new data/evidence is obtained. To explain this, we first need some additional probabilities.

As defined so far, the probabilities that measure degrees of belief are *unconditional*. A *conditional* probability of the form $p(q_1|q_2)$, i.e., the probability of q_1 given q_2, can be defined in an analogous manner by considering conditional bets, i.e., bets that are called off in case the condition is not fulfilled (see Ramsey 1926, p. 180 for this idea; see Döring 2000 for discussion). Using an additional Dutch book argument (see e.g., Gillies 2000, pp. 62–64 and Döring 2000, p. 397), one can then show that conditional and unconditional probabilities relate to each other as follows:

$$p(q_1 \wedge q_2) = p(q_1|q_2)p(q_2) \; . \tag{7.2}$$

(this is sometimes taken as the definition of conditional probability in case $p(q_2) \neq 0$). Given this result, it is easy to prove *Bayes' theorem* (e.g., Joyce 2016):

$$p(q_1|q_2) = \frac{p(q_2|q_1) \times p(q_1)}{p(q_2)} \; . \tag{7.3}$$

Although this theorem is much used in Bayesian epistemology, it is not characteristic of it, but rather a theorem from mathematical probability theory.

To turn back to updating, assume that Connie is rational. Denote her probabilities as $p(\cdot)$. Suppose further that she observes that e is the case (here e is taken to constitute evidence and thus abbreviated with “e”). According to Bayesian epistemology, Connie should update her *prior probabilities* $p(\cdot)$ by replacing them by *posterior probabilities* $p'(\cdot)$ that take into account the evidence. It is often suggested that this should be done via a rule called *Bayesian conditionalization*: For any proposition q, the posterior probability is set to be

$$p'(q) = p(q|e) = \frac{p(e|q)}{p(e)} \times p(q) \; . \tag{7.4}$$

Here and below, we assume that probabilities in denominators do not equal zero.

According to Bayesian conditionalization, the posterior probability is a multiple of the prior probability. If the latter is zero, the posterior probability will be zero too. Otherwise, the prior probability is multiplied with the factor $\frac{p(e|q)}{p(e)}$. Here, $p(e|q)$ is called the *likelihood*. It measures how likely the evidence is taken in view of a hypothesis q. If a hypothesis entails the evidence, i.e., if the evidence follows from the hypothesis, then the likelihood is 1.[8]

Bayesian conditionalization has some desirable features. Note first that, qua being conditional probabilities, the posterior probabilities are in fact probabilities, i.e., they obey the axioms of the probability calculus. Second, the posterior probability for the evidence equals 1. In this sense, the evidence is taken into account. Third, any proposition q that is incompatible with the evidence e is given zero probability. Thus, falsification, i.e., the rejection of a hypothesis in terms of observations that are excluded by it, is a special case of Bayesian conditionalization.

All this is desirable, but does not suffice to justify Bayesian conditionalization. Lewis (1997) has thus provided a Dutch book argument for Bayesian conditionalization (see also Teller 1973), but the argument has remained controversial (see Mellor 2005, pp. 119–120 for a brief discussion). So, Bayesian conditionalization has remained controversial too (see e.g., Howson and Urbach 2006, pp. 80–85). Nevertheless, Bayesians agree that degrees of beliefs should be updated in some way when evidence comes in. In the following, we will stick with the simple rule of Bayesian conditionalization mentioned above.[9]

Suppose now that Connie's prior belief in q was strengthened by updating with evidence e. This is to say that

$$p'(q) = p(q|e) > p(q) \; . \tag{7.5}$$

We may then say that q has been *confirmed* by the evidence e. Likewise, q is disconfirmed via e if, and only if,

$$p'(q) = p(q|e) < p(q) \; . \tag{7.6}$$

Note that confirmation, as defined here, is a matter of degree and doesn't amount to proof (see Hempel 1945, Sect. 1).

The condition for Bayesian confirmation holds if, and only if, the factor $\frac{p(e|q)}{p(e)}$ is larger than 1, i.e.,

$$p(e|q) > p(e) \; . \tag{7.7}$$

This is to say that the hypothesis q makes e more probable than it would otherwise be. Thus, e confirms a hypothesis q if, and only if, the latter makes e more likely.

[8] Quite generally, if q_1 entails q_2, then $p(q_2|q_1) = 1$. The reason is that, in this case, $p(q_1 \wedge q_2) = p(q_1)$.

[9] A simple generalization of Bayesian conditionalization was provided by Jeffrey (1967), Chap. 11. Howson and Urbach (2006), pp. 80–85 argue for a different approach, which reduces to Bayesian conditionalization under certain conditions.

To summarize, according to Bayesian epistemology, rational beliefs come in degrees, which are restricted by the axioms of the probability calculus and conditionalization (or some alternative form of updating).

7.2.2 A Brief Discussion of Bayesian Epistemology

But how well does Bayesian epistemology fare as a theory in epistemology of science (see Earman 1992, Hájek and Hartmann 2010a, pp. 87–102 or Bartelborth 2013 for a much more detailed evaluation)?

On the positive side, Bayesian epistemology starts out with a more nuanced account of belief than traditional epistemology because it allows for degrees of belief. Since it takes degrees of rational belief to be measurable in terms of probabilities, Bayesian epistemology can draw on the rich resources from probability theory. This allows for powerful applications of Bayesian epistemology. One example is Bayesian statistics, which is a rival to the classical/error-statistical/frequentist approach to statistics (see Howson and Urbach 2006, Chap. 8 and Lee 2012 for Bayesian statistics; see Mayo 1996 for a philosophical defense of some sort of error statistics). Other examples of applications include measures of confirmation (see e.g., Huber 2007, Sect. 6b for an overview) and measures of coherence (see e.g., Bovens and Hartmann 2004). For another advantage, Bayesian epistemology is intimately connected to Bayesian decision theory (see Beisbert 2011 for an introduction). As a consequence, Bayesian degrees of belief have a characteristic role to play in decisions. Taken together, Bayesian epistemology and Bayesian decision theory provide a picture of rationality that covers both the rationality of belief formation and of action. Further, Bayesian epistemology can account for a number of intuitions about rational belief formation and scientific methodology. For instance, Bayesianism can explain why a hypothesis q seems particularly strongly confirmed if it predicts some event that seems very unlikely, but then happens to occur.

But the power of Bayesian epistemology to account for intuitive judgments about rational belief formation is limited. One famous problem in this regard is the *problem of old evidence* (Glymour 1980): Suppose that we have recently constructed a new theory q without drawing on evidence e, which we did know though when constructing the theory. As a consequence, if q happens to account for e by entailing it, this does not boost our degree of belief in q. For

$$p(q|e) = p(e|q) \times p(q)/p(e) \; . \tag{7.8}$$

Since e is already known, $p(e) = 1$; and since q entails e, $p(e|q) = 1$ too. Thus, q is not confirmed by e. This is not as it should be because, e.g., the credibility of

Einstein's General Theory of Relativity was increased because it was able to account for well-known data.[10]

There are other problems about Bayesianism. A first complaint is that it includes too many degrees of freedom to be useful (Bartelborth 2013, pp. 28–30). The reason is that an agent needs to start with a joint probability distribution over all propositions under consideration. Note, however, that the number of degrees of freedom shrinks considerably if groups of propositions are supposed to be independent from each other. Here, two propositions q_1, q_2 are probabilistically independent if, and only if,

$$p(q_1 \wedge q_2) = p(q_1) \times p(q_2) \ . \tag{7.9}$$

Relations of probabilistic independence can well be represented using Bayes nets; see Neapolitan (2004).

Second, Bayesianism is demanding in that it expects logical omniscience from a rational agent. The reason is that, following the axioms of the probability calculus, every tautology is supposed to have a probability of 1.

The most important objection against Bayesian epistemology is presumably that it is too subjectivist. The principles of rational belief formation acknowledged by Bayesians only fix probabilities up to some priors. Thus, different agents may rationally start out with very different priors and thus arrive at different posteriors. As a further consequence, the agents may disagree on whether a hypothesis q is confirmed by evidence e. For it is ultimately a matter of the priors whether Eq. (7.5) or (7.6) holds in a particular case.

In response, Bayesians can to some extent alleviate the worry about subjectivism by using convergence results: Under suitable conditions, the posteriors converge to each other when they are updated using the same data (see Savage 1972, pp. 46–50 and Blackwell and Dubins 1962 for such results). But convergence may be slow, so two rational agents may significantly differ on their degrees of belief although they have incorporated the same evidence.

To avoid subjectivism, Bayesians can also try to turn objectivist by expanding their repertoire of principles of rational belief formation. Some further principles, e.g., the Principal Principle (see Lewis 1980, 1994, pp. 483–490), are uncontroversial among Bayesians, but not sufficient to uniquely fix all priors. One famous candidate that might do this is the Principle of Insufficient Reason (also called Principle of Indifference).[11] Following this principle, the prior probabilities of n exhaustive, mutually exclusive hypotheses should be set to $1/n$, if there are no reasons to favor some hypotheses over others. But the application of the principle leads into paradox (see e.g., Gillies 2000, pp. 37–49). So the principle is not widely adopted, nor is objective Bayesianism.

[10] See e.g., Wagner (1997), Sprenger (2015), Wenmackers and Romeijn (2016) for attempts to solve the problem. See also Chap. 41 by Frisch in this volume.

[11] See e.g., Keynes (1921), p. 42 for an influential statement. The maximum entropy approach famously advocated by E. T. Jaynes generalizes the principle Jaynes (1957, 1968, 1979).

The only strategy left then is to bite the bullet and to argue that alternative epistemological frameworks do not really deliver more objectivity, since they are ultimately built on some arbitrary assumptions too (Howson and Urbach 2006, p. 265). The latter point is often made in the foundations of statistics, where Bayesian statistics is rivaled by frequentist or error statistics. Under the latter account of statistics, certain hypotheses should be rejected given suitable evidence, independently of what the working scientists think. Bayesians reply that error statistics is ultimately based upon decisions that cannot be justified. For example, the so-called *p*-value or the level of significance, which determines when a hypothesis is rejected under the error-statistical view, is only set conventionally, or so the charge is (see Howson and Urbach 2006, Chap. V for a thorough analysis of the assumptions underlying frequentist statistics).

To summarize, we may say that Bayesian epistemology has both astonishing achievements and dramatic shortcomings. So, a nuanced attitude toward Bayesianism is called for (see Hájek and Hartmann 2010b, pp. 100–101 and Bartelborth 2013, pp. 61–66 for examples). In this respect, it seems promising to restrict the application of Bayesian epistemology to local settings, in which only few rival hypotheses are under consideration and thus have nonzero probability. This circumvents the problem of the many degrees of freedom. Likewise, in a more local setting, there may be good reasons to set the priors in this or this way, for instance, because the priors reflect expert judgment.

7.3 Bayesian Epistemology and the Validation of Computer Simulations

Turn now to the validation of computer simulations. How should scientists reason during validation, if we adopt a Bayesian perspective?

To answer this question, Bayesians will, in general, proceed as follows: They identify suitable propositions related to a simulation, consider rational degrees of belief in these propositions and apply the Bayesian principles. The resulting posterior probabilities express the degree of trust that they should rationally invest in the propositions. These probabilities may eventually be used to maximize an expected utility related to the model, but in what follows, we will neglect this possible step.

What then are propositions that are relevant to the validation of a simulation? Natural candidates state the *results* of a computer simulation. In what follows, these may be results that *have been* obtained by actually running the program, or results that *may be* obtained in this way.

In our example, Connie's simulation produces the result that the mean global temperature increases to a certain extent in a business as usual scenario. More generally, propositional results can be obtained from a simulation as follows: The computer simulation outputs numbers that are interpreted to be the values of characteristics about the target system, e.g., mean global temperature. Further, the numbers were obtained given some input, which corresponds to some assumptions about, e.g., the initial state. So a propositional result might be that, for such and such an initial state, the value of some characteristic is such and such.

If the possible values of a characteristic are from a continuous rather from a discrete set, e.g., from the interval [1, 2], then there is almost no chance that a result as constructed so far is true, because the number output is affected by all kinds of errors. What is only sensible to expect is that the result holds up to some error, or with some accuracy (see Chap. 2 by Beisbart in this volume). Clearly, that the value of a characteristic is in a certain range of values is a proposition too that can be considered by Bayesians.

But Bayesians have an alternative to deal with characteristics with a continuous range of values. The trick is to consider *probability densities* rather than probabilities. Very roughly, for an infinitesimal interval $\mathrm{d}T$, the probability to find the temperature in the interval $[T, T + \mathrm{d}T]$ equals the probability density times $\mathrm{d}T$. For a finite interval or temperatures $[T_1, T_2]$, the probability is accordingly obtained by integrating over the probability density. Probability densities are easily generalized to higher dimensional spaces (i.e., vector spaces).[12] Of course, in Bayesian epistemology, probability densities too reflect degrees of belief. Rules such as Bayes' theorem hold for them too. To keep the presentation simple in what follows, we will thus use $p(\cdot)$ to either denote a probability or a probability density, as required in the example.

Often, simulation scientists will content themselves with more qualitative propositions about the simulated systems as result, e.g., that the temperature increases. In the terms of (Bogen and Woodward 1988), we may say that scientists construct phenomena out of the (simulated) data. Related results can be formulated in terms of propositions too.

Let us thus assume that results have been formulated in a set of propositions.[13] The next step then is to set values on the probabilities of the propositions.

7.3.1 Data-Driven Validation

A natural way to arrive at such probabilities is to use updating. For this purpose, simulation results need to be compared with data. In what follows, we'll first present a very simple toy example of updating and then expound a more sophisticated method.

For the toy example, assume that scientists are interested in a finite number of qualitative results. We can conjoin the related propositions in one big proposition, call it r, and consider the probability of r. To update this probability, scientists can obtain evidence on some of the results, call it e. Using Bayesian conditionalization, we obtain for the posterior probability

[12] See Papoulis and Pillai (2002), Sect. 4.2 for an exact definition of probability densities.

[13] See Beisbart (2012) for a more extended discussion of how computer scientists obtain propositional content from simulation output.

$$p'(r) = p(r|e) = \frac{p(e|r)}{p(e)} \times p(r)\ . \tag{7.10}$$

In this setting, there are only two possibilities: Either the evidence is incompatible with the results r, then $p(e|r)$ is a multiple of $p(e \wedge r) = 0$, so the results are falsified and obtain zero credibility. Alternatively, the evidence is compatible with r. Then, $p(e|r) = 1$ because the results imply e, and the credibility of the whole results r becomes:

$$p'(r) = \frac{p(r)}{p(e)}\ . \tag{7.11}$$

This probability is larger than the prior, if the prior probability of the evidence was less than one (which is natural; otherwise, there was no point in gathering evidence). To determine quantitatively by how much the probability of the results is boosted, we need to obtain $p(e)$. This is most naturally done under the assumption of a restricted range of possible results (or hypotheses) r_j ($j = 1, ..., o$). The results r_j are assumed to be mutually incompatible and jointly exhaustive, which is to say that other hypotheses have no credibility. The r_j may result from other computer simulation models. We then have

$$p(e) = \sum_j p(e|r_j)p(r_j)\ . \tag{7.12}$$

So using Bayesian conditionalization (or, maybe, some other type of updating), we obtain a natural method for *data-driven validation.*

If we do not want to rely on a set of mutually exclusive hypotheses that exhaust the space of what we take to be credible, we may at least use Bayesian updating to compare two hypotheses regarding their posterior probabilities. If we form the ratio of the probabilities, then $p(e)$ drops out, and we obtain a comparative assessment of the credibility of both hypotheses (results from different simulations). It turns out, however, that, in our toy example, the ratio does not change for two results that are compatible with the existing data. This is good reason to turn to a more powerful Bayesian method of data-driven validation, as proposed by Bayarri et al. (2007).[14] We here present a rough outline of the method.

The results considered by the method are of the type that a certain characteristic takes an output number as value. Thus, if the possible values of the characteristic are continuous, we need probability densities in what follows. We let $p(x)$ denote the probability (density) that a certain characteristic X takes value x.

Absorb all input that is needed to run the simulation program once into a vector $\mathbf{x}$.[15] Here, the input comprises initial conditions and, maybe, some parameter values. Collect some subset of the simulation output that is supposed to be of interest in another vector $\mathbf{y}$. If the simulation is deterministic, i.e., if merely one output is possible given an input, the computer simulation induces a function that maps input

[14] Earlier versions of this work have been around as Bayarri et al. (2002, 2005).

[15] As usual in, e.g., physics, vectors and vector-valued functions are denoted using boldface letters.

to relevant output:

$$\mathbf{f}^{\mathrm{sim}} : \mathbf{x} \mapsto \mathbf{y} = \mathbf{f}^{\mathrm{sim}}(\mathbf{x}) \ . \tag{7.13}$$

The method to be described presently can naturally be generalized to nondeterministic, so-called stochastic simulations.[16]

Assume that, for each simulation input **x**, there is a real-world counterpart for the output. It comprises those values of the selected characteristic that would arise in reality, if **x** described the initial conditions and parameter values in reality:

$$\mathbf{f}^{\mathrm{real}} : \mathbf{x} \mapsto \mathbf{y} = \mathbf{f}^{\mathrm{real}}(\mathbf{x}) \ . \tag{7.14}$$

We assume once more that the output is unique, which means that the system is deterministic too.

The difference between the simulated output and the real counterpart is called *bias*:

$$\mathbf{b}(\mathbf{x}) := \mathbf{f}^{\mathrm{real}}(\mathbf{x}) - \mathbf{f}^{\mathrm{sim}}(\mathbf{x}) \ . \tag{7.15}$$

Bias is a function of the input.[17]

Assume now that there are measured data $\{\mathbf{y}_i^d\}$ for a set of inputs to the simulation program $\{\mathbf{x}_i\}$ $(i = 1, ..., n)$. Since there are experimental errors $\mathbf{e}_i$, the data points will in general not coincide with the true values for **y**:

$$\mathbf{y}_i^d = \mathbf{f}^{\mathrm{real}}(\mathbf{x}_i) + \mathbf{e}_i \ . \tag{7.16}$$

The proposed method uses these data to obtain a posterior probability density for $\mathbf{f}^{\mathrm{real}}(\mathbf{x}^{\mathrm{new}})$ for some arbitrary new input to the simulation program $\mathbf{x}^{\mathrm{new}}$. So the idea is that the program is run with a new input $\mathbf{x}^{\mathrm{new}}$ and produces $\mathbf{y}^{\mathrm{new}}$ as output. Our aim is to derive the posterior probability density for the real counterpart to that output:

$$p'(\mathbf{f}^{\mathrm{real}}(\mathbf{x}^{\mathrm{new}})) \ . \tag{7.17}$$

Now, due to Eq. (7.15),

$$\mathbf{f}^{\mathrm{real}}(\mathbf{x}^{\mathrm{new}}) = \mathbf{f}^{\mathrm{sim}}(\mathbf{x}^{\mathrm{new}}) + \mathbf{b}(\mathbf{x}^{\mathrm{new}}) \ . \tag{7.18}$$

So, the posterior probability density for $\mathbf{f}^{\mathrm{real}}(\mathbf{x}^{\mathrm{new}})$ is basically the one of the bias, shifted by the simulation output $\mathbf{f}^{\mathrm{sim}}(\mathbf{x}^{\mathrm{new}})$. Thus, to obtain the probability density

[16] In this case, $\mathbf{f}^{\mathrm{sim}}$ and possibly $\mathbf{f}^{\mathrm{real}}$ defined below should be regarded as random functions. The method doesn't essentially change if this is so because we will below assume probabilities for the functions anyway.

[17] The assumption that there is a real-world counterpart $\mathbf{f}^{\mathrm{real}}(\mathbf{x})$ to the model for each input **x** is not appropriate for all models. We do not expect a real counterpart, for instance, if **x** contains parameters that do not really correspond to characteristics in reality.

for the value of $p'(\mathbf{f}^{\text{real}}(\mathbf{x}^{\text{new}}))$, we need only the posterior probability density for the bias.[18]

To obtain this, Bayesians have to assume a prior probability for the unknown quantities, viz., the measurement errors $\{\mathbf{e}_i\}$ and the bias $\mathbf{b}$. Since the bias is a function of input, the bias has to be thought of as a random function, for which we need a probability density.[19] We condense the priors of the unknown quantities in one joint probability density

$$p\left(\mathbf{b}, \{\mathbf{e}_i\}\right) \ . \tag{7.19}$$

The posterior for these quantities thus is by Bayes' Theorem

$$p\left(\mathbf{b}, \{\mathbf{e}_i\}|\{\mathbf{y}_i^d\}\right) = \frac{p\left(\{\mathbf{y}_i^d\}|\mathbf{b}, \{\mathbf{e}_i\}\right) \times p\left(\mathbf{b}, \{\mathbf{e}_i\}\right)}{p\left(\{\mathbf{y}_i^d\}\right)} \ . \tag{7.20}$$

Here, $\mathbf{y}_i^d$ is uniquely determined, given a value for the bias and the error, once $\mathbf{f}^{\text{sim}}(\mathbf{x}_i)$ is known, which can be obtained by running the simulation program. This means that $p\left(\{\mathbf{y}_i^d\}|\mathbf{b}, \{\mathbf{e}_i\}\right)$ reduces essentially to a delta function. The denominator can in principle be determined as in Eq. 7.12, if we restrict ourselves to some set of models.

From $p\left(\mathbf{b}, \{\mathbf{e}_i\}|\{\mathbf{y}_i^d\}\right)$, we obtain a probability density over the bias by integrating over the $\{\mathbf{e}_i\}$. Evaluating the bias function at $\mathbf{x}^{\text{new}}$ gives the bias of the model we can expect for input $\mathbf{x}^{\text{new}}$. By adding it to $\mathbf{f}^{\text{sim}}(\mathbf{x}^{\text{new}})$, which can be known from running the simulation program, we obtain a degree of belief for the true counterpart to $\mathbf{y}^{\text{new}}$, $\mathbf{f}^{\text{real}}(\mathbf{x}^{\text{new}})$. We may use it to calculate the mean value of $\mathbf{f}^{\text{sim}}(\mathbf{x}^{\text{new}}) + \mathbf{b}(\mathbf{x}^{\text{new}})$ and the variance or specify credible intervals, i.e., intervals in which the true counterpart must be with high probability.

More technical details about the method can be found in Bayarri et al. (2007). Note that we have neglected two complications of their work. First, they also include calibration, which is not part of validation, properly speaking. Second, they assume that a so-called emulator of the simulation program is used (see Chen et al. 2006, 2008; Wang et al. 2009 for very similar approaches). That is, a probabilistic model for the function that maps simulation input to simulation output is introduced and updated using actual outputs from a few runs. The model can then be used instead of the computer program to save computational costs. Although the method is Bayesian, it is not relevant for our purposes because it is not a necessary ingredient of validation.

Of course, the method specified above only works given a prior $p\left(\mathbf{b}, \{\mathbf{e}_i\}\right)$. It is natural to assume that the prior of the unknown quantities, Eq. (7.19), factorizes because the experimental errors and the bias function are independent. Bayarri et al. (2007) assume that the bias function is a Gaussian random field with some unknown parameters, for which prior probabilities are assumed. The idea here is to have a

[18] Here possible errors in measuring $\mathbf{x}$ have been neglected because this would complicate things further.

[19] Technically speaking, this probability density lives on a different space than the probability densities of outputs, but this detail need not detain us.

broad set of bias functions available and to use the data via updating to restrict the range of the bias.

All in all, we have a Bayesian method for a data-driven validation of a whole simulation program. Note that the application of the method works even if the $\mathbf{y}_i^d$ cover only some aspects of the relevant results. In this way, the application of the method need not include a comparison between simulation output and data for those characteristics that are of ultimate interest (i.e., the mean global temperature in our example); rather, the comparison may be carried out using different characteristics (say precipitation in our example).[20]

7.3.2 *The Problem of the Priors in Validation*

The rational degrees of belief obtained with the method proposed by Bayarri et al. (2007) depend on priors. As mentioned, the authors make a proposal how to set the priors, which suffices for a concrete application (see our p. 14). But the priors were not determined in a principled way, so different priors are possible. Of course, in some sense, this problem will not disappear because, at some point, the choice of priors is inevitable in Bayesian epistemology, unless some version of objective Bayesianism is adopted. Nevertheless, we may ask whether the priors can arise from the application of principles from Bayesian epistemology on yet other priors. Ultimately, the question is whether validation (in the broad sense defined in the introduction) can draw on sources that are different from mere comparison between simulation outputs and data. This question is of course not specific to a Bayesian approach to validation.

A very natural answer is that most results from computer simulations do have credibility that does not derive from a direct comparison with data, because the simulations are based upon theories and further assumptions that enjoy independent credibility. For instance, Connie's simulations assume the Navier–Stokes equations, which clearly have independent credibility. This credibility ultimately depends on data too, because theories and other assumptions from the natural and social sciences only obtain credibility if they can account for some data. But these data are for the most part very different from data to which simulations may be compared to. For instance, the Navier–Stokes equations have been tested in a variety of settings that are not at all in the scope of applications of Connie's simulations, e.g., in small-scale experiments that are not resolved in global circulation models.

[20] There are a few more concrete proposals how to use the notions from Bayesian epistemology for tasks related to validation. Kennedy and O'Hagan (2000) offer a Bayesian analysis that allows simulation scientists to work with approximations to complicated models. Kennedy and O'Hagan (2001) study calibration from a Bayesian perspective. Sensitivity analysis is investigated from a Bayesian perspective by Oakley and O'Hagan (2004). All of these works use so-called statistical emulators for the original computer simulation. A useful tutorial to a Bayesian analysis of simulation outputs is given by O'Hagan (2006). In this tutorial, the application of Bayesian methods to validation is named as an example, but not further pursued (p. 1299).

Bayesian epistemology can in principle account for the prior credibility of theories and assumptions that enter a simulation. In the Bayesian framework, the prior credibility of a theory or an assumption can be measured using a rational degree of belief from a scientist.[21] And this degree of belief will be informed by all kinds of data. Given that theories or assumptions like the Navier–Stokes equations have been tested in numerous ways, it is in fact realistic to expect that scientists more or less agree on the credibility of many assumptions.[22] Of course, a simulation may incorporate very controversial assumptions, maybe the point being that the simulations are used within a test of the assumptions, but in this case, the results of the simulations should be conditioned on the controversial assumptions.

Let us thus explore how Bayesians can quantify the prior credibility of simulation results. As before, call r a proposition that conjoins many results from a simulation. Let us for definiteness assume that the simulation is built on a theory t and auxiliary assumptions a. Apart from the theory, we include all assumptions on which the simulation is built, e.g., assumptions about parameter values and initial conditions, in a. Let us call the conjunction of t and a the underlying model m. Assume that the scientists from a field further agree on priors for the theory and the assumptions, $p(t)$ and $p(a)$. Suppose further, that the theory and the assumptions are taken to be independent, which is to say that $p(t \wedge a) = p(t) \times p(a)$. So, $p(m) = p(t) \times p(a)$.[23] Now the computer simulation is supposed to trace the consequences of the model assumptions (see Beisbart 2012 for discussion). If it does, the results are entailed by the model and thus $p(r|m) = 1$. As a consequence, the credibility of the results is at least as high as the credibility of the model:

$$p(r) \geq p(r \wedge m) = p(r|m) \times p(m) = p(m) = p(t) \times p(a) \; . \tag{7.21}$$

$p(r)$ will be strictly larger than $p(m)$ if there are alternative credible models that imply the same results.

But this analysis is threatened by two problems. *First,* even computer simulations that are not designed to test controversial assumptions often rely on assumptions that are not thought to be realistic. Most simulations are based upon abstraction, idealization and approximation; as a consequence, some of their assumptions are strictly speaking taken to be false. For instance, simulations that presuppose the Navier–Stokes equations are ultimately built on Newtonian physics, which is strictly speaking not correct, since relativistic effects are neglected (despite this, Newtonian physics is still a very useful approximation). But if an assumption is taken to be false, its prior probability is zero. And this would mean that the prior probability of

[21] In what follows, we adopt the standard assumption that a theory can be expressed in terms of one or more propositions.

[22] The idea need not be that each scientist has updated her belief many, many times. Rather, she may have taken it from other scientists in her training. Taking priors from textbooks or other reliable sources is certainly a reasonable way to set one's priors.

[23] More realistically, there will be some model assumptions that are very credible, while there are uncertainties about some other model assumptions. In this case, a probability model for the uncertainties seems more useful.

our whole model $p(m)$ is zero, which would certainly not provide a good basis for putting credibility on the results of the simulations.

This problem arises if we are too strict about the model assumptions. In practice, e.g., model equations are only taken to be credible up to some errors. For instance, the Navier–Stokes equations are not supposed to hold literally, but only to some approximation. Accordingly, solutions to the Navier–Stokes equations will not state the literal truth about the dynamics of a fluid, but this is not an issue as long as some limited imprecision in the solutions is tolerated. A Bayesian analysis has to take this into account.

The *second problem* with the analysis above is as follows. What scientists have trust in are typically conceptual models. But a computer simulation does not exactly trace implications of a conceptual model because the latter is at best approximated in the simulations. The reasons are well known: Differential equations are approximated using difference equations; other types of numerical approximations are used, e.g., in Fourier transformations; round off errors arise; the computer program may contain so-called bugs; and the hardware may not function as intended. All this yields errors of the result with respect to the conceptual model, and this is the reason why we describe the simulation itself by a computational model (see Chap. 5 by Roy in this volume). So r may not follow from m. But then $p(r|m) \neq 1$. In fact, if we understand the result as saying that a characteristic takes the particular value provided by the output of the simulation, and if, due to errors, this value does not coincide with what the model implies, then r and m are incompatible and $p(r|m) = 0$. Again, this is not a good basis for obtaining prior credibility for the results.

This problem arises to some extent if we are too strict about the results. Typically, the outputs of a simulation are taken with some grain of salt. For instance, if the output for the temperature is T, then a sensible result is that temperature is in a range around T, and the idea is that this result follows from the conceptual model.

To solve both problems, we may try a thorough quantitative error analysis as follows. We relax the model assumptions to avoid the first problem. For instance, if the Navier–Stokes equations are part of the model m, we relax them by assuming that they only hold up to correction terms of size ϵ_m. To do so requires additional knowledge about the sizes of possible corrections, but this knowledge is often available.

We then need to obtain bounds on the effects that corrections to the model assumptions smaller than ϵ_m make on the results. Further, to address the second problem, all other sources of errors need to be taken into account. For instance, if a discretization of differential equations leads to deviations from the original model no larger than ϵ_d, then the effects on the results are traced too. In this way, we may in principle derive an upper bound ϵ_r on the effects of all kinds of errors. We then interpret r as stating that the simulation outputs only approximate the true values of the characteristics up to ϵ_r. We then have

$$p(r|m) = 1 \tag{7.22}$$

again by construction, which gives us a lower bound on the prior probability of r as before. Note here that talk of bounds assumes suitable *metrics* that allow to measure distances in the spaces of outputs and inputs. For instance we need to be able to tell how far an output is from the true counterpart, if we wish to say that the corrections are smaller than ϵ_r in the results r.[24]

Unfortunately, the thorough quantitative analysis just sketched is impossible in practice. For one thing, the deviations from the model equations that are arise due to the various sources of errors are often not known in terms of upper bounds. For instance, bugs are typically not known at all nor are their effects. For another thing, the propagation of the errors and their interactions is very complicated and difficult to describe. Computer simulations are run precisely because we do not know what a certain model implies. That is, we do not know how the model assumptions "propagate" through the calculations of a simulation to effect results. It is roughly as difficult to know how errors propagate through the calculations or to obtain bounds on this. So we can only use simulations themselves to learn about these effects.

From a Bayesian point of view, a natural method is to *model* the errors from the simulations. We can proceed in a similar way as we did before when we followed Bayarri et al. (2007).

We assume that, using a metric, we can take distances $\delta(\cdot,\cdot)$ between arbitrary possible outputs. For an arbitrary input $\mathbf{x}$, let $\mathbf{f}^{\text{model}}(\mathbf{x})$ be the prediction of the conceptual model for input $\mathbf{x}$. Define the *model bias*, $b^{\text{model}}(\mathbf{x})$ (i.e., the bias of the simulations with respect to the conceptual model), as the distance between the simulation output and the output that the conceptual model as such would produce for input $\mathbf{x}$,

$$b^{\text{model}}(\mathbf{x}) = \delta(\mathbf{f}^{\text{sim}}(\mathbf{x}), \mathbf{f}^{\text{model}}(\mathbf{x})) \ . \tag{7.23}$$

Here, the bias function is not a vector anymore, but rather real-valued, since it just expresses how far the results of the simulation are from the model predictions in terms of our metric.

We assume a prior probability density $p^b(b^{\text{model}}(\mathbf{x}))$ over the model bias that expresses the uncertainty over model bias at $\mathbf{x}$. This prior probability density may then be updated by running the program for specific inputs $\mathbf{x}$, for which the solutions to the conceptual model are known. We thus obtain some values of the model bias to which our model for the bias can be adjusted via updating. This can be done in a way that is completely analogous to the method specified by Bayarri et al. (2007).

Suppose now that, for new input $\mathbf{x}^{\text{new}}$ to the simulation, we tolerate corrections up to ϵ_r in the results. Let r be the claim that the results r constructed from the actual output of the simulation run have corrections no larger than ϵ_r. Then we have for the probability for the result r

$$p'(r) \geq \int_{-\epsilon_r}^{\epsilon_r} \mathrm{d}\delta p^b(\delta, \mathbf{x}^{\text{new}}) p(m) \ . \tag{7.24}$$

So we obtain a lower bound on the prior for the result. The bound may be used as a basis for data-driven validation. We have a lower bound because other models with nonzero credibility may imply the same results.

[24] A metric does not presuppose quantitative output and may be applied to qualitative output too.

Call this method *Bayesian verification*. “Verification” is here meant to refer to activities that show that a simulation program appropriately traces solutions to the conceptual model. Note, however, that—as any method of verification—this one does not take into account errors in the conceptual model. The proposed method of verification is simpler than the Bayesian method for validation given by Bayarri et al. (2007), because the bias function is now real-valued and because no experimental errors need to be taken into account.

An immediate objection against Bayesian verification is that it only pushes the problem one step back. We were interested in obtaining rational priors for a result, but it turns out that we now need priors for a model bias. However, the new model for model bias may not be as contentious as m and easier to investigate, so we can reasonably expect more agreement about it.[25]

Independently of the details, Bayesian epistemology faces an interesting conceptual problem concerning verification. According to Bayesians, a rational agent has coherent probabilities in that the axioms of the probability calculus are fulfilled. Suppose now that results r follows from model assumptions m as a matter of fact. The agent should then set the probability $p(r|m)$ at 1. But in typical examples of models used in simulations, the agent does not know what the consequences of the model assumptions are. If she then assumes a probabilistic model for the bias following Bayesian verification, she will typically set a nonzero probability on results that do not follow from the model. She thus violates the axioms of the probability calculus.

Accordingly, to handle verification of computer models, Bayesians have to build what may be called a nonideal theory.[26] Such a nonideal theory allows that some principles of Bayesian epistemology are violated. Nevertheless, other principles are assumed to hold and still applied.

To summarize this section, we have first shown that there is a distinct Bayesian methodology for data-driven validation, i.e., validation that compares outputs of a simulation with data. A problem though is that Bayesian data-driven validation assumes priors. We have then seen that we can constrain such priors. The crucial idea is that the modeling assumptions on which the simulation is based have credibility. A problem is that this credibility doesn’t straightforwardly flow to the results because the simulation program at best approximates solutions to the model equations. We can use what we have called Bayesian verification to solve this problem.

As far as validation is concerned, we have so far concentrated on the validation of *results* from simulations. But we may also ask to what extent a whole simulation model (including its underlying assumptions) can be validated. The Bayesian answer is that we can update model probabilities via

[25] A slightly different regress problem arises if so-called Monte Carlo simulations are needed to calculate quantities that are used in Bayesian verification (as is quite often the case if posterior probabilities are calculated). The threat is that the Monte Carlo simulation needed to obtain some quantity is not verified. However, in practice, this is not a problem because the verification of the simulation needed for the integral is easier than the verification of the model of the simulation.

[26] The term was coined by Rawls (1971), e.g., pp. 9, 247.

$$p'(m) = p(m|e) = \frac{p(e|\, m) \times p(m)}{p(e)} \;, \tag{7.25}$$

Thus, the credibility of a model can be enhanced using data. If we want to apply something like the method used by Bayarri et al. (2007), then it is most natural to conceive of a model as a function from input to output; and the method detailed above gives us effectively the posteriors for the model as a function of $\mathbf{x}^{\mathrm{new}}$.

7.4 Discussion

What then has Bayesian epistemology to offer for Connie and to other simulation scientists? What possible problems are there? And do rival epistemological frameworks better?

To begin with the *first question*, Bayesian epistemology offers i. the notion of degrees of belief, which allows for the use of mathematical probability theory, and ii. principles that constrain the formation of rational degrees of belief. These components form a theoretical framework that may be used as a toolbox to determine rational degrees of belief on propositions related to a computer simulation program. In Sect. 7.3 of this chapter, we have used the tools from this box. There are different ways to use them, depending on the choice of the propositions that are evaluated for credibility.

That Bayesian epistemology approaches simulations and their results in terms of rational belief or credibility is very natural. For it is a natural question to ask to what extent we should believe the results of simulations.[27] Also, knowledge and rational belief belong to the aims constitutive of science.

A possible objection is that credibility is not the proper standard for assessing simulation models. For instance, Parker (2009) argues that climate models and simulations should be evaluated in terms of *adequacy for purpose*. A dependence on purposes is manifest in the use of a validation metric (see Chap. 13 by Marks and Chap. 18 by Saam in this volume). Roughly, the validation metric quantifies the distance between the outputs and data taken from the target system. There are in principle many ways how a distance may be taken, so researchers decide on the basis of their purposes which distance they choose.

But assuming that the ultimate criterion for computer simulations is adequacy for purpose need not push us outside the confines of Bayesian epistemology. Note first that, in her paper, Parker does talk about the confirmation of hypotheses (e.g., p. 233). These hypotheses hold that a simulation is adequate for a certain purpose. Clearly, we can treat such propositions in the terms offered by Bayesian epistemology and explain their confirmation in Bayesian terms. Second, and more importantly, adequacy for purpose often boils down to credibility of some propositions. For instance, if the goal is to predict precipitation to some accuracy, then the goal is reached as soon as we

[27] The title of Knutti (2008) is very telling in this respect.

obtain results with the desired accuracy that have high credibility. More generally, the main goals pursued using simulations are not arbitrary, but restricted to the gain of information constructed from simulation outputs. Since information may be more or less uncertain, the point of a simulation can only be to obtain information about a specific question with high accuracy and low uncertainty. Thus, when Bayesians quantify uncertainties about errors using probabilities, as shown in Sect. 7.3, they address adequacy for a particular purpose.

What if there are other purposes and standards that matter during the validation of computer simulations? It may, for instance, be argued that the simplicity of simulation models should be taken into account too (see Chap. 2 by Beisbart and Chap. 40 by Hirsch-Hadorn and Baumberger for discussion). But additional purposes and standards do not pose a problem for Bayesians. First, if standards such as simplicity are supposed to matter to credibility for some agents, the latter can set higher priors on models that fulfill the standards. Second, if a simulation is evaluated in terms of purposes that are not meant to matter to credibility, Bayesians can describe these purposes in terms of utilities and apply Bayesian decision theory. They may, e.g., quantify the benefits (costs) of (true) false predictions from several simulation models and then choose to work with that model that yields maximal expected net gains (see Chap. 20 by Jiang et al. for this approach). Note also that it is doubtful whether evaluating simulations following simplicity still counts as part of validation (see Chap. 2 by Beisbart for a discussion).

As far as validation metrices are concerned, they can be integrated in a Bayesian framework. Bayesians can use their framework to assess the credibility of propositions to the effect that the outputs of the simulations are such and such far from what is true about the real world in terms of a validation metric. Conversely, choices similar as those that determine a validation metrics are also implicit in the Bayesian approach. They enter via the propositions that are assessed for credibility. Further, some Bayesian quantities, e.g., likelihoods may be considered to be validation metrices themselves (this point of view is taken by Jiang et al. in their Chap. 20 in this volume).

In the literature, we find a different objection against Bayesian validation. Oberkampf and Barone (2006) suggest that Bayesian epistemology is not a suitable framework for validation because it does not take validation to be an assessment. They write (p. 10):

> From this very brief description of parameter estimation and Bayesian inference, it should be clear that the primary goal of both approaches is "model updating" or "model calibration." Although this goal is appropriate and necessary in many situations, it is a clearly different goal from that used to evaluate a validation metric. Our emphasis in validation metrics is in blind assessment of the predictive capability of a computational model (how good is the model?), as opposed to optimizing the agreement between a given model and experimental measurements.[28]

In the terms of the method proposed by Bayarri et al. (2007), the problem may seem to be that the method returns a posterior probability (density) for the true value at

[28] Cf. also Oberkampf and Roy (2010), Sect. 13.8.

$\mathbf{x}^{\text{new}}$. So the focus seems to be what we should think about the true value and not how good the model is.

But this is not the only way of looking at the method. What is first updated is not a probability of the model itself or of the value of some model parameters, but of the bias. The model is left fixed, as it were, and Bayesians identify a different (merely statistical) model that specifies how far the simulation is from reality in some respects. The larger the bias is, the worse is the original model.

One may still object against the Bayesian account of validation that it does not really evaluate a simulation because it does not lead to the acceptance or rejection of a simulation. But we did at least see the rejection of a result of a simulation in Bayesian terms on p. 11. Also, if the method proposed by Bayarri et al. (2007) leads to a probability density that is narrowly peaked about a considerable amount of the bias, then the results of the simulation are left with small credibility. This has significant effects for the credibility of the model itself. This is because

$$p(r) = p(r|m) \times p(m) \; . \tag{7.26}$$

Given that the model entails the results or at least makes them very likely (i.e., $p(r|m) \approx 1$), the credibility of the model cannot be high if the results are not credible. Further, as indicated above and as discussed in Chap. 20 by Jiang, Bayesians can use Bayesian decision theory to reject a model for further use.

This discussion naturally leads to our *second question*, viz. what possible problems there are for doing validation in a Bayesian framework.

A *first*, practical problem is that the application of a Bayesian methodology requires priors. But there are at least two strategies to set priors:

- Obtain the prior credibility of simulation results from the prior credibility of the underlying model using Bayesian verification (see Sect. 7.3.2).
- Begin with priors that cover a broad range of hypotheses, e.g., about biases, and use enough data to obtain posteriors with information specific enough for the case at hand. This strategy is used by Bayarri et al. (2007).

In some cases, expert judgment may be used too. The idea is that experts have degrees of belief in the results of simulations and that these degrees are well-informed by experience (see e.g., Bayarri et al. 2007, p. 141).

The subjectivity inherent to Bayesianism may be thought to give rise to a *second*, more conceptual problem. If validation depends on priors, which are subjective for most Bayesians, then validation itself becomes subjective and agent-relative. In particular, how good a particular simulation scores in terms of credibility, ultimately depends on the prior probabilities that were chosen. Or, if we want to say that a simulation may be more or less validated, then Bayesians will have to say that a result or model can only be validated relative to some priors. This is counterintuitive, and a more objective notion of validation would be preferable.[29]

[29] As we have indicated above, every validation presumes a standard or a validation metric, so validation is relative to the choice of such a metric. More precisely, then, the charge against Bayesians is that their validation is even relative, once a validation metric has been fixed.

In response to this problem, Bayesians may either turn objectivist or bite the bullet. In the latter case, they can alleviate some worries about subjectivity by noting that, in many practical applications, relativity is not an issue. The reason is that scientists often agree on the priors (at least to some approximation) such that agent relativity does not really matter. Bayesians who stick with subjectivism can further try to argue that, in some hidden way, alternative approaches to validation suffer from exactly the same problem because they are rooted in subjective decisions and thus relative too, as indicated above. To explore the prospects of this strategy is beyond the scope of this chapter. Note, in any case, that validation is but one among other notions that become relative if (subjectivist) Bayesianism is accepted. Another example from this class is confirmation.

A *third* problem is the old evidence problem. This problem is relevant to validation because the data-driven validation of many simulations draws on data that were already available when the simulation was produced. If the related critique of Bayesianism is right, then Bayesians cannot use old evidence to validate simulations. So Bayesianism is in stark contrast with actual validation practice, or so the charge is.

I don't think that this problem is a reason to reject Bayesianism. As indicated above in footnote 10 on p. 8, there are some proposals to solve the problem from a Bayesian perspective. To briefly indicate a direction that seems fruitful, we may say the following: The problem arises because it is assumed that, as soon as agents have come up with a new model (or theory), they immediately have a joint probability over the new model and all evidence that bears on it. But this is not realistic since agents are not logically omniscient. In particular, they do not immediately recognize the consequences of their models. So it seems adequate that agents, who construct a new model, first set some low probability on it and then update it with evidence that was already available at the time they constructed the model, but not used. The thought then is that the problem may be solved in the framework of nonideal Bayesian epistemology. In Chap. 41 in this volume, Frisch discusses some solutions to the problem too and relates them to the distinction between validation and calibration.

A *fourth* problem seems to be that validation is often done by groups of scientists, while the Bayesian account of validation seems to be individualistic and focused on the belief of individual scientists. It may thus be objected that a Bayesian account of validation is doomed to neglect a host of complications that arise due to the division of the labor of validation within larger groups, where researchers with different backgrounds and power need to collaborate. This objection is not well-founded, however. Note first that Bayesian epistemology has a clear focus on the question of how scientific inference should be done. So, Bayesian epistemology is not meant to describe all sorts of complications that arise when science is done in practice. As a normative account of scientific inference, Bayesian epistemology can be applied both at the level of an individual person or at the level of a group if this group can be said to have beliefs (see e.g., Gilbert 1987 for an account of group belief). In science, there are good reasons to think that there are group beliefs. First, publications by research teams seem to express group belief. As far as validation is concerned, in publications, the uncertainties are often expressed using probabilities, and on this basis, a certain result is reached. This is naturally understood in terms of degrees of beliefs by a group. Second, as Gillies has shown, it is in the interest of a group of scientists to bet on group beliefs that they have agreed upon (see Gillies 2000, Chap. 8). If they do not bet in this way, they can fall prey to Dutch books. This leaves open *how* the scientists from a team agree on probabilities. Questions like these are dealt within social epistemology.

A *fifth*, more profound problem is as follows: Bayesianism is built on the assumption that uncertainties are well expressed using probabilities. This assumption matters for validation because uncertainties, e.g., about the bias function are articulated in terms of probabilities. But there are good reasons to assume that uncertainty is not well expressed using precise probabilities. Consider first complete ignorance, which may be taken as a limiting case of uncertainty. Even in this case, Bayesians would try to elicit a precise probability from an agent. But this seems inappropriate. Suppose, for instance, that you travel to a country you have never been to and discover a fruit on a tree. You have no idea whether it is edible or not. Nevertheless, Bayesians would ascribe you a fixed degree of belief on the proposition that it is edible, say of 0.5. This is very different from a case, in which you put a probability of 0.5 on a specific outcome of a coin toss because you have updated your degree of belief many times on the basis of statistical evidence (see Huber 2009, Sect. 3.1 for this point). Second, Frigg et al. (2014) have argued that the use of probabilities is not efficient to trace the consequences of certain types of model uncertainty, as is done in sensitivity analysis. A related point has been made against Bayesianism more specifically by Albert (1999). All this is a good motivation to think that Bayesians do not represent uncertainty in an appropriate way and to move to a framework that is more differentiated, e.g., imprecise probabilities (see Bradley 2016 and Chap. 21 by Bradley in this volume).[30]

In response, Bayesians should admit that imprecise probabilities provide an even more differentiated option to think about uncertainties. They may hope that Bayesianism nevertheless suffices for many applications, that it is sometimes preferable because it is less complicated than imprecise probabilities and that imprecise probabilities keep a lot of features from Bayesianism.

Let us now turn to our *third question*: Are there any serious competitors for a Bayesian way of thinking of validation?

In epistemology, Bayesianism is sometimes contrasted with what may be called traditional epistemology (e.g., Bartelborth 2013). The latter is a reconstruction that captures what is common to many pre-Bayesian epistemologies, but different from Bayesian epistemology. The most basic tenet is that belief is an all-or-nothing affair, i.e., ungraded. This tenet may be complemented by principles of rationality that e.g., exclude contradictory beliefs. Conceived in this way, traditional epistemology is not a powerful rival, when it comes to validation (see Bartelborth 2013) for a

[30] Roy and Oberkampf (2010), Sect. IV.B recommend imprecise probabilities for certain kinds of uncertainties.

general comparison), simply because it does not have interesting implications about validation specifically.[31]

In philosophy of science, the most serious rival of Bayesian epistemology is falsificationism as prominently defended by Popper (1934).[32] We here take it that falsificationism wants to keep science clear from inductive inference, which is thought to be unjustifiable. Thus, the basic type of scientific inference is supposed to be a modus tollendo tollens, by which a hypothesis is falsified because one of its consequences is incompatible with an accepted basic sentence that reports observations. As far as statistical or probabilistic hypotheses are concerned, we take it that falsificationism adopts the principles of error statistics, roughly the rejection of hypothesis using a *p*-value (see Mayo 1996 for a philosophical defense). As far as validation is concerned, falsificationism recommends that validation metrics be used to formulate validation hypotheses, which should then be tested using error-statistical methods.

At least at first sight, error statistics has the advantage of avoiding recourse to priors. The question though is whether the first impression can stand reflection. It is clear that, in excluding priors, error statistics incorporates assumptions on which Bayesians do not draw. The big question then is whether these assumptions can be justified in some way. This question is at the center of a general debate between Bayesians and falsificationists.

Even if error statistics justifiably avoids recourse to priors, there are significant downsides, which correspond to advantages of Bayesian epistemology.

First, unlike Bayesian epistemology, falsificationism about validation does not allow for a positive evaluation of results from a computer simulation. These results, the simulation programs and models behind them may be rejected. But if results from a simulation are not falsified during testing, neither the results nor the simulation as such is confirmed in any way. Otherwise, falsificationism would have to allow for a type of induction, which it does not do.[33]

Second, and relatedly, falsificationists cannot allow that validation takes into account the prior credibility of the model assumptions used. The reason is simply that no positive notion of credibility is available to falsificationists. By contrast, Bayesians can draw on the prior credibility of the model assumptions, as was shown in Sect. 7.3.2.

We may conclude then that a Bayesian approach to validation has many virtues and significant advantages, if compared to a falsificationist outlook.

[31] Maybe, imprecise probabilities allow for a yet different rival to Bayesianism. A discussion of this issue is beyond the scope of this chapter.

[32] See Chap. 6 by Beven and Lane in this volume for a falsificationist view of validation. Note that this chapter moves quite far away from falsificationism. We will here concentrate on a very simple version of falsificationism.

[33] Falsificationists cannot solve the problem by introducing the notion of corroboration (see Popper 1934, Sect. 4, Chap. X and App. *IX for this notion and Putnam 1974, Salmon 1981 for a critical perspective) or that of verisimilitude (see Popper 1969, pp. 385–99 and Keuth 2000, Chap. 7 for discussion).

7.5 Conclusions

Bayesian epistemology is a powerful theory of scientific inference. As we have seen in this chapter, it can be usefully applied to validation too. Bayesianism is particularly helpful because it can tell how trust should be built up in simulations and their results: Simulations and their results can have prior probability because the model assumptions on which they are built have credibility. They can be confirmed by updating corresponding beliefs using measured data. As we have seen, some general problems about Bayesian epistemology are pertinent to Bayesian validation too, but there are proposals for solutions. Bayesian validation raises some interesting questions of its own, but I have argued that they do not lead to legitimate worries that Bayesianism can be useful in validation.

The topic of Bayesian validation raises a number of research questions that need further scrutiny. For instance, the idea of using Bayesian epistemology for verification, which is closely related to validation, needs elaboration. From a philosophical point of view, it would be interesting to see how a Bayesian approach to validation can be connected to solutions of the old evidence problem (see Chap. 41 by Frisch in this volume). The most urgent need for future research seems the application of Bayesian methods in validation. Only more detailed examples can teach us how well Bayesian validation fares in practice.

Acknowledgements This chapter has benefited from helpful comments by Seamus Bradley, interesting critical suggestions and comments by Nicole J. Saam, and an extensive and extremely valuable commentary by Stephan Poppe. Thanks to all of them!

References

Albert, M. (1999). Bayesian learning when chaos looms large. *Economics Letters*, *65*, 1–7.

Bartelborth, T. (2013). Sollten wir klassische Überzeugungssysteme durch Bayesianische ersetzen? *Logos*, *3*, 2–68.

Bayarri, M. J., Berger, J. O., Higdon, D., Kennedy, M. C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C. H., & Tu, J. (2002). A framework for validation of computer models. Tech. rep. Obtained from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.125.7166&rank=5.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., James Cavendish, C.-H. L., et al. (2005). *A framework for validation of computer models*. NISS, Technical Report Number 162.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., et al. (2007). A framework for validation of computer models. *Technometrics*, *49*(2), 138–154.

Beisbart, C. (2011). A rational approach to risk? Bayesian decision theory. In S. Roeser, R. Hillerbrand, P. Sandin, & M. Peterson (Eds.), *Handbook of risk theory* (pp. 375–404). Berlin: Springer.

Beisbart, C. (2012). How can computer simulations produce new knowledge? *European Journal for Philosophy of Science*, *2*, 395–434.

Blackwell, D., & Dubins, L. (1962). Merging of opinions with increasing information. *Annals of Statistical Mathematics*, *33*, 882–886.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review*, *97*, 303–352.

Bovens, L., & Hartmann, S. (2004). *Bayesian epistemology*. Oxford: Oxford University Press.

Bradley, S. (2016). Imprecise probabilities. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 ed.

Chen, W., Xiong, Y., Tsui, K.-L., & Wang, S. (2006). Some metric and Bayesian procedure for validating predictive models in engineering design. In *Proceedings of IDETC, CIE 2006 ASME 2006 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference September 10–13, 2006*. Philadelphia: Pennsylvania, USA.

Chen, W., Xiong, Y., Tsui, K.-L., & Wang, S. (2008). A design-driven validation approach using Bayesian prediction models. *Journal of Mechanical Design*, *130*(2), 021101.

de Finetti, B. (1931a). Probabilismo. *Logos*, *14*, 163–219. Translated as de Finetti, B. (1989). Probabilism. A critical essay on the theory of probability and on the value of science. *Erkenntnis*, *31*, 169–223.

de Finetti, B. (1931b). Sul significato soggetivo della probabilità. *Fundamenta Mathematica*, *17*, 298–329.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, *7*, 1–68. Here quoted from the English translation: de Finetti, B. (1964). Foresight: Its logical laws, its subjective sources. In H. E. Kyburg, & H. E. Smokler (Eds.), *Studies in subjective probability* (pp. 53–118). Wiley.

Döring, F. (2000). Conditional probability and Dutch books. *Philosophy of Science*, *67*(3), 391–409. http://www.jstor.org/stable/188624

Earman, J. (1992). *Bayes or bust*. Cambridge (MA): MIT Press.

Easwaran, K. (2011a). Bayesianism I: Introduction and arguments in favor. *Philosophy Compass*, *6*(5), 312–320.

Easwaran, K. (2011b). Bayesianism II: Applications and criticisms. *Philosophy Compass*, *6*(5), 321–332.

Frigg, R., Bradley, S., Du, H., & Smith, L. A. (2014). Laplace's demon and the adventures of his apprentices. *Philosophy of Science*, *81*(1), 31–59.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123.

Gilbert, M. (1987). Modeling collective belief. *Synthese*, *73*, 185–204.

Gillies, D. (2000). *Philosophical theories of probability*. London and New York: Routledge.

Glymour, C. N. (1980). *Theory and evidence*. Princeton: Princeton University Press.

Hájek, A. (2012). Interpretations of probability. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. (Winter 2012 ed).

Hájek, A., & Hartmann, S. (2010a). Bayesian epistemology. In J. D. et al. (Ed.) *A companion to epistemology*, (pp. 93–106). Oxford: Blackwell.

Hájek, A., & Hartmann, S. (2010b). Bayesian epistemology. *A companion to epistemology* (pp. 93–105). Oxford: Blackwell.

Hartmann, S., & Sprenger, J. (2010). Bayesian epistemology. In D. Pritchard & S. Bernecker (Eds.), *Routledge companion to epistemology* (pp. 609–620). London: Routledge.

Hempel, C. G. (1945). Studies in the logic of confirmation (I.). *Mind*, *54*, 1–26. reprinted in L. Sklar, (2000). *Philosophy of science. Probability and confirmation* (pp. 245–270). Garland, New York.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). Open Court: La Salle.

Huber, F. (2007). Confirmation and induction. In J. Fieser, & B. Dowden (Eds.), *Internet encyclopedia of philosophy*.

Huber, F. (2009). Belief and degrees of belief. In F. Huber, & C. Schmidt-Petri (Eds.), *Degrees of belief*. Springer.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *4/3*, 227–241. http://bayes.wustl.edu/etj/articles/prior.pdf.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*, 620–630.

Jaynes, E. T. (1979). Where do we stand on maximum entropy? In R. D. Levine & M. Tribus (Eds.), *The maximum entropy formalism* (pp. 15–118). Cambridge (MA): M. I. T. Press.

Jeffrey, R. (1967). *The logic of decision* (2nd ed.). New York: McGraw-Hill.

Jensen, K. K. (2011). A philosophical assessment of decision theory. In S. Roeser, R. Hillerbrand, P. Sandin, & M. Peterson (Eds.), *Handbook of risk theory* (pp. 405–439). Berlin: Springer.

Joyce, J. (2016). Bayes' theorem. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University (winter 2016 ed.).

Kennedy, M. C., & O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, *87*(1), 1–13.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(3), 425–464.

Keuth, H. (2000). *Die Philosophie Karl Poppers*. Tübingen: UTB, Mohr und Siebeck.

Keynes, J. M. (1921). *A treatise on probability*. London: Macmillan. (Reprint 1973).

Knutti, R. (2008). Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1885), 4647–4664. https://doi.org/10.1098/rsta.2008.0169.

Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (Second English edition). Chelsea.

Lee, P. M. (2012). *Bayesian statistics. An introduction* (4th ed.). Chichester: Wiley.

Lewis, D. (1980). A subjectivist's guide to objective chance. In *Studies in inductive logic and probability* (vol. II, pp. 263–293). Berkeley and Los Angeles: University of California Press. (Here quoted from Lewis (1986) pp. 83–113).

Lewis, D. (1986). *Philosophical papers* (Vol. II). New York: Oxford University Press.

Lewis, D. (1994). Humean supervenience debugged. *Mind*, *103*, 473–490. Reprinted in Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.

Lewis, D. (1997). Why conditionalize? In D. Lewis (Ed.), *Papers in metaphysics and epistemology* (pp. 403–407). Cambridge: Cambridge University Press.

Mayo, D. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mellor, H. D. (2005). *Probability. A philosophical introduction*. London and New York: Routledge.

Myrvold, W. C. (2003). A Bayesian account of the virtue of unification. *Philosophy of science*, *70*(2), 399–423.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. Upper Saddle River, NJ: Pearson Prentice Hall.

Oakley, J. E., & O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(3), 751–769.

Oberkampf, W. L., & Barone, M. F. (2006). Measures of agreement between computation and experiment: Validation metrics. *Journal of Computational Physics*, *217*, 5–36. http://portal.acm.org/citation.cfm?id=1167051.1167053

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge: Cambridge University Press.

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering & System Safety*, *91*(10), 1290–1300.

Papoulis, A., & Pillai, S. U. (2002). *Probability, random variables, and stochastic processes* (4th ed.). Boston etc.: McGraw-Hill.

Parker, W. S. (2009). Confirmation and adequacy-for-purpose in climate modelling. *Aristotelian Society Supplementary*, *83*(1), 233–249.

Plato. (2015). *Theatetus and sophist*. Cambridge: Cambridge University Press. (Edited by C. Rowe).

Popper, K. R. (1934). *Die Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft*. Wien: Julius Springer. (English version: The logic of scientific discovery, London: Hutchinson, 1959).

Popper, K. R. (1969). *Conjectures and refutations* (3rd ed.). London: Routlege and Kegan Paul.

Putnam, H. (1974). The 'corroboration' of theories. In P. A. Schilpp (Ed.), *The philosophy of karl popper* (pp. 221–240). La Salle (IL): Open Court.

Ramsey, F. (1926). Truth and probability. First printed in: Braithwaite, R. B. (Ed.), *Foundations of mathematics and other essays* (pp. 156–198). Routledge and P. Kegan, London 1931. Reprinted in Ramsey, F. (1990). Philosophical papers. In D. H. Mellor (Ed.). (pp. 52–94). Cambridge: University Press.

Rawls, J. (1971). *A theory of justice*. Harvard University Press: Cambridge (MA). (Quoted from the revised edition 1999).

Roy, C. J., & Oberkampf, W. L. (2010). A complete framework for verification, validation, and uncertainty quantification in scientific computing. In *AIAA 2010-124, 48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*. Published by American Institute of Aeronautics and Astronautics, Inc.

Salmon, W. (1981). Rational prediction. *British Journal for the Philosophy of Science*, *32*, 115–125.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York: Dover. (1st ed.) Wiley: New York (1954).

Sprenger, J. (2015). A novel solution to the problem of old evidence. *Philosophy of Science*, *82*(3), 383–401.

Strevens, M. (2006). The Bayesian approach to the philosophy of science. In D. M. Borchert (Ed.), *Encyclopedia of philosophy* (2nd ed., pp. 495–502). Macmillan Reference.

Talbott, W. (2016). Bayesian epistemology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University (winter 2016 ed.)

Teller, P. (1973). Conditionalization and observation. *Synthese*, *26*, 218–258.

von Plato, J. (1989). De Finetti's earliest works on the foundations of probability. *Erkenntnis*, *31*, 263–282.

von Plato, J. (1994). *Creating modern probability*. Cambridge: Cambridge University Press.

Wagner, C. G. (1997). Old evidence and new explanation. *Philosophy of Science*, *64*(4), 677–691.

Wang, S., Chen, W., & Tsui, K.-L. (2009). Bayesian validation of computer models. *Technometrics*, *51*(4), 439–451.

Wenmackers, S., & Romeijn, J. (2016). New theory about old evidence. *Synthese*, *193*(4).

Chapter 8
Validation of Computer Simulations from a Kuhnian Perspective

Eckhart Arnold

Abstract While Thomas Kuhn's theory of scientific revolutions does not specifically deal with validation, the validation of simulations can be related in various ways to Kuhn's theory: (1) Computer simulations are sometimes depicted as located between experiments and theoretical reasoning, thus potentially blurring the line between theory and empirical research. Does this require a new kind of research logic that is different from the classical paradigm which clearly distinguishes between theory and empirical observation? I argue that this is not the case. (2) Another typical feature of computer simulations is their being "motley" (Winsberg in Philos Sci 70:105–125, 2003) with respect to the various premises that enter into simulations. A possible consequence is that in case of failure it can become difficult to tell which of the premises is to blame. Could this issue be understood as fostering Kuhn's mild relativism with respect to theory choice? I argue that there is no need to worry about relativism with respect to computer simulations, in particular. (3) The field of social simulations, in particular, still lacks a common understanding concerning the requirements of empirical validation of simulations. Does this mean that social simulations are still in a prescientific state in the sense of Kuhn? My conclusion is that despite ongoing efforts to promote quality standards in this field, lack of proper validation is still a problem of many published simulation studies and that, at least large parts of social simulations must be considered as prescientific.

Keywords Computer simulations · Validation of simulations · Scientific paradigms

E. Arnold (✉)
Bavarian Academy of Sciences and Humanities, Munich, Germany
e-mail: eckhart.arnold@posteo.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_8

8.1 Introduction

Kuhn (1976) famously introduced the term *paradigm* to characterize the set of background beliefs and attitudes shared by all scientists of a particular discipline. According to Kuhn these beliefs and attitudes are mostly centered around *exemplars* of good scientific practice as presented in the textbook literature, but classical texts, specific methodological convictions or even ontological commitments can also become important for defining a paradigm. Furthermore, paradigms comprise shared convictions as well as unspoken assumptions of the group of researchers (Kuhn 1976, postscript). An important function of paradigms is that they both define and limit what counts as relevant question and legitimate problem within a scientific discipline.

Kuhn's concept of a paradigm is closely connected with his view of how science develops. According to Kuhn phases of *normal science* where science progresses within the confinements of a ruling paradigm are followed by *scientific revolutions* which, in a process of creative destruction, lead to a paradigm shift. Scientific revolutions are triggered by the accumulation of problems that are unsolvable within the ruling paradigm (so called *anomalies*). With an increasing number of anomalies scientists grow unsatisfied with the current paradigm and start to look for alternatives—a state of affairs that (Kuhn 1976, Chap. 7/8) describes as the *crisis* of the ruling paradigm. Then, a paradigm shift can occur that consists in a thoroughgoing conceptual reorganization of a scientific discipline or, as the case may be, the genesis of a new sub-discipline. Unless there is a crisis, the search for alternative paradigms is usually suppressed by the scientific community.

This theory could be relevant for computer simulations and their validation. Because computer simulations are sometimes characterized as a revolutionary new tool that blurs the distinction between model and experiment, the question can be asked if this tool brings about or requires new paradigms of validation. Under *validation* I understand a process which allows to test whether the results of a scientific procedure adequately capture that part of reality which they are meant to explain or to enable us to understand. It is widely accepted that for theories or theoretical models, the process of validation consists in the empirical testing of their consequences by experiment or observation, which in this context is also often described as *verification* or *falsification* or, more generally, as *confirmation*.[1] The question then is, if the same still holds for computer simulations, that is, if computer simulations also require some form of empirical validation before they can be assumed to inform us about reality.

For the purpose of this paper, I understand empirical validation in a somewhat wider sense that does not require strict falsification, but merely any form of matching theoretical assumptions with empirical findings. In this sense, a historian checking

[1] In the realm of computer simulations the term *verification* is, somewhat confusingly, reserved for checking wether the simulation software is free from programming errors (so-called "bugs") and whether it is faithful to the mathematical model or theory on which it is based. The term *validation* is used for the empirical testing of the simulation's results. See also Chap. 4 by Murray-Smith in this volume.

an interpretation against the historical sources can also be said to validate that interpretation. However, I assume that proper validation always includes an empirical component and I therefore use the terms "validation" and "empirical validation" interchangeably in the following.

In the following, I first summarize Kuhn's philosophy of science (Sect. 8.2). Then I list some of the dramatic changes that computer simulations have brought about in science and—in order to forestall possible misunderstandings—explain why these changes are not scientific revolutions in the sense of Kuhn (Sect. 8.3). In the main part of this chapter (Sect. 8.4), I then examine the validation of simulations from a Kuhnian perspective. Relating to the discussion about the relation between computer simulations and experiments I argue that computer simulations can clearly be distinguished from real experiments and, therefore, do not require a new paradigm of validation. In principle, validating simulations is just like validating theory. I continue by examining whether computer simulations aggravate the problem of theory choice that is associated with the so called "Duhem-Quine-thesis" (Harding 1976), which I deny. Finally, I examine some of the issues that the validation of social simulations and in particular agent-based-models raises from the point of view of Kuhn's philosophy of science. For the lack of commonly accepted standards of validation, it seems unclear whether this field has already reached a state of "normal science" with established paradigms of validation. Because the practices of validation vary greatly in this field, a general conclusion is not possible, however. I therefore confine myself to discussing the issue with respect to selected examples.

8.2 Kuhn's Philosophy of Science

A crucial aspect of Kuhn's concept of scientific revolutions is the alleged *incommensurability* of paradigms (Kuhn 1976, Chap. 12, postscript 5.) (Sismondo 2010, Chap. 2) (Bird 2013, Sect. 4.3f.). Incommensurability means that theories rooted in different paradigms cannot easily be compared with respect to their scientific merits, because of

1. *methodological incommensurability*, which means that the criteria of evaluation depend on and change with the paradigm,
2. the *theory-ladenness of observation*, due to which an assessment based on empirical evidence may not be able to resolve the dispute,
3. *semantic incommensurability*, which means that the differences of the respective conceptual reference frameworks and taxonomies may render the translation between the nomenclatures of different paradigms difficult and error-prone.

Kuhn did not go as far as the proponents of the strong program of sociology of science who maintain that the resolution of inter-paradigm-disputes is primarily, if not exclusively, determined by social factors such as group allegiance and power-structures (Bird 2013, Sect. 6.3). However, he did deny that the choice between different theories is guided by a scientific meta-method such as systematic falsification or by any other particular set of rules. In this respect one can describe Kuhn's stance as a *mild relativism*. Kuhn's relativism is restricted by his belief that a common ground for theory choice can still be found in such general characteristics as empirical accuracy, consistency, breadth of scope, simplicity or parsimony, fruitfulness for future research (Kuhn 1977, Chap. 13). And he furthermore holds that the comparison and mutual evaluation of paradigms is possible on the pragmatic basis of their problem-solving capacity.

Although Kuhn regarded scientific revolutions and the paradigm shifts they bring about as scientifically perfectly legitimate processes, that is processes that are primarily driven by a scientific motivation and not just by social power, he nonetheless found that in almost any paradigm change some things get lost—if only that certain questions will not be considered worthwhile any more. An example is the question how physical bodies influence each other over a distance, which cannot be answered by Newton's theory of Gravity and therefore simply was not asked any more, although, before Newton it was considered important (Kuhn 1976, Chap. 12). The phenomenon that accepted questions, problems and even solutions can become orphaned after a paradigm shift has subsequently been called *Kuhn loss* (Bird 2013, Sect. 2).

Also, even though Kuhn allowed for paradigm shifts to make sense scientifically, this does not always need to be the case, but one should expect that sometimes paradigm shifts are primarily due to social factors. Not in the least because of the popularity of Kuhn's theory of scientific revolutions, it has become seductive for scientists to stage a paradigm shift to promote their scientific agenda. In order to distinguish illegitimate paradigm-shifts terminologically, the derogatory term *scientific imperialism* can be used, which has been coined to describe the take-over of a branch of science by a single paradigm (Dupre 1994) by unfair means. Following Kuhn's line of thought the problem solving capacity could be a criterion by which to qualify a paradigm shift as either legitimate or imperialistic. Because of the incommensurability issues described before, an objective judgment about this can, of course, be difficult.

A contemporary of Kuhn that is often mentioned in the same breath, is Paul Feyerabend, who is (in-)famous for the slogan "anything goes". In popular folklore this is sometimes understood as meaning that Feyerabend advocated that in science any method is as good as any other. However, what Feyerabend actually demonstrated in his book "Against Method: Outline of an Anarchist Theory of Knowledge" (Feyerabend 1975/1983) and other works was that even from the most humble historical beginnings, a serious scientific theory or school of thought can still emerge. Feyerabend's work gains its thrust from the fact that he can show that some of the game changers in the history of science such as, for example, Galileo's theory of motion, violated accepted scientific standards of their time (Feyerabend 1975/1983, Chap. 9). Just as Kuhn he denies that the historical development of science is or can be guided by methodological or epistemological rules. Similar to Kuhn, Feyerabend's philosophy has a certain relativistic flair, which Feyerabend other than Kuhn was ready to accept (Preston 2016, Sect. 5).

Nonetheless, despite what the subtitle of his major work suggests, Feyerabend's analyses do not warrant a strong relativism. Almost all of Feyerabend's examples concern theories that—later in their historical development—would be considered as scientific even by conventional standards. Thus, what we can learn from Feyerabend is a certain tolerance against the methodological chaos of new scientific approaches in their infant stages. This can be important, for example, when evaluating social simulations, which according to some authors suffer from a lack of proper empirical validation (Heath et al. 2009). The question is then not so much whether these simulations adhere to a particular scientific standard but rather whether the respective scientific community learns from its failure to do so and will be able to develop appropriate methodological standards in the future.

Another point that deserves clarification, because it is—at least in the philosophical discussion—almost habitually mentioned in context with Kuhn, is the *Duhem–Quine thesis* (Harding 1976). The Duhem–Quine thesis draws on the fact that if the logical consequence of a whole system of premises turns out to be false then it is still unclear which one or more of the premises are false.[2] This means that if a theory is empirically disconfirmed, we do not (yet) know which part of the theory is wrong. The Duhem–Quine thesis can be seen as supporting a certain degree of arbitrariness, if not relativism in theory choice. And it corresponds well to Kuhn's view that the way scientists cope with anomalies is not strictly guided by methodological rules. It may be a matter of creative choice. As we shall see later, this choice is in practice much less arbitrary than it may appear in the formal logical representation of a theory as a system of propositions.

Despite all reservations, Kuhn's picture of the history of science is still one of linear development, where normal science and revolutionary phases follow each other in time. For Kuhn the prolonged coexistence of several competing paradigms was the mark of a prescientific stage where much intellectual energy is wasted in disputes between rivaling schools of thought. Recent research, however, has emphasized that the coexistence of different paradigms within one and the same science is much too common to be dismissed as prescientific (Kornmesser 2014; Schurz 2014). This is particularly true of the social sciences, where hardly ever one paradigm can claim to solve all puzzles so successfully that it is able to gather the entire scientific community under its flag. That Kuhn may have underestimated the amount of coexistence of paradigms in science does not invalidate his analyses, though. The concepts of *normal science* and *scientific revolutions* can still be employed as ideal-types to characterize the scientific proceedings within an established paradigm on the one hand and the discourse between different coexisting paradigms on the other hand.

[2] See also Chap. 39 by Lenhard in this volume.

8.3 A Revolution, but not a Kuhnian Revolution: Computer Simulations in Science

Kuhn's theory of scientific revolutions is so popular that his concept of a paradigm has by now become part of the common vocabulary. Inevitably, it is often used in a sense that is different from what Kuhn had in mind. It may therefore help to make clear what is not a revolution or paradigm change in Kuhn's sense. A most salient example in this context is that of the introduction of computer simulations to science, because it can with some justification be said that computer simulations have revolutionized many areas of science.

Computer simulations can roughly be defined as the imitation of a natural process (or, in the case of social simulations, a social process) by a computer program (Hartmann 1996). Undoubtedly, computer simulations have brought about considerable changes in scientific practice and theoretical outlook. Here are but some examples:

- In engineering, simulations have been used before long to simulate the properties of machinery and processes. A large class of simulations is based on the method of finite elements which has as far reaching applications as structural engineering, car crash tests and even cardiovascular simulations (Carusi et al. 2013).
- In chemistry simulations are employed in order to simulate chemical processes on a quantum-mechanical bases, some of which are even outside the reach of direct experimentation (Arnold 2013).
- In climate science the simulations are used to simulate the possible future development of the world climate. Naturally, experimentation with the world climate is not possible. By the same token, unfortunately, these simulations cannot be validated directly.
- The theory of nonlinear dynamical systems ("chaos theory") can even be said to owe much of its origin to computational methods (Gleick 2011). At any rate its development has certainly been propelled by the use of computers, though it might not necessarily have been computer simulations in the narrower sense of imitations of a natural process in the computer.
- In social science there exists a now already long-standing tradition of simulating social processes. However, the social simulations community still struggles for the acceptance within the broader social sciences community (Squazzoni and Casnici 2013).

Some of these examples certainly warrant the characterization as "revolutionary". Are they revolutionary in a Kuhnian sense, though? And would it be reasonable to call simulation-based science in general a new paradigm of science?

For one thing, the way Kuhn used the term paradigm, paradigms are always tied to specific scientific disciplines. Even though we are not tied to Kuhn's definition and the term *paradigm* has indeed been used more liberally by other authors since its original introduction, it would appear a bit vague to speak of a paradigm of computer simulations, because it is not at all clear what would be the content of this paradigm.

Even more importantly, Kuhn reserves the concept of scientific revolutions for changes that are caused by a crisis of the conceptual framework of a scientific discipline and that lead to a reconstruction of the conceptual system that is incommensurable with the previous reference framework. Not any dramatic change in science is a revolution in the Kuhnian sense. A prominent example for a dramatic change that is not a Kuhnian revolution is the discovery of the structure of the DNA-molecule by Watson and Crick. While this discovery was a door-opener for molecular genetics, it neither required nor effected a conceptual reconstruction and there was no question of it being incommensurable with the previously held views on hereditary biology. Quite the contrary, it fit in nicely with the existing body of knowledge. The discovery of the DNA was normal science at its best, not a Kuhnian revolution.

Similarly, the introduction of computer simulations into a particular branch of science alone is not a Kuhnian revolution, no matter how dramatic the changes in scientific practice and the extension of our knowledge through computer simulations might be. Only, if the use of computer simulations leads to a revision of established fundamental concepts, it is a Kuhnian revolution. A possible candidate from the list above might be chaos theory, in so far as it has modified the received picture of causality.

8.4 Validation of Simulations from a Kuhnian Perspective

Can Kuhn's concept of paradigm illuminate the validation of computer simulations? And, if so, how? In the following, I am going to state several questions that can be raised in this context and then try to give answers to these questions based on the current discussion on computer simulations in the philosophy of science. The questions that in my opinion deserve consideration are

1. Notwithstanding the question (discussed earlier) to what extent computer simulations have prompted paradigm shifts *in* science, another question is, whether computer simulations have lead to, or require new paradigms in the logic of scientific discovery. Classical research logic assumes a clear distinction between theoretical research based on deductive inference and empirical research based on experiment and (potentially theory-laden) observation.[3] Most importantly, there is a hierarchy between the theoretical and empirical realm. Theoretical assumptions

[3] Because theory-ladenness of observation is an often misunderstood topic, two remarks are in order: (1) Theory-ladenness of observation as such does not blur the distinction between theory and observation. At worst we have a distinction between pure theory (without any observational component) and theory-laden observation. (2) Theory-ladeness of observation does not lead to a vicious circle when confirming theories by empirical observation. This is true, as long as the observations are not laden with the particular theories for the confirmation of which they are used. There are areas in science where no sharp distinction between theoretical reasoning and reporting of observations is made. However, as far as computer simulations are concerned, it is clear that because Turing Machines do not make observations, a computer program is always a theoretical entity—not withstanding the fact that a computer program may represent an empirical setting or

are confirmed or disconfirmed by empirical tests—not the other way round. Computer simulations are sometimes depicted as being located somewhere between empirical and theoretical research, and—as the common metaphor of "computer experiments" suggests[4]—blurring the lines between the two (Morrison 2009).

2. In a similar vein, computer simulations often rely on a rich mixture of assumptions and technicalities that are drawn from diverse sources. In the philosophical literature on simulations this has been described as their being "motley"(Winsberg 2015) and not simply falling from theory. This can raise worries concerning the prospects of empirical validation of computer simulations. In particular, the question can be asked if the sort of problems associated with the Duhem–Quine thesis increase with computer simulations: You may know that your simulation contains many abstractions, simplifications, and presumptions, but you cannot be sure which of these are potentially dangerous.
3. Finally, some thoughts shall be given to the validation of simulations in the social sciences. Because the social sciences are multi-paradigm-sciences the validation of simulations raises specific problems in this area. Given that it is still not common practice to validate simulations, one can even ask whether the field of social simulations has already emerged from a prescientific state.

8.4.1 Do Computer Simulations Require a New Paradigm of Validation?

While Kuhn's theory of scientific revolutions is mainly concerned with the supersession of scientific theories, his concept of paradigms can also be applied to other aspects of scientific practice. For example, it might be applied to changes in the logic of scientific research. The question whether computer simulations bring about (or require) a new kind of research logic is particularly salient, because it has been argued recently that computer simulations somehow blur the line between models and experiments (Winsberg 2009). But if this means that computer simulations are—just like experiments—somehow empirical, the question naturally arises whether the validation of computer simulations can still be understood along the lines of what has earlier been described as classical research logic. Or, if a new paradigm of validation is necessary to assess whether a simulation adequately captures its target system or not?

Before the recent discussion about the relation of simulations and experiments, this question seemed to be rather trivial and its answer obvious: Computers are calculating machines and computer simulations are nothing but programed mathematical models that run on the computer. Therefore, computer simulations can just like models produce no other than purely inferential knowledge, that is, knowledge that fol-

make use of empirical data. In the latter respect it can be compared with a physical theory that may in fact represent empirical reality as well as contain natural constants (i.e., empirical data).

[4] See also Chap. 37 by Beisbart in this volume.

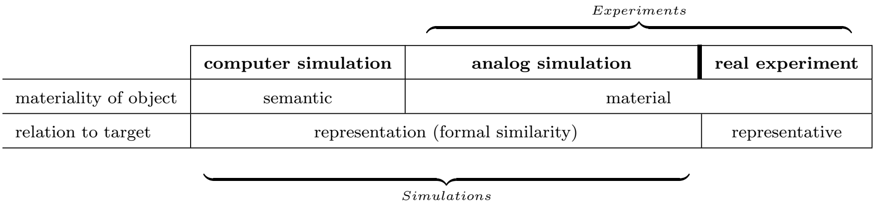
lows deductively from the premises built into the simulation. In particular, computer simulations cannot produce genuine empirical knowledge like experiments or observations can. It is true that computer simulations can produce new knowledge, because they yield logical consequences of the built-in premises that were not formerly known to us (Imbert 2017, Sect. 1.3.4). It is also true that computer simulations can—like any model—produce knowledge about empirical reality, because the premises built into them have empirical content and so have their logical consequences. But this is far cry from the empirical knowledge that experiments or observations yield and which—because it is of empirical origin—is genuine. But then computer simulations have just the same epistemic status as theories and models and therefore follow the same research logic and require just the same kind of validation. Now, in order to validate a model or a theory it must be tested empirically, and so must computer simulations.

What I have just described is more or less the picture of computer simulations that was pertaining in the general literature on simulations up to the beginning of the millennium. It had by that time been fleshed out with two distinctions that make the difference between computer simulations and empirical research procedures extraordinarily clear: First, by the distinction of the modus operandi. Is it a *formal* procedure (computer simulation) or a *material* process (experiment)? Second, by the distinction of their relation to the target system. Accordingly, this relation could be characterized as one of *formal similarity* (Guala 2002) with the object of the simulation being a *representation* (Morgan 2003) of the target system or, in the case of experiments, one of *material similarity* with the object of experimentation being a *representative* of the target system.

In recent years, however, there has been a persistent discussion among philosophers of science during the course of which the distinction between simulations and experiments has been seriously called into question. Most notably, some authors have claimed that it is impossible to make a sharp distinction between simulations and experiments—at least as far their epistemic reach or inferential power is concerned. (Winsberg 2009; Parker 2009; Morrison 2009; Winsberg 2015). Others have advocated the weaker claim that while there is a distinction between the two categories, the transition between them is smooth and that there are borderline cases for which it is difficult to determine into which category they fall (Morgan 2003).

Now, if this were true, then the generally accepted research logic of empirical science, which relies on the ability to distinguish clearly between empirical observation and theoretical reasoning would find itself in a serious crisis and we would have to expect and, in fact, need to hope for new paradigms of research logic and, in particular, for the validation of computer simulations to emerge.

However, the case for the non-discriminability of simulations and experiments rests almost entirely on conceptual confusions and an ambiguous use of the term "experiment". The examples with which supporters of the non-discriminability thesis demonstrate their claim concern almost exclusively atypical kinds of experiments, where the object of experimentation is not really a representative of the target system. For example, (Winsberg 2009, p. 590), discusses "tanks of fluid to learn about astrophysical gas-jets" as an instance of an experiment. But this is an atypical experiment,



	computer simulation	**analog simulation**	**real experiment**
materiality of object	semantic	material	
relation to target	representation (formal similarity)		representative

Fig. 8.1 Conceptual relation of simulations and experiments (Kästner and Arnold 2013)

because the tanks of fluid are not representatives of the target system (astrophysical gas-jets). This kind of experiment is indeed in no better position to produce genuine empirical knowledge about the target system than any computer model. But the fact that there are such atypical experiments does not contradict the fact that there exist real experiments that can produce genuine empirical knowledge about their target system and that this is a feature that distinguishes real experiments from models.

The conceptual confusion that exists in the philosophical discussion about the relation of simulations and experiments can easily be clarified by the schema on Fig. 8.1, which depicts the overlap in the use of the words "simulation" and "experiment". The kind of experiments that Winsberg and other authors advocating the non-discriminability between simulations and experiments discuss over and over again, has been termed "analog simulation" in the schema. As all experiments do, "analog simulations" operate on a material object, but this object does not have a material similarity to its target system and therefore is only a representation, but not a representative of its target system. The latter is required for an experiment to produce genuine empirical knowledge about its target system.

That simulations are not experiments—save for the ambiguity and overlap in the use of words—becomes furthermore clear if we consider the kind of experiments that give rise to anomalies and which in retrospect are declared crucial experiments that decide the choice between conflicting theories. Because the laws of the scientific theories are programmed into computer simulations, they cannot be used to test these very theories. If it really was as difficult to distinguish between simulations and experiments as some philosophers of science believe, then it should—at least in principle—be possible to substitute experiments with simulations in any context.

However, if we draw the demarcation line between analog simulations and real experiments and not, as the authors advocating the non-discriminability-thesis implicitly do, between computer simulations and analog simulations, then we are able to distinguish clearly those scientific procedures that can generate genuine empirical knowledge about their target system from those that cannot. Simulations and, in particular, computer simulations belong to the latter category and therefore have—with respect to validation—the same epistemic status as theories and models. They need to be validated empirically, but they cannot provide empirical validation.[5]

Summing it up, computer simulations do not break the received paradigm of research logic of empirical science. Therefore, a new paradigm of validation specifically for simulations is not needed.

8.4.2 Validation of Simulations and the Duhem–Quine Thesis

Another point frequently emphasized in the philosophy of simulation literature is that computer simulations can become highly complex. This is also one of the major differences between computer simulations and thought experiments, to which they are otherwise quite similar. At least in the natural sciences computer simulations can often be based on comprehensive and well tested theories, such as quantum mechanics, general relativity, Newton's of gravitation or—in engineering—the method of finite elements. But even in the natural sciences simulations cannot always be based on a single theory, but they sometimes rely on different theories from different origins. Climate simulations are a well-known example for this. And even where simulations are based on a single theory, they usually also draw on various sorts of approximations, local models and computational techniques. None of these can be derived from theory, so that they need independent credentials. This situation has been described in the philosophy of simulation literature as their being motley and partly autonomous (Winsberg 2003). This description echos a recent trend in the philosophy of science which emphasizes the importance and relative independence of models from theory (Morgan and Morrison 1999; Cartwright and Press 1983).

So, if simulations are knit together from many independent set pieces of theories, models, approximations, algorithmic optimizations etc., then the Duhem–Quine thesis could point out a potential problem. A possible reading of the thesis assumes that if validation fails (for example, because an empirical prediction was made that turned out to be wrong), then one cannot know which part of the chain of theoretical reasoning failed that leads to the empirical prediction. In the case of computer simulations this means that one does not know whether the theory on which the simulation is based, the simplifications that may have been made in the course of modeling or, finally, the program code has failed.

By the same token, if this reading of Duhem–Quine is accurate, simulation scientists would—for better or worse—enjoy a great freedom of choice concerning where to make adjustments if a simulation fails, i.e. if it leads to unexpected, obviously false or no results at all. Some philosophers have even argued that scientists sometimes deliberately employ assumptions that are known to be false to make their

[5] In simulation-science the term *empirical* is sometimes used to distinguish simulation and numerical methods from mathematical analysis (Phelps 2016 is an example of this.). But this is just a different use of words and should not be confused with "empirical" in the sense of being observation-based as the word is understood in the context of empirical science.

simulations work. Among these are artificial viscosity (Winsberg 2015, Sect. 8), or—another often cited example—"Arakawa's trick" (Lenhard 2007). Arakawa based a general circulation model of the world climate on physically false assumptions to make it work, which by the scientific community was accepted as a technical trick of trade.

However, this reading of Duhem–Quine paints a somewhat unrealistic picture of scientific practice, because in case of failure there usually exist further contextual cues where the error causing the failure has most likely occurred. While in the abstract formal representation of theories that is sometimes used to explain Duhem–Quine, the premises are represented as propositions with no further information, scientists usually have good reasons to consider the failure of some premises as more likely than others. In science and engineering, the premises are usually ordered in a hierarchy that starts with the fundamental physical, chemical, or biological theories, ranges over various steps of system description and approximation down to the computer algorithms and, ultimately, the programm code. If a simulation fails one would start to examine the premises in backward order. And this is only reasonable, because prima facie, it is more likely that your own program code contains a bug than, say, that the theory of quantum mechanics is false or that some of the tried and tested approximation-techniques are wrong. Though, of course, this is not completely out of the question, too.[6] It should be understood that the credibility of the various premises occurring in this hierarchy does not follow their generality, but depends on their respective track record of successful applications in the past. It can safely be assumed that this situation is typical for normal science.[7]

It must be conceded, though, that during a scientific revolution or within cross-paradigm-discourse, there might be no hierarchy of premises to rely on, because some of the premises higher up in the hierarchy, like the fundamental theories, are not generally accepted any more. In this situation, there might, as Kuhn suggested, only be vague meta-principles left to rely on and we must face the possibility of not being able to resolve all conflicts of scientific opinion.

What about the conscious falsifications like artificial viscosity and "Arakawa's trick" that—according to some philosophers of science—are introduced by simulations scientists in order to make their simulations work? This reading has not gone unchallenged, and it has been called in to question whether the artificial viscosity that Winsberg mentions is more than just another harmless approximation (Peschard 2011) or whether "Arakawa's trick" not merely compensates for errors Error made at another place, which would make it an example of a simulation the success of which

[6] See Arnold (2013, Sect. 3.4) for a case-study containing a detailed description of this hierarchy of premises.

[7] But see Lenhard in Chap. 39 in this volume, who paints a very different picture. I cannot resolve the differences here. In part they are due to Lenhard using examples where "due to interactivity, modularity does not break down a complex system into separately manageable pieces." (Lenhard and Winsberg (2010), p. 256) To me it seems that as far as software design goes, it is always possible—and in fact good practice—to design the system in such a way that each unit can be tested separately. As far as validation goes, I admit that this may not work as easily because of restrictions concerning the availability of empirical data.

is badly understood rather than one that is very representative of simulation-based science (Beisbart 2011, 333f.). It seems that these philosophically certainly interesting examples concern exceptions rather than what is the rule in the scientific practice with simulations. For the time being that is to say, because it is well imaginable that in the future development of science these tricks become more common.

Summing it up, with respect to the Duhem–Quine thesis there are neither additional challenges nor additional chances for the validation of simulations. Under *normal science*-conditions it does not play a role at all. Other than that it merely reflects the greater methodological imponderabilities during a revolutionary phase or in an inter-paradigm context.

8.4.3 Validation of Social Simulations

Most of the discussion so far and all of the examples were centered around science and engineering. Therefore, in the following I am going to briefly discuss questions concerning the validation of simulations that are more specific for the social sciences.

8.4.3.1 Where Social Simulations Differ

In the context of validation of social simulations two features of the social sciences become relevant that distinguish them from most natural sciences: First, the social sciences are multi-paradigm-sciences. It is the normal state of these sciences that there exist multiple more or less mutually incommensurable paradigms at the same time. This multi-paradigm-character is well described in the textbook by Moses and Knutsen (2012). For Kuhn such a state of affairs was a sign of a prescientific phase. But given that the social sciences are—within inevitable confinements—nonetheless able to produce convincing explanations at least for some social phenomena, the qualification as prescientific seems inadequate. Also, if considered in isolation, most of these paradigms expose typical features of normal science, like a textbook literature, role models and exemplars, etc.

Deviating from Kuhn, I therefore suggest, that the qualification as prescientific should be reserved to those sciences or branches of a science that—given their state of development—have not yet been able at all to produce results that can be validated or confirmed by some reasonable procedure. The qualification as prescientific is in so far justified as without a common understanding and practice of validation one can never be sure whether the results are indeed reliable.

Secondly, the social sciences include qualitative paradigms, including paradigms that rely on hermeneutical methods. It is safe to assume that these can neither be completely ignored nor always be resolved to quantitative or otherwise formal methods and paradigms.[8] As computer simulations are quantitative, the decision to use computer simulations is also a decision for a quantitative paradigm.

Here, I understand the term "quantitative" in a wide sense, including anything that is described in a formal language. This can be formal logic, mathematics, or a programming language. This wide sense of using the term "quantitative" is motivated by the fact all formal descriptions share the same epistemic risks of either losing important information, because the expressive power of formal languages is limited in comparison to natural language, or adding arbitrary assumptions in form of modeling decisions. A simulation model forces its author to provide detailed mechanics of all processes that are included in the model, because otherwise the model would not run. However, if the mechanics are not known, this amounts to theoretical speculation. A purely verbal description, in contrast, allows its author to remain silent or at least adequately vague about underlying mechanics the details of which are not known. On the other hand, because of their strict specification, formal models cannot as easily be misunderstood as verbal descriptions. And they enforce logical consistency.

Both of these features affect the validation of social simulations. Because, when trying to validate a simulation study, say, on the evolution of cooperation, it might become necessary to compare its findings with those of biological field research or, depending on the envisaged application cases, those of cultural history. Thus, different scientific disciplines with different paradigms might be affected. And, it might become necessary to translate between a qualitative descriptive language used in empirical research and the formal languages used in simulation research.

One possible objection when discussing social simulations in the connection with Kuhn, is that it is not a scientific discipline, but a field that runs across several disciplines. However, since this field is shaped by shared attitudes, well-known exemplars (Axelrod 1984; Axtell et al. 2002; Epstein and Axtell 1996; Schelling 1971) and an emerging textbook-literature (Railsback and Grimm 2012; Gilbert and Troitzsch 2005), looking at it from a Kuhnian perspective does not seem too far-fetched.

8.4.3.2 Are Social Simulations Still in a Prescientific Stage?

One of the most surprising features to the outside observer of the field of social simulations is the widespread absence of empirical validation, sometimes combined with a certain unwillingness to see this as a problem.

In a meta-study on agent-based-modeling (ABM), which is one very important sub-discipline of social simulations, Heath et al. (2009) find that the models in 65% of surveyed articles have not properly been validated, which they consider "a practice

[8] There are scientists who deny even this and who also believe that without formal models no explanation of any sort is possible in history or social science. I am a bit at loss for giving proper references for this point of view, because I have mostly been confronted with it either in discussions with scientists or by anonymous referees of journals of analytic philosophy. The published source I know of that comes closest to this stance is the keynote "Why model?" by Epstein (2008), which I have discussed in Arnold (2014).

that is not acceptable in other sciences and should no longer be acceptable in ABM practice and in publications associated with ABM" (4.11). While some of these not-validated simulations can serve a purpose as thought experiments that capture some relevant connection in an idealized and simplified form (Reutlinger et al. 2017), many of them are merely follow-ups to existing simulations and bear little relevance of their own. The practice of publishing simulations without empirical validation and seemingly little (additional) theoretical relevance is so widespread that it has been termed the YAAWN-Syndrome where YAAWN stands for "Yet Another Agent-Based Model ... Whatever ... Nevermind" (O'Sullivan et al. 2016). The fact that such a term has been coined is an indication that the ABM-community is growing weary of unvalidated or otherwise uninteresting simulations. Thus, the situation may change in the future. For the time being, lack of validation is still a problem.

To be sure, agent-based-modeling is a broad field. On the one hand side there are very theoretical simulations that set out from abstract concepts but without any particular application case in mind. And on the other hand, there exist simulations that are right from the start related to a particular empirical setting. The latter kind of simulations is typically found in corporate or political consulting. I am going to look at the theoretical simulations first and then consider the more applied kinds of simulations later.

Naturally, unvalidated simulations are much more prevalent among the theoretical simulations, where the lack of empirical validation is sometimes not even perceived as a problem. This may be illustrated by a quotation from an interview with a philosopher who has produced models of opinion dynamics (Hegselmann and Krause 2002) that have frequently been cited in other modeling studies but that have not been empirically validated

> None of the models has so far been confirmed in psychological experiments. Should one really be completely indifferent about that? Rainer Hegselmann becomes almost a bit embarrassed by the question. "You know: In the back of my head is the idea that a certain sort of laboratory experiments does not help us along at all." (Grötker 2005, p. 2)

But if laboratory experiments do not help us along, how can models that have never been confirmed empirically either by laboratory experiments or by field research help us along? This lack of interest in empirical research is all the more surprising as opinion dynamics concern a field with an abundance of empirical research. Naively, one should assume that scientists have a natural interest in finding out whether the hypotheses, models and theories they produce reflect empirical reality. That this is obviously not always the case, confirms Kuhn's view that the criteria by which scientific research is judged are also set by the paradigm that guides the thinking of the researchers and that there is no such thing as a "natural" scientific method independent of paradigms. However, even Kuhn's mild relativism would rule out science without any form of empirical validation as unrewarding.

The lack of empirical concern within the field of social simulations can furthermore be attributed to another working mechanism of paradigms that Kuhn identified, namely, the role of *exemplars*. As mentioned earlier, according to Kuhn scientific practice is not guided by the abstract rules of a logic of scientific discovery. Instead, scientists follow role models or *exemplars* of good scientific practice.

Some very influential role models in the field of social simulations concern simulations that have never successfully been validated. The just mentioned opinion dynamics simulation by Hegselmann and Krause is one example for this kind of role model. But the arguably most famous unvalidated model that serves as an exemplar in Kuhn's sense is Robert Axelrod's "Evolution of Cooperation" (Axelrod 1984). Despite the fact that the reiterated Prisoner's Dilemma simulations that Axelrod used as a model for the evolution of cooperation had turned out to be a complete empirical failure by the mid-1990s (Dugatkin 1997) and despite the devastating criticism Axelrod's approach had received from theoretical game theory (Binmore 1994, 1998), it continues to be passed down as a role model of social simulations until this day. In a journal article from 2010 in the prestigious *Science*-journal, where a similar research design as Axelrod's was employed, it is mentioned as a role model that has been "widely credited with invigorating the field" (Rendell et al. 2010, 2008f.). And one can easily find recent studies (Phelps 2016) that naively pick up Axelrod's study as if no discussions concerning its robustness, its empirical validity or its theoretical scope had ever taken place in the meantime. If simulation research designs without proper validation such as Axelrod's continue to be treated as exemplars, it is no surprise that many social simulations lack proper validation.

Now, there are two caveats: First, in some cases unvalidated simulations can serve a useful scientific function, among other things as thought-experiments. Of a thought experiment one usually does not require empirical validation. Thus, if Axelrod's evolution of cooperation or Hegselmann's and Krause's opinion dynamics could be considered thought experiments their status as role models in connection with their lack of empirical validation could not be taken as an indication that social simulations still remain in a prescientific stage. However, the way that both these simulations functioned as role models was not by their (potential) use as thought-experiments, but as a research programme. Indeed, it would be hard to justify the literally dozens if not hundreds of follow-up simulations to Hegselmann-Krause or Axelrod as thought experiments without invalidating the category of a thought-experiment as a useful scientific procedure. But it has to be kept in mind that not any kind of unvalidated simulation is an indication of prescientific fiddling about.

Second, and more importantly, not all simulation traditions have, of course, remained as disconnected from empirical research as Axelrod's Evolution of Cooperation and Hegselmann's and Krause's opinion dynamics simulations. One example is the Garbage-Can-Model (GCM) by (Cohen et al. 1972) which describes decision making inside organizations with a four component model, taking "problems", "solutions", "participants" and "opportunities" into account. This model is highly stylized and, because of this, would be difficult to validate directly. Nevertheless, it is frequently referred to in studies on organizational decision making, including empirical studies.

But why, one may ask, could the connection to empirical research, or more generally, other kinds of research on organizational decision-making be established in this case while it failed in the aforementioned cases? There are several possible reasons:

- Modeling organizational decision-making is a much more restricted topic than, say, modeling evolution of cooperation in general. This makes it easier to find the right abstraction level for modeling. While biologists complained about simulations of the evolution of cooperation that "Most repeated animal interactions do not even correspond to repeated games." (Hammerstein 2003, p. 83), researchers from organizational science have no such difficulties in relating to the Garbage-Can-Model in their case studies (Fardal and Sornes 2008; Delgoshaei and Fatahi 2013).
- Within organization theory working with stylized descriptions is generally accepted. Thus, the target that the simulation model had to match was an already highly stylized verbal description. (Nonetheless, the simulation model did not represent the verbal description faithfully (Fioretti and Lomi 2008, p. 1.4)) It is much easier to cast a stylized verbal description convincingly into a simulation model than, say, a thick historical narrative as in one of Axelrod's suggested application cases.
- For the study of organizational decision-making the Garbage-Can-Model seems to serve as a kind of vantage point. It helps to analyze and communicate organizational decision making problems by relating a particular decision-making situation to the model—even if the model is only used as a conceptual reference framework and the actual simulation results are ignored.[9] Because of its popularity the Garbage-Can-Model could even be considered an exemplar in Kuhn's sense. To serve as a vantage point, a model does not need to be empirically validated or even testable. It stands to reason, though, that it still needs to be "realistic enough" in some weaker sense to serve this purpose.
- While for the latter purpose (vantage point) a stylized verbal description could suffice, simulation models have the advantage that they can be run. This allows to generate hypotheses about the simulated process which can help to establish the basic plausibility of the model, if the simulation itself and its results are plausible in view of the prior knowledge about the simulated process.[10] In the case of the GCM the model establishes the connection between a certain structure of the decision-making process and certain characteristics of the outcome, like how efficiently problems will be solved. In a verbal description this connection can be maintained, but not be demonstrated. A simulation can show that such a connection exists, even if only within the model.

In view of the possible functions of communication and hypotheses-generation, one can argue that models like the Garbage Can Model can be useful in the context

[9] This seems to be the standard case for applying the GCM in organizational science. See Fardal and Sornes (2008) and Delgoshaei and Fatahi (2013) for example. It will be interesting to see whether the more refined simulation models of the GCM that have been published more recently (Fioretti and Lomi 2008) will bring about an increased use of simulation models in applied studies referring to the GCM or not.

[10] This is precisely where Axelrod's simulations was lacking, because (a) his tournament of reiterated Prisoner's Dilemmas is too far removed from the phenomenology of either animal or human interaction to be prima facie plausible, and (b) his results were—unbeknownst to him—highly volatile with respect to the simulation setup and thus also lack plausibility.

of empirical research even without being empirically validated themselves. Still, the question remains what characteristics a model of this kind must have to be considered useful or suitable, or how one can tell a good model from a bad model. There seems to exist an intuitive understanding within the scientific communities habitually using these models, but it is hard to find any explicit criteria. This strengthens the impression that a paradigm of validation is not yet in place, at least not for the more theoretical simulations.

What about applied simulations, though? Agend-based models are, among other things, used to give advice about particular policy measures, like introducing a new pension plan (Harding et al. 2010) or determining the best procedures for research funding (Ahrweiler and Gilbert 2015). Obviously, validation is of considerable importance if simulations are used for political consulting. So, how do scientists who apply social simulations get around the restriction that the simulation results often cannot directly be compared with measurable empirical data? In particular, how can simulations be validated that are meant to evaluate the possible consequences of policy measures that might never be implemented?

In their discussion of the validation of the SKIN-model, which simulates knowledge dynamics in innovation networks, Ahrweiler and Gilbert (2015, Sect. 1.1.2) do not even assume that there exist objective observations independent of a concrete research goal or question.[11] At least for the sake of the argument they even accept the view that the observation of a social process is a construct of this process or "what you observe as the real world" (Ahrweiler and Gilbert 2015, Sect. 1.2), just like the simulation of the same process is another construct of this process. However, since the authority over what is observed as the real world lies with the "user community" (Ahrweiler and Gilbert 2015, Sect. 1.3), the output of a simulation can meaningfully be compared with the observations.

Since the construction of the simulation as described by (Ahrweiler and Gilbert 2015, Sect. 2.4) is a process in which the user community is deeply involved, it is tempting to raise the question how unbiased this kind of validation really is. After all, an administration assigning the task of examining the potential for enhancement of their administrative procedures to a team of simulation scientists might be more interested in the vindication of certain administrative procedures than in their unbiased assessment. However, the "user community view" as described by (Ahrweiler and Gilbert 2015) depicts only the outline of the construction and validation process of applied agent-based models. A more detailed analysis of the validation of applied agent-based-models as provided by (Harding et al. 2010) reveals that there exists a whole array of validation procedures which, if executed properly, limits the risk of producing biased or arbitrary results. For the Australian Population and Policy Simulation Model Harding et al. (2010) report, among other measures: (i) the calibration and benchmarking of the simulation with available cross-sectional and

[11] They discuss this under the heading of "theory-ladenness of observations", though their examples suggest that the issue at stake is rather different interpretations of observations or a focus on different observations depending on the research questions than different observations due to a different theoretical background.

longitudinal data, (ii) the comparison of the simulation model's projection with that of other models, (iii) the modular structure and separate evaluation of each module, (iv) the examination, if both the individual agent's simulated life histories and the summary statistics yield reasonable results. The impact of proposed policy measures as revealed by the simulation can by its very nature not beforehand be compared with empirical data. However, one can contend that in the context of policy advise a simulation is sufficiently validated, if it leads to policy decisions that are better grounded than they would be without running a simulation model.

Where does this leave us? Are social simulations still in a prescientific stage with respect to their validation? On the one hand there is a widespread lack of proper validation and the impression that the increasing number of published agent-based models does not necessarily pay off in terms of further deepening our understanding of the simulated processes. While other quality issues of agent-based models, such as their reproducibility and mutual comparability, have been addressed in recent years,[12] there is still no common understanding concerning how agent-based models should be validated. So far, the textbooks on agent-based simulations have little to say about validation. With the central issue of validation still being unresolved, the field of social simulations does yet seem to have matured into a normal science in the sense of Kuhn. The situation can positively be a described as a phase of humble beginnings in the sense of the interpretation of Feyerabend's anarchic epistemology that was given earlier.

On the other hand, scientists that apply agent-based models to particular empirical processes typically invest considerable time and effort into the validation of their simulations and employ a diverse set of validation procedures to ensure the credibility of their simulations. So, we might indeed be witnessing a paradigm of validation of applied agent-based models in the making. It is, so far, only in the making, because the various validation procedures and criteria used by the practitioners do not yet seem to have been consolidated to a degree where they become textbook knowledge.

8.5 Summary and Conclusions

Putting it all together, we arrive at fairly conservative conclusions: Kuhn's theory of scientific revolutions and his concept of a paradigm does not have any particular consequences for the validation of simulations. At least it does not have consequences that are any different from those it has for the validation of theories or non-simulation models. And neither do computer simulations require us to reconsider Kuhn's theory or related topics like the Duhem–Quine thesis. This result is somewhat unspectacular, but it may be clarifying. With regard to the discussion about the novelty of computer simulations it means that, whatever the novelty may be, neither the introduction of computer simulations nor their validation is or requires a Kuhnian revolution.

[12] A most notable initiative in this respect has been the introduction of the ODD Protocol for the standardized description of agent-based-models (Railsback and Grimm 2012).

The coexistence of multiple paradigms in the social sciences is a challenge for Kuhn's theory in its original form. But, again, the validation of simulations does not raise any specific problems in this context. Presently, many social simulations suffer from the fact that for the lack of proper validation they are quite uninformative about their target system. Although, there are also examples where social simulations do contribute to the understanding of the target system, the field as a whole does not yet seem to have become normal science in the sense of Kuhn. This is most notably due to the fact that—as of now—there exists no commonly shared understanding of the validation requirements of social simulations.

References

Ahrweiler, P., & Gilbert, N. (2015). The quality of social simulation: An example from research policy modelling. In M. Janssen, M. A. Wimmer, & A. Deljoo (Eds.), *Policy practice and digital science: Integrating complex systems, social simulation and public administration in policy research* (pp. 35–55). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-12784-2_3. ISBN: 978-3-319-12784-2.

Arnold, E. (2013). Experiments and simulations: Do they fuse? In E. Arnold, & J. Duran (Eds.), *Computer simulations and the changing face of scientific experimentation* (pp. 46–75). Newcastle: Cambridge Scholars Publishing. 978-1443847926.

Arnold, E. (2014). What's wrong with social simulations? *The Monist*, *97*(3), 361–379. https://doi.org/10.5840/monist201497323. ISSN: 0026-9662.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Axtell, R. L., et al. (2002) Population growth and collapse in a multiagent model of the Kayenta Anasazi in Long House Valley. *Proceedings of the National Academy of Sciences*, *99*(3), 7275–7279. https://doi.org/10.1073/pnas.092080799. ISSN: 0027-8424.

Beisbart, C. (2011). *A transformation of normal science. Computer simulations from a philosophical perspective*. unpublished.

Binmore, K. (1998) *Game theory and the social contract II. Just playing*. Cambridge, Massachusetts/London, England: MIT Press.

Binmore, K. (1994). *Game theory and the social contract I. Playing fair*. Fourth printing. Cambridge, Massachusetts/London, England: MIT Press.

Bird, A. (2013). Thomas Kuhn. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Fall 2013. Metaphysics Research Lab, Stanford University.

Cartwright, N. (1983). *How the laws of physics lie*.Clarendon paperbacks. Oxford University Press. Clarendon Press. ISBN: 9780198247043.

Carusi, A., Rodriguez, B., & Burrage, K. (2013). Model systems in computational systems biology. In E. Arnold, & J. Duran (Eds.), *Computer simulations and the changing face of scientific experimentation*. (Chap. 6).

Cohen, M. D., March, J. G., & Olsen, J. P. (1972). A grabage can model of organizational choide. *Administrative Science Quarterly*, *17*, 1–25.

Delgoshaei, B., & Fatahi, M. (2013). Garbage can decision-making in a matrix structure. A Case study of linköping university. Linköping University/Department of Management and Engineering. urn:nbn:se:liu:diva-95612.

Dugatkin, L. A. (1997). *Cooperation among animals*. Oxford University Press.

Dupré, J. (1994). Against scientific imperialism. *Philosophy of Science Association Proceedings*, *2*, 374–381.

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 12. ISSN: 1460-7425.

Epstein, J. M., & Axtell, R. L. (1996). Growing artificial societies: Social science from the bottom up. MIT Press.

Fardal, H., & Sørnes, J. O. (2008). Is strategic decision-making. A garbage can view. *Issues in Informing Science and Information Technology*, 5.

Feyerabend, P. (1975/1983). *Wider den Methodenzwang*. Suhrkamp Verlag.

Fioretti, G., & Lomi, A. (2008). An agent-based representation of the garbage can model of organizational choice. *Journal of Artificial Societies and Social Simulation*, *11*(1), 1. ISSN: 1460-7425.

Gilbert, N., & Troitzsch, K. (2005). *Simulation for the social scientist*. New York: Open University Press.

Gleick, J. (2011). *Chaos: Making a new science*. Open Road Media.

Grötker, R. (2005). *Reine Meinungsmache*. German: Technology Review (heise Verlag). http://%5C-www.%5C-heise.%5C-de/%5C-tr/%5C-artikel/%5C-Reine-%5C-Meinungsmache-%5C-277359.%5C-html.

Guala, F. (2002). Models, simulations and experiments. In L. Magnani & N. Nersessian (Eds.), *Model-based reasoning: Science, technology, values* (pp. 59–74). Kluwer Academic Publishers.

Hammerstein, P. (2003). Why is reciprocity so rare in social animals? A protestant appeal. In P. Hammerstein (Ed.), *Genetic and cultural evolution* (pp. 83–94). Cambridge, Massachusetts/London, England: MIT Press in cooperation with Dahlem University Press. (Chap. 5).

Harding, S. G. (Ed.). (1976). *Can theories be refuted? Essays on the Duhem-Quine thesis*. Kluwer.

Harding, A., Keegan, M., & Kelly, S. (2010). Validating a dynamic population microsimulation model: Recent experience in Australia. *International Journal of Microsimulation*, *3*(2), 46–64.

Hartmann, S. (1996). The world as a process: Simulations in the natural and social sciences. In R. Hegselmann et al. (Ed.) *Simulation and modelling in the social sciences from the philosophy of science point of view* (pp. 77–110).

Heath, B., Hill, R., & Ciarallo, F. (2009). A survey of agent-based modeling practices (January 1998 to July 2008). *Journal of Artificial Societies and Social Simulation (JASSS)*, *12*(4), 9.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, *5*(3), 1.

Imbert, C. (2017). Computer simulations and computational models in science. In L. Magnani, & T. Bertolotti (Eds.), *Springer handbook of model-based science* (pp. 735–781). https://doi.org/10.1007/978-3-319-30526-4.

Kästner, J., & Arnold, E. (2013). When can a computer simulation act as substitute for an experiment? A case-study from chemisty. In E. Arnold (Ed.), *Homepage Eckhart Arnold*, preprint.

Kornmesser, S. (2014). Scientific revolutions without paradigm-replacement and the coexistence of competing paradigms: The case of generative grammar and construction grammar. *Journal for General Philosophy of Science*, *45*(1), 91–118. https://doi.org/10.1007/s10838-013-9227-3. ISSN: 1572-8587.

Kuhn, T. S. (1976). Suhrkamp: Die Struktur wissenschaftlicher Revolutionen.

Kuhn, T. S. (1977). *The essential tension. Selected studies in scientific tradition and change*. The University of Chicago Press.

Lenhard, J. (2007). Computer simulation: The cooperation between experimenting and modeling. *Philosophy of Science*, *74*(2), 176–194. https://doi.org/10.1086/519029.

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Modern Physics, 41,* 253–262.

Morgan, M. S. (2003). Experiments without material intervention. Model experiments, virtual experiments, and virtually experiments. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 216–233). University of Pittsburgh Press.

Morgan, M. S., & Morrison, M. (Eds). (1999). *Models as mediators. Perspectives on natural and social science*.

Morrison, M. (2009). Models, measurement and computer simulation: The changing face of experimentation. *Philosophical Studies*, *143*, 33–57. https://doi.org/10.1007/s11098-008-9317-y.

Moses, J. W., & Knutsen, T. L. (2012). *Ways of knowing. Competing methodologies in social and political research* 2nd (first ed. 2007). London: Palgrave Macmillan.

O'Sullivan, D., et al. (2016). Short communication. Strategic directions for agent-based modeling: Avoiding the YAAWN syndrome. *Journal of Land Use Science*, *11*(2), 177–187.

Parker, W. S. (2009). Does matter really matter? Computer simulations, experiments, and materiality. *Synthese*, *169*, 483–496. https://doi.org/10.1007/s11229-008-9434-3.

Peschard, I. (2010). *Review of Eric Winsberg's. Science in the age of computer simulation*. University of Chicago Press. Notre Dame Philosophical Reviews (Mar. 31, 2011).

Phelps, S. (2016). An empirical game-theoretic analysis of the dynamics of cooperation in small groups. *Journal of Artificial Societies and Social Simulation*, *19*(2), 4. https://doi.org/10.18564/jasss.3060. ISSN: 1460-7425.

Preston, J. (2016). Paul Feyerabend. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Winter 2016. Metaphysics Research Lab, Stanford University.

Railsback, S. F., Grimm, V. (2012). *Agent-based and individual-based modeling. A practical introduction*. Princeton University Press.

Rendell, L., et al. (2010). Why copy others? Insights from the social learning strategies tournament. *Science*, *328*, 208–213. https://doi.org/10.1126/science.1184719.

Reutlinger, A., Hangleiter, D., & Hartmann, S. (2017). Understanding (with) toy models. *The British Journal for the Philosophy of Science*, axx005. https://doi.org/10.1093/bjps/axx005.

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, *1*(2), 143–186. https://doi.org/10.1080/0022250X.1971.9989794. ISSN: 0022-250X.

Schurz, G. (2014). Koexistenz und Komplementarität rivalisierender Paradigmen: Analyse, Diagnose und kulturwissenschaftliches Fallbeispiel. In S. Kornmesser, & G. Schurz (Eds.), *Die multiparadigmatische Struktur der Wissenschaften* (pp. 47–62). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-00672-3_2. ISBN: 978-3-658-00672-3.

Sismondo, S. (2010). *An introduction to science and technology studies* (2nd ed.). Wiley.

Squazzoni, F., & Casnici, N. (2013). Is social simulation a social science outstation? A bibliometric analysis of the impact of JASSS. *Journal of Artificial Societies and Social Simulation*, *16*(1), 10. ISSN: 1460-7425.

Winsberg, E. (2015). Computer simulations in science. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Summer 2015. Metaphysics Research Lab, Stanford University.

Winsberg, E. (2003). Simulated experiments: Methodology for a virtual world. *Philosophy of Science*, *70*, 105–125.

Winsberg, E. (2009). A tale of two methods. *Synthese*, *169*, 575–592. https://doi.org/10.1007/s11229-008-9437-0.

Chapter 9
Understanding Simulation Validation—The Hermeneutic Perspective

Nicole J. Saam

Abstract The thesis of a hermeneutic perspective on validation in simulation has existed ever since Kleindorfer et al. (*Manag Sci* 44:1087–1099, 1998) published their overview of various positions in the philosophy of science. This chapter introduces the distinction between a hermeneutics *in* validation and a hermeneutics *of* validation. I argue that the hermeneutic perspective according to Kleindorfer, O'Neill and Ganeshan, which qualifies as a hermeneutics *in* validation perspective, is rather fruitless. Instead, a hermeneutics *of* simulation validation is proposed on the basis of Gadamer's philosophical hermeneutics. The goal of the hermeneutics *of* validation is to understand simulation validation. The challenge is to set up a hermeneutic situation in the first place. Hermeneutic aims to demonstrate how simulation validation is historically situated, revealing the hidden prejudice (prejudgement) in validating, and distinguishing between legitimate prejudice and prejudice that has to be overcome. Understanding simulation validation is a dialogic, practical, situated activity.

Keywords Simulation validation · Philosophical hermeneutics · Understanding · Interdisciplinary dialogue

9.1 Introduction

In 1998, Kleindorfer et al. (1998) published an article in which they examined how well various positions in philosophy of science can account for validation of computer simulations. Remarkably, they not only considered standard positions from the history of philosophy, such as rationalism and classical empiricism, or from more recent, analytical philosophy such as logical positivism, diverse falsificationist positions, Kuhnianism and Bayesianism. Rather, they ended up favouring a hermeneutic perspective on the validation of simulations. Hermeneutics is presented as a solution

N. J. Saam (✉)
Institut für Soziologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
e-mail: nicole.j.saam@fau.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_9

to the debate between objectivists and relativists—a debate which they consider to be 'the underlying plate tectonics in the simulation validation problem' (Kleindorfer et al. 1998, p. 1088). Naylor and Finger's (1967) article is presented as the classic view of positive simulation validation, which they term 'objectivist', while Barlas and Carpenter's (1990) article is presented as the 'relativist' antithesis, drawing on the philosopher Thomas Kuhn (1970). Kleindorfer et al. (1998) follow American philosopher Bernstein (1983) in arguing that there is a Cartesian legacy in the debate, stating that 'many simulation modellers apparently believe that model validation is an 'either/or' proposition' (Kleindorfer et al. 1998, p. 1088), and they seek a means for breaking out of this dichotomy. Following Bernstein, they present the hermeneutics of German philosopher Hans-Georg Gadamer as 'a philosophical fulcrum' for transcending the objectivist versus relativist debate (p. 1097). They argue that whereas, in general, philosophy of science has begun to turn away from the Cartesian legacy, the discussion of simulation validation still assumes an 'either/or' situation. In this state of affairs, Kleindorfer et al. (1998, p. 1087) 'set out a perspective'. The hermeneutic position is favoured, since it refers to Ancient practical wisdom (*phronesis*) and requires that practitioners conduct 'meaningful dialogue on a model's warrantability' (Kleindorfer et al. 1998, p. 1098).

While this article is much quoted and its statements on the hermeneutic account have often been reproduced, it has never been discussed or elaborated upon in depth. A description of the hermeneutic approach has yet to be extended beyond the initial sketch of two pages. Meanwhile, the state of two other related philosophical debates is unfavourable for hermeneutics: (1) the much broader attempt to establish a hermeneutics of the natural sciences (e.g. Crease 1997; Heelan 1998; Feher et al. 1999)—separately from a Kuhnian history of science perspective—seems to have failed, as the articles by Markus (1987), Eger (1997) and Kisiel (1997) indicate. Most recently, this failure is reflected in the lack of a chapter on hermeneutics and the natural sciences in part IV, 'Hermeneutic Engagements', of Malpas and Gander's (2014) *Routledge Companion to Hermeneutics*. (2) The current philosophical debate on understanding simulation models (Humphreys 2004, 2009; Reutlinger et al. 2018; Saam 2017) refrains from any reference to hermeneutic perspectives—whether traditional or modern, although understanding is a topic at the centre of hermeneutics. Responding to this state rather reminiscent of a standstill or even a drawback of hermeneutics as applied to (natural) science or scientific methods, this chapter discusses the fruitfulness of a hermeneutic perspective on validation in simulation.

To this end, I first introduce the distinction between a *hermeneutics in validation* and a *hermeneutics of validation* (Sect. 9.2). I then show that the perspective of a hermeneutics *in* simulation validation as proposed by Kleindorfer et al. (1998) is rather fruitless by arguing that their perspective rests on conditions that are not fulfilled and on some misunderstandings (Sect. 9.3). I proceed by introducing the perspective of a hermeneutics *of* simulation validation. Connecting Gadamer's (2013 [first German edition 1960]) philosophical hermeneutics and his ideas of prejudice (German *Vorurteil;* please note that Gadamer has a positive conception of prejudice in terms of prejudgment), circularity and historicity to insights from the hermeneutics of the natural sciences discourse, I argue for four claims of a hermeneutics of validation: understanding simulation validation requires the setup of a hermeneutic situation. The simulating scientist shows a hermeneutic naiveté vis-à-vis her validation practices, as opposed to the philosopher of science and the methodologist. This naiveté is overcome in interdisciplinary dialogue. Major hermeneutic tasks are showing how simulation validation is historically situated, revealing the hidden prejudices in validating, as well as distinguishing legitimate prejudice from the prejudice that has to be overcome (Sect. 9.4). In the discussion, I consider the limitations to, and the significance of, a hermeneutics *of* validation (Sect. 9.5). The conclusion suggests issues for future hermeneutic dialogues.

This chapter uses Schlesinger's SCS definition of model validation ('*the substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model*'; Schlesinger 1979, p. 104) as a point of reference for defining simulation validation. Prior to any further analysis, this definition offers no idea of how a hermeneutic perspective might contribute to simulation validation.

9.2 Hermeneutics *in* Versus Hermeneutics *of* Validation

In this chapter, hermeneutics will be used to refer to a philosophical discipline concerned with analysing the conditions of understanding. Hermeneutics emerged as a crucial branch of text studies. Later on, it came to include the study of ancient and classic cultures, as well as of day-to-day life, and existence as such. As Ramberg and Gjesdal (2005) emphasize, the term *hermeneutics* covers both the art of understanding and interpretation of linguistic and non-linguistic expressions (call this first-order hermeneutics) as well as the theory thereof (second order). As a philosophical discipline, based on Gadamer's (2013) account of hermeneutics, three levels may be distinguished: hermeneutics as an art aiming at the understanding (1) of any kind of text; (2) of human life in general, in particular as takes place in language and (3) of existence as such. All understanding is, according to Gadamer, interpretative, i.e. disclosure of meaning.

It is important to understand Gadamer's concept of text. He uses text as a model. Everything is mediated in the universal medium of language. In everyday life, 'text' refers to an object that can be read, something written. A broader understanding of the text recognizes that everything that is mediated in the universal medium of language—utterances, verbal communication, e.g. regarding simulation validation practices, even thoughts—can be transformed into text. Texts can be seen as objectifications of human experience. Finally, all of human life that takes place in language can be studied as text. In this way, Gadamer uses the textual model to develop his hermeneutic conception, which is by no means restricted to 'texts' alone. The concept of validation texts referred to below is based on this Gadamerian understanding of text and includes not only any sort of text, such as textbooks or scientific articles on simulation validation, but all sorts of validation practices and validation knowledge which are mediated in the medium of language. In the same way, the endeavour of

Table 9.1 Hermeneutics *in* versus hermeneutics *of* validation

	Hermeneutics *in* validation	Hermeneutics *of* validation
Object to be understood	Simulation models and their results relative to target	Simulation validation procedures and related practices
Question	How can the model and its results be validated?	How can simulation validation be understood? How are acts of understanding involved in the validation of models and their results?
Interpreter	Working scientist	Philosopher of science, methodologist, working scientist
Hermeneutics applied	First-order art	First-order art and second-order theory

understanding simulation validation is not restricted to the means of reading literature on validation.

If we want to apply the hermeneutic perspective to simulation validation this, therefore, means asking how understanding is related to it. Here, I introduce a major distinction (see Table 9.1): the difference between a 'hermeneutics *in*' and a 'hermeneutics *of*' addresses the position of the interpreter. If the interpreter is a simulation scientist who uses the hermeneutic perspective *when validating* her simulation model, we shall refer to this as hermeneutics *in* validation (asking the question of how the model and its results can be validated). This situation has to be distinguished from a hermeneutic perspective that is taken from the position of an observer. Philosophers of science, methodologists or simulating scientists then *reflect* on simulation validation in order to understand this scientific activity and its related practices (asking the question of how simulation validation can be understood). Revisiting Ramberg and Gjesdal's distinction (2005), 'hermeneutics *in*' refers to the first-order art while the 'hermeneutics *of*' validation refers to the second-order theory of understanding and interpretation.

9.3 Hermeneutics *in* Validation

The first attempt to outline a hermeneutic position in simulation validation was presented by Kleindorfer et al. (1998), based on a conference paper by Kleindorfer and Geneshan (1993), and qualifies as a hermeneutics *in* validation perspective.

In the following, I reconstruct their claims in Sect. 9.3.1. In Sect. 9.3.2, I interpret their outline as an effort to directly apply hermeneutics to validation. I argue that this effort is rather fruitless, since their perspective rests on conditions that are not given or are based upon some misunderstandings. The second option would be to apply hermeneutics in the sense of seeking analogies. However, there prove to be important disanalogies. This raises the question of whether the remaining claims are essentially hermeneutic or rather are supported by different philosophical perspectives too (Sect. 9.3.3).

9.3.1 Hermeneutics According to Kleindorfer, O'Neill and Ganeshan

In their article, Kleindorfer et al. (1998) provide a description of various philosophical positions and summarize the problems and the kinds of arguments these positions each allow in arriving at defensible simulation models. The motivation for Kleindorfer et al. (1998, p. 1087) philosophical sketch is a perceived 'doubt and even anxiety among simulation modellers as to what the methodologically correct guidelines or procedures for validating simulating models should be'. Referring to Bernstein (1983), they describe this anxiety as Cartesian. It is related to an *either (confirmed)/or(refuted)* distinction in validation, while in practice confirmation is a matter of degree. As Oreskes et al. (1994, p. 643) emphasize: 'In practice, few (if any) models are entirely confirmed by observational data, and *few* are entirely refuted'. Kleindorfer et al. (1998, p. 1087) intention is to 'free the practitioner to pursue a varied set of approaches to validation with diminished burden of methodological anxiety'. Consequently, they do not prescribe a particular technique or algorithm, but offer hermeneutics as a perspective that frees the validating simulation scientist.

These statements resonate with Gadamer's (2013) major theme in his most important book *Truth and Method*. There, he developed his philosophical hermeneutics which provides an account of the proper grounds for understanding. He rejects the attempt to found understanding on any ('scientific') method or set of rules, arguing that there is no methodology that describes the means by which to arrive at an understanding of human life. Neither is there any such methodology that is adequate for understanding nature. Insisting on the limited role of method, he emphasizes that understanding is a dialogic, practical, situated activity. It seems plausible that Kleindorfer, O'Neill and Ganeshan felt attracted by Gadamer's claim concerning the limited role of method and the priority that should be given to dialogue.

Kleindorfer et al. (1998, p. 1090) summarize their paper by saying that the epistemological focus of hermeneutics rests on interpretation and understanding through dialogue and practice.[1] They contrast the epistemological focus of hermeneutics to

[1] The presentation of the hermeneutic perspective on simulation validation by Kleindorfer et al. (1998, pp. 1096–1098 and one row in Table 9.1, p. 1090) amounts to no more than two pages in total. Gadamer's hermeneutics as put forward in *Truth and Method* (Gadamer 2013 [original 1960]) serves as a major reference, although it is the reading of Gadamer's hermeneutics by Bernstein (1983) which the authors actually adopt. This becomes explicit on p. 1098, where they refer to 'Bernstein's hermeneutics'. They introduce Bernstein as a philosopher who presents Gadamer's hermeneutics as a philosophical fulcrum for transcending the polarity of the foundationalist versus anti-foundationalist debate. The authors apply Bernstein's hermeneutics to validation in simula-

other foci: the logical justification of knowledge claims (attributed to rationalism, classical empiricism, logical positivism), theories as frameworks for prediction and testing (instrumentalism, dogmatic and methodological falsificationism) , consistent treatment of probabilistic induction (Bayesianism) and progressive historical growth of knowledge (Kuhnianism, Lakatos' methodology of scientific research programmes).

In order to facilitate the discussion of Kleindorfer, O'Neill and Ganeshan's hermeneutic perspective on simulation validation, I will now reconstruct their most important statements in five claims. First of all, Kleindorfer, O'Neill and Ganeshan emphasize the contribution that openness and reason make to the growth of knowledge. This claim is transferred to simulation validation. They put forward claim C-Open to address the issue of openness:

C-Open: the model builders are free to establish and increase the credibility of the model by any reasonable means.

Kleindorfer et al. (1998, p. 1098) state that the validation of a model can be achieved in any reasonable manner, and they explain that by reasonable means/manner they mean 'historically situated dialogue, judgment and practical discourse'. This openness includes, for instance, the possibility of meaningfully comparing different models and the involvement of further model stakeholders. They refer back to Bernstein's concept of rationality, which they describe as 'historically situated and practical, involving choice, deliberation and judgement' (Kleindorfer et al. 1998, p. 1097), and to *phronesis*, the term that Aristotele—and Gadamer and Bernstein—used for 'practical wisdom'. A mere glance at Bernstein's (1983) final part IV reveals that the concept of judgment refers to the political philosophy of Hannah Arendt, developed in particular in *The Human Condition* (1958), and that the concept of practical discourse alludes to the discourse theory of Habermas (1984, 1996). Kleindorfer, O'Neill and Ganeshan claim that practical judgement and interactive orientation bring an ethical dimension to scientific validation. They contend that in this way 'we are able to discern the difference between the good and the bad, the worthwhile and the frivolous, the "true" and the "false"' (Kleindorfer et al. 1998, p. 1098). They claim that human judgement and decision enter the process of validation; judgement and decision making cannot be avoided. Quoting Forrester (1961, p. 118), they argue that a choice is made concerning that part of the available knowledge that is to be relied upon. As an example, they turn to the court system, putting forward claim C-Court:

C-Court: the court system is a framework for simulation validation consistent with Bernstein's hermeneutics.

Kleindorfer, O'Neill and Ganeshan argue that to obtain a conviction the guilt of the defendant does not have to be proved. Rather, guilt would have to be established beyond reasonable doubt. Biases and prejudice on the part of the jurors would presumably contribute to what is considered to be 'reasonable'. In the next paragraph,

tion, taking two quotations from Barlas and Carpenter (1990) and Carson (1989) to support their arguments.

they describe the court as a model or metaphor (Kleindorfer et al. 1998, p. 1098). They relate the court metaphor to the openness of the meaningful dialogues on a model's warrantability. They put forward claim C-Part to address the involvement of further model stakeholders beyond the model builders such as the model users and referees of journal articles:

C-Part: the simulation validation procedure favoured by hermeneutics is based upon participation by all interested in the outcome.

Notably, Kleindorfer, O'Neill and Ganeshan's hermeneutic perspective on validation does not rest on the concept of the so-called hermeneutic circle.[2] However, they connect the hermeneutic circle to their concept of understanding simulation results:

C-HC: in simulation we experience cognitive processes as described by the hermeneutic circle.

Kleindorfer et al. (1998, p. 1097) argue that in simulation, there is a persistent play back and forth 'whereby our understanding of general principles is increased as we interpret the particulars in a given application. In the light of that understanding, we simultaneously begin to see the particulars more sharply and are better able to give them meaning'. The term 'general principles' is not specified and may serve as a substitute for the principles governing the modelled system as a whole. Immediately after this statement, they turn to the metaphor of play. They seem to refer to a familiar saying of modellers who describe aspects of their scientific work in simulation as 'playing' with a theory or model. Without any further explanation, they report on the recognition that this 'playing' is perceived as a way of effecting model validation:

C-Play: 'playing' with a theory or simulation model is a way of effecting its validation.

The presentation of Kleindorfer, O'Neill and Ganeshan's hermeneutic perspective by Feinstein and Cannon (2003) basically repeats these claims.

9.3.2 A Reply to Kleindorfer, O'Neill and Ganeshan

Before we discuss the claims C-Open through C-Play, two preliminary remarks seem necessary. They address (i) the theoretical status of the 'hermeneutic position' and (ii) the lack of elaborate claims.

(i) Kleindorfer et al. (1998) seem hesitant to establish a genuine philosophical position. While they announce in the abstract that they will 'set out' a hermeneutic

[2] Several formulations of the term 'hermeneutic circle' are known. The classic notion refers to the back-and-forth movement of thought from the whole to a part of the object of investigation and back to the whole again, each new understanding of the latter modifying the understanding of the former, and vice versa. The objective is to recover the meaning of the object of the investigation.

perspective, they write in the introduction that they will ‘describe’ the implications of hermeneutics to the validation problem in simulation. Altogether, it strikes me that Kleindorfer et al. (1998) use the subjunctive in establishing their claims C-Open, C-Court and C-Part (‘The hermeneutic position would assert… would be consistent with … would be free … would not preclude’), while the indicative mode is used within claims C-HC and C-Play. This gives the sketch of the hermeneutic position an ambiguous status. I have decided to adopt the theoretical claim of Bernstein’s hermeneutics on which Kleindorfer, O’Neill and Ganeshan rely to gain a definite philosophical position wherever the subjunctive mode is used by the authors. Consequently, I take the respective claims (C-Open, C-Court and C-Part) to be descriptive sentences. In contrast, recognizing that the authors are social scientists, I consider claims C-HC and C-Play to be empirical sentences.

(ii) In the absence of elaborate claims, one argument is always pertinent, but not scientifically fruitful: Kleindorfer, O’Neill and Ganeshan’s claims could be rejected because they are explained in insufficient fashion and are much too general. I will follow a different path. My objections will be based on an effort to provide at least some missing explanations in the light of the few hints that Kleindorfer, O’Neill and Ganeshan give. I have chosen this approach in order to begin the discussion of their theses, which is still lacking. Ultimately, my criticism addresses my own reconstructions of what Kleindorfer, O’Neill and Ganeshan have argued, based on Bernstein (1983) as my primary source, since he obviously also served as such for Kleindorfer, O’Neill and Ganeshan. As my goal is not to elaborate Kleindorfer, O’Neill and Ganeshan’s hermeneutic perspective, I will try to keep it brief and only provide the relevant link to Bernstein’s hermeneutics.

9.3.2.1 Dialogue, Judgment and Practical Discourse (C-Open)

To begin with, I do not wish to refute claim C-Open in general. I recognize that historically situated dialogue, judgment and practical discourse may have the liberating effect the authors seek to highlight. However, I disagree with subsuming all three procedures under a hermeneutic position. While the authors seem to follow Bernstein (1983, e.g. p. 110, 112, 176, 219, 229), who argues in favour of the convergence of Gadamer’s hermeneutics, Habermas’ discourse theory and Arendt’s political philosophy, I contend that the latter two have objected to basic hermeneutic assumptions (see the Gadamer Habermas debate; for a recent review of that debate and its outcomes see Smith 2014) or have taken their inspiration from Kant’s *Critique of Judgment* (Arendt) such that the force of the better public argument that they support is not founded on a hermeneutic position. Thus, these thinkers recommend judgment and practical discourse from other philosophical positions beyond hermeneutics. Judgment and practical discourse are not only recommended from a hermeneutic position.

What I miss in relation to claim C-Open is any explanation as to when either historically situated dialogue, judgment or practical discourse may be suitably applied to simulation validation. Are they all reasonable for every problem it entails? This point also holds for the contention that practical judgment and interactive orientation provide an ethical dimension to the practice. When and how can the 'good and the bad' be discerned? What does 'the worthwhile and the frivolous' mean? Why is 'the "true" and the "false"' placed in quotation marks? What practical wisdom is required and applied? While Kleindorfer, O'Neill and Ganeshan quite convincingly relate particular problems in simulation validation to other philosophical positions described and discussed in their previous sections, they do not relate any distinct problem to the application of their hermeneutic position. It seems that there is no problem for their proposed new perspective—apart from the very general Cartesian anxiety. It thus seems that claim C-Open is too broad to give useful advice to practitioners. To illustrate my counterargument, I look more closely at the conditions of the possibility of achieving the validation of a model via historically situated dialogue.

Against claim C-Open, I contend that Gadamer's conditions to enter into the dialogue with the matter at issue are *not* fulfilled in simulation validation practice. In simulation validation, the situatedness is bracketed: according to Gadamer, all understanding directed at the grasp of some particular subject matter is based on a prior understanding—a prior hermeneutic situatedness. There are always 'fore-structures' of understanding, meaning anticipatory structures that allow what is to be interpreted or understood to be grasped in a preliminary fashion. This situatedness is historically determined. However, a reflexive hermeneutic awareness of this historically determined situatedness may be lacking—a situation which has been called 'historical amnesia' by Markus (1987). who had observed that natural scientists are acculturated to write their reports with a depersonalized objectivity that decontextualizes the situational contingencies. 'Bluntly put, the natural sciences, in practice, seem to be in no need of a hermeneutics—they succeed quite well without it' (Markus 1987, p. 8). Two important reasons he presents as to why this should be so are: (1) the success of the practice very much depends on tacit knowledge that is incorporated, e.g. in laboratory activities. There is no pragmatic benefit in reflecting on the implicit hermeneutics operative in these craft skills. (2) Traditions embodied in validation terminology and methods are subject to an accelerated rate of obsolescence, rendering the use of Gadamer's concept of tradition shallow. Markus (1987, p. 46) concludes in his analysis that 'a *reflexive* hermeneutic awareness [is] unnecessary for the successful practice of the natural sciences'. Suppose that a hydrologist evaluates the validity of her groundwater flow simulation model's results (see Chap. 27 by Roache in this volume). She compares the model results and their uncertainties with observational (often experimental) results and their uncertainties. She considers the errors in the simulation result and the experimental result. She reflects on the proposed purpose and domain of applicability of her model. In these evaluations, she will neither consider how the concepts of water and velocity she refers to are historically situated, nor will she reflect on how her concepts of uncertainty and error and her observational methods are historically situated. Instead, the final judgment on the validity of the hydrological model's results will be based on having bracketed these questions. Former methods and techniques of validation, as well as historical concepts of water and velocity, are irrelevant for her present evaluation, because the model assumptions are based on one, present, state-of-the-art concept of water and assumptions from computational physics. The irrelevance of historical concepts of water and velocity is a consequence of the underlying rules of model construction. In one and the same model, a certain theoretical concept should only be defined and implemented in one and the same way (and if this rule is violated during model construction, it is a task for the validator to find that mistake). The validation of this model's results depends on validation methods and techniques that primarily reflect the state of the art and only secondarily on the history of scientific methods. Validation of a model is not achieved via historically situated dialogue—which might indeed free the hydrologist to pursue a varied set of approaches to validation with diminished burden of methodological anxiety—rather it is achieved via thorough evaluations that reflect the state of the art in the methods and techniques that are applied. Stating this, I do not question that dialogue is historically situated. It is. I only argue that the conditions for the possibility of a dialogue are not fulfilled.

Additionally, I point to what I want to call the social scientific misunderstanding of dialogue. Kleindorfer et al. (1998) seem to assume that dialogue requires the discursive encounter of scientists. However, Gadamer's concept of dialogue is philosophical and much broader. It requires an interpreter and a text. To Gadamer, a dialogue is not necessarily a social encounter. This misunderstanding is relevant in Kleindorfer et al. (1998) claims C-Open and C-Court. In their interpretation of the court system as a model and of the court metaphor it becomes obvious that they put forward a social concept.

9.3.2.2 The Court System and Its Openness (C-Court)

At the centre of their court system claim is the justification of validity claims—a major topic in the Habermasian discourse model—and not the Gadamerian idea of pluralistic dialogue between different horizons. Kleindorfer, O'Neill and Ganeshan do not aim at an understanding that occurs as a hermeneutic 'fusion of horizons', nor do they—as Gadamer does—envision a process in which the subject is altered (because the interpreter's horizon is enlarged and enriched). It is not sufficient, however, to claim C-Court. Kleindorfer, O'Neill and Ganeshan's claim is much better suited to Habermas' than to Gadamer's model.

The court metaphor contradicts Kleindorfer, O'Neill and Ganeshan's assumption of an openness in which meaningful dialogue can be conducted. As Doublet (2003, p. 62) argues, legal hermeneutics is dogmatic. There is an authorized interpretation of law from the side of the legislator. Although modern legal hermeneutics also acknowledges alternative perspectives such as textualist accounts (see e.g. Poscher 2014), so-called intentionalist accounts of legal interpretation remain a strong current. This raises doubt as to whether the court metaphor—which is a vague conceptualization anyway—can serve as a framework for simulation validation and warrant the favoured openness.

Not legal (i.e. a hermeneutics directed to the understanding of legal texts), but social science hermeneutics (directed to the understanding of social action) may be applied in court when actions, statements or motivations of the accused person are interpreted. Social science hermeneutics, for which I prefer to use the concept of sociology of understanding, may be considered to be more open-ended than legal hermeneutics. However, it seems that Kleindorfer et al. (1998, p. 1098) rather have in mind legal hermeneutics, as their summarizing statement shows: 'By and large, it is the merits of the case as defined within the parameters of the law that determine a trial's outcome'.

9.3.2.3 Participation and Judgement (C-Part)

Claim C-Part calls for the participation of all who are interested in the outcome—the stakeholders, to use a modern term. Indeed, there are stakeholder approaches to simulation validation, however, these approaches are restricted to action research v and to particular conditions that have been explained based on a pragmatic perspective: action researchers consider the ways in which social reality is an ongoing accomplishment of social actors rather than something that is external to them and that totally constrains them. In particular, social realities are perceived as being local, specific and socially constructed. The local community whose problem is being addressed by the action research is considered to be experts on their own experience. Their local knowledge is explored through communication with the action researcher (see Chap. 17 by Saam in this volume). Kleindorfer, O'Neill and Ganeshan do not provide any specification or qualification addressing social reality as an ongoing accomplishment of social actors. If they had such a constructionist perspective, they would have to make explicit: what is the knowledge which is contributed by the stakeholders to simulation validation? What is its epistemic state compared to the knowledge of the simulating scientist? When should this knowledge be contributed? What are the conditions for the participation of the stakeholders?

Ultimately, claim C-Part seems to be an adaptation of Hannah Arendt's political philosophy as discussed by Bernstein (1983, pp. 210–221). Interested in politics and the public sphere, she contends that each person must be given the opportunity to participate in politics (Arendt 1969, p. 233). A second source is Gadamer's hermeneutics (Bernstein 1983, p. 137). However, it is in Arendt's *Crisis of the Republic* that (political) judgment and participation are related (see the discussion in Bernstein 1983, pp. 207–223). The question of how Arendt's analysis of judgment as an intrinsically political mode of thinking can be transferred to simulation validation is not addressed by Kleindorfer, O'Neill and Ganeshan. As the criterion of equality among citizens cannot simply be transferred from politics to science, the claim is not convincing without further explanation.

9.3.2.4 The Hermeneutic Circle (C-HC)

Claim C-HC is related to simulation validation in an indirect way. Understanding simulation results may be considered a necessary condition for the validation of simulation models and their results. I want to point out that Kleindorfer, O'Neill and Ganeshan's claim is based on an insufficient application of Gadamer's concept of the hermeneutic circle. I do not deny that there is some type of circularity in understanding simulation results. I agree that simulation scientists improve their understanding of the model's results based on their foreknowledge that directs the specification of further simulation experiments. But I am hesitant to apply Gadamer's concept of prejudice and understanding here. What is essential for Gadamer's understanding of the hermeneutic circle is that the hidden prejudice—the kind of prejudice really relevant to hermeneutics—is effective for us via history. Prejudice is revealed as prejudice only in the encounter with tradition. Gadamer (2013, p. 310) argues that 'Understanding is, essentially, a historically effected event'. The cognitive processes described by Kleindorfer, O'Neill and Ganeshan lack this historical dimension. The encounter with my foreknowledge prior to the previous simulation runs is not an encounter with tradition. Second, the understanding that results from this encounter is not a hermeneutic understanding. According to Gadamer, all understanding is disclosure of meaning (*Sinn*). But understanding simulation results are not related to conceiving the meaning of some sort of results. Rather, understanding simulation results is related to giving well-founded answers to what-if-things-had-been-different questions (Saam 2017) or grasping the model (Reutlinger et al. 2018). Thus, while there is some type of circularity in understanding simulation results, this understanding does not qualify as *hermeneutic* understanding and it is not based on the *hermeneutic* circle.

9.3.2.5 Play (C-Play)

In claim C-Play, play is used as a metaphor. It can best be explicated by Bernstein's preferred understanding of play as a 'to-and-fro movement' (Bernstein 1983, p. 121, 171), which he takes from one of Gadamer's analyses of play. I do not deny that there is some sort of to-and-fro movement from simulation results to target, and vice versa, as well as from simulation model to theory, and vice versa, in simulation validation. In simulations that are not based on theory, there may also be such a to-and-fro between model assumptions and experimental results. However, even in this explicated way the claim is much too vague to contribute to a hermeneutics *in* validation, all the more so because the concept of play is not prominent in hermeneutics and cannot be reduced to a to-and-fro movement. Hence, while I do not wish to deny that there is some element of play in simulation (see Saam and Schmidl 2018), I claim that C-Play is inadequate for describing empirical validation practice.

Altogether, Kleindorfer, O'Neill and Ganeshan's sketch of a hermeneutic perspective *in* validation is not convincing. My claim is that the Habermasian discourse model (Habermas 1984, 1996) fits their basic intention much better. This discourse model relies on the force of the better argument among all competent on an issue. It explicitly addresses validity claims and conforms with claim C-Open. Thus, core features outlined by Kleindorfer et al. (1998), such as openness, rationality, judgment, understanding, interpretation, participation and critique, characterize the Habermasian model.

I consider the perspective of a hermeneutics *in* simulation validation as proposed by Kleindorfer et al. (1998) to be rather fruitless, since their perspective rests on conditions that are not given and based upon some misunderstandings. Gadamer's philosophical hermeneutics provides neither a logical justification of knowledge claims, nor theories as frameworks for prediction, testing or probabilistic induction. In contrast to logical positivism, variants of falsificationism (see Chap. 6 by Beven in this volume) and Bayesianism (see Chap. 7 by Beisbart in this volume), its contribution is limited to the level of second-order reflexion. If hermeneutics can make a contribution to simulation validation, it must be on another level. In Sect. 9.4, I will therefore propose a hermeneutics *of* validation.

9.3.3 Claim C-Open—A Second View

To reiterate, I do not seek to refute claim C-Open in general. Notably, Gadamer's (2013) refusal to found understanding on any ('scientific') method or set of rules has some parallel in Feyerabend's (1975) polemic *Against Method*. Thus, the claim that the model builders are free to establish and increase the credibility of the model by any reasonable means is supported by different philosophical perspectives. I recommend seeking support and evidence for this claim from different philosophical perspectives, rather than subsuming it too superficially under a hermeneutics *in* validation or an epistemic anarchism *in* validation. The claim has a liberating effect; the more so the better we understand when and why it is supported.

9.4 Hermeneutics *of* Validation

I take Gadamer's (2013 [1960]) hermeneutics as a starting point for a hermeneutics *of* validation because in his philosophical hermeneutics he establishes a claim to the universality of hermeneutics. As Steinmann (2007, p. 102) has put it, Gadamer's claim as to the universality of hermeneutics indicates the attempt to establish hermeneutics as a 'radically modern epistemology' . According to Gadamer, science, but also art, culture, history and philosophy, comes to understanding only in the universal medium of language. Gadamer (2013, p. 491) postulates that 'man's relation to the world is absolutely and fundamentally verbal in nature, and hence intelligible'. Following Heidegger, he perceives language as a universal ontological structure. Language is 'the basic nature of everything towards which understanding can be directed' (Gadamer 2013, p. 490). Understanding is the ever-present enactment structure of human life, the very mode of human existence. Understanding has to be considered a basic hermeneutic experience, founding all kinds of cognition and their respective ways of knowing and acting. Although Gadamer often refers to the example of understanding a text, his approach is by no means restricted to texts alone. Rather, it holds for everything within the limits of possible human experience. For Gadamer, understanding is not just a kind of knowledge specific to the human sciences. He rejects the methodological reduction and limitation by traditional hermeneutics (up to and including Dilthey). Understanding is '*a universal aspect of philosophy*' (Gadamer 2013, p. 491).

In this way, simulation validation too can become the focus of understanding. Following the distinction of levels in Sect. 9.2, the hermeneutics *of* validation considers validation as a human activity which is linguistically mediated. This practice can be understood referring to the hermeneutics of levels (1) and (2), as introduced in Sect. 9.2 above. I adopt Gadamer's basic hermeneutic ideas of *prejudice*, *circularity* and *historicity*.

In elaborating on this hermeneutic perspective, I first include some findings from the research on a hermeneutics of the natural sciences. Second, there has been a tendency to dissolve hermeneutics and to reduce it to a Kuhnian history of science perspective (see for instance, the last section in D'Agostino 2014) or to a social constructivist or cultural studies of sciences perspective (see the review by Kisiel 1997). My aim is to conserve the originality of the hermeneutic perspective. Gadamer emphasized that he sought to investigate the conditions of possibility for understanding as such.

In the following, I will elaborate four theses: (1) Understanding simulation validation requires a hermeneutic situation. (2) Simulation scientists show a hermeneutic naiveté vis-à-vis their validation practices. (3) Interdisciplinary dialogue constitutes a hermeneutic situation in which the hermeneutic naiveté is lost. (4) Hermeneutic tasks are: showing how simulation validation is historically situated, revealing the hidden prejudices in validating and distinguishing between legitimate prejudice and prejudice that has to be overcome.

As indicated in Sect. 9.2, I will distinguish three groups of interpreters: simulation scientists, methodologists and philosophers of science.

9.4.1 The Requirement of a Hermeneutic Situation

A hermeneutics *of* validation requires the setup of a hermeneutic situation, in Gadamer's definition (2013, pp. 316 f.), a situation in which we encounter the past having to understand the tradition from which we come (such as the concept of text—see Sect. 9.2 above—the concept of the past can be understood in a broad way. For example, in his hermeneutics of conversation, Gadamer describes how two speakers, say ego and alter, exchange opinions and try to understand each other. Ego tries to understand alter based on alter's latest articulated opinion and ego's prior understanding of the whole conversation. The same holds vice versa for alter. Here, the 'past' not only refers to distant history and to our cultural tradition but also to the beginning of that conversation). The awareness of such a hermeneutic situation is not self-evident. It requires an awareness of effective history (German *Wirkungsgeschichte*), that is the awareness of that particular relation between past and present in which past tradition is constitutive of present orientation. Understanding occurs as a hermeneutic fusion of horizons (German *Horizontverschmelzung*):

> 'a hermeneutic situation is determined by the prejudices that we bring with us. They constitute, then the horizon of a particular present, for they represent that beyond which it is impossible to see [...] In fact the horizon of the present is continually in the process of being formed because we are continually having to test all our prejudices. An important part of this testing occurs in encountering the past and in understanding the tradition from which we come [...] understanding is always the fusion of those horizons supposedly existing by themselves' (Gadamer 2013, p. 316 f.).

Understanding simulation validation thus requires on the part of the interpreter an awareness of the prejudice and of the tradition on which the validation concepts are based. I claim that this awareness varies with the interpreter's role and discipline.

9.4.2 Hermeneutic Naiveté Versus Hermeneutic Consciousness

Following Markus (1987, p. 9) I claim that the simulating scientist shows a hermeneutic naiveté vis-à-vis her validation concepts, methods, procedures and related practices. She is engaged with the validation of her model. However, she lacks the hermeneutic 'self-consciousness' (Markus 1987, p. 9) that is typical of many social scientists. As Markus would put it, simulation validation works on the basis of an ideology 'which regards any acceptable scientific text as totally self-sufficient as to its meaning' (Markus 1987, p. 9).

However, the hermeneutic naiveté of the simulation scientist is overcome in interdisciplinary dialogue. Two other groups of researchers who may be interested in simulation validation share this hermeneutic consciousness: philosophers of science engaging in comparative research on simulation validation, and methodologists conducting research on their respective disciplines' methods. Philosophers of science and methodologists are those researchers who have to develop the hermeneutic consciousness as part of their professional engagement with science.

9.4.3 Interdisciplinary Dialogue

Interdisciplinary dialogue constitutes a situation in which a hermeneutic situation is set up. Here, the simulation scientist loses her hermeneutic naiveté vis-à-vis her validation concepts. She perceives validation concepts different from her discipline's tradition, which makes her increasingly aware of the historical situatedness of her own validation concepts.

Three approaches may be followed in empirical interdisciplinary dialogues in general, and apply interdisciplinary exchange about simulation validation more specifically: (1) Initially, interdisciplinary dialogue often amounts to the projection of one's self onto the other. In this case, validation concepts of one's own discipline are projected onto the other discipline. Whatever this approach yields, it is not interdisciplinary understanding. (2) Others emphasize disciplinary alterity, trying to resist the impulse to subsume other disciplines under their methodological tradition. During the dialogue, they correct their view of other disciplines. Even if this approach is intended as a guide to the beginning of the conversation between disciplines, it falls short of interdisciplinary understanding in a philosophical sense. (3) The third approach consists in comparing specific concepts, e.g. validity concepts, across disciplines. This approach focuses on specific items which bear both similarities and differences to a researcher's discipline. For instance, the concept of validity has a range of meanings in physics, and the question then becomes how it is used in another discipline.

Versions of these approaches to interdisciplinary dialogue exist in simulation validation, but hermeneutics has another perspective. Hermeneutic understanding requires us to look at our prejudice and uncover the misunderstandings that we bring with us. Understanding involves as much an engagement with one's own discipline and the situatedness of its validation concepts as it is about that discipline which is being understood.

In an interdisciplinary dialogue, hermeneutics can be understood in at least two ways: (1) as a means for understanding elements of another discipline's simulation validation concepts, in order to uncover its standing in relation to that discipline's tradition and (2) as a means for understanding elements of a discipline's own simulation validation concepts, in order to uncover its own standing in relation to its own tradition. Thus, the 'other' that is encountered may be another discipline's tradition or it may be one's own discipline's tradition or history.

Philosophical hermeneutics allows a self-critical and self-constitutive encounter with alterity as embedded in validity concepts in diverse academic disciplines. There cannot be a universal understanding because all understanding depends on prejudice and tradition. The hermeneutic perspective preserves pluralism in simulation validation: it helps establish the awareness that there is (1) no universal strategy for validation and (2) no single interpretation of a specific tradition of validation. Understanding simulation validation occurs as a hermeneutic 'fusion of horizons' in which the interpreter's horizon is enlarged and enriched. Hermeneutics' perspective is opposed to those perspectives that seek universally shared scientific values in simulation validation.

Considering the relevance of tacit knowledge (see Markus above), I claim that an adequate understanding of simulation validation texts cannot be acquired in an intercourse with the text alone. I therefore propose a more sociological reading of interdisciplinary dialogue that makes accessible the tacit knowledge. Rather than in philosophical dialogue, a social encounter will disclose the tacit knowledge and promote our understanding of simulation validation.

9.4.4 The Hermeneutic Tasks

Hermeneutic aims in interdisciplinary dialogue are showing how simulation validation is historically situated, revealing the hidden prejudice in validating and distinguishing between legitimate prejudice and prejudice that has to be overcome.

9.4.4.1 The Historical Task

The first hermeneutic aim in theorizing simulation validation is to show how simulation validation is historically situated. At present, there is no one unique definition of simulation validation. However, Schlesinger (1979) definition of simulation validation serves as a major reference for many simulating scientists who discuss the question of how to define simulation validation. I will use their definition to illustrate the task. In Schlesinger et al.'s definition, the computerized model, the domain of applicability, the intended application of the model and the scientific value they refer to—accuracy—are historically situated. I concentrate on the most obvious aspect here: why do they refer to accuracy? Compare Schlesinger et al.'s definition to that by Caldwell and Morrison ('Validation is a proactive, diagnostic effort to ensure that the model's results are reasonable and credible' and 'to assess whether the model's outputs are reasonable for their intended purposes', Caldwell and Morrison 2000, pp. 202 f.). The relevance of the scientific value of accuracy is historically contingent. The relevance of scientific values is also situated and may depend on the discipline. The task is to show how the computerized model, the domain of applicability, the intended application of the model and the scientific values are historically situated. Without going into too much detail here, we may anticipate that an interpreter will understand the use of the concept of accuracy after it has been revealed that Schlesinger et al.'s background is in engineering and the natural sciences, where the present state of the art allows for quantitative evaluations of validity. It can be understood, then, that they somehow forgot other disciplines in which the state of the art does not allow for the meaningful use of measures of accuracy. Economists Caldwell and Morrison, instead, face a state of their art in which only qualitative evaluations of the validity of their microsimulation model can be given. In particular, there are no true experimental data against which to validate the predictions of their simulation model. It can be understood, then, that in their definition they refer to more open concepts, such as reasonableness and credibility.

9.4.4.2 The Epistemological Tasks

Gadamer's formulation of the hermeneutic circle (Gadamer 2013, p. 279) claims that the foreknowledge of an interpreter creates expectations in regard to a certain interpretation. In the hermeneutic situation, prejudice (prejudgements) undeniably structure the human understanding of the self and the world. The foreknowledge is the condition of the possibility of understanding. Gadamer rejects the negative connotation of prejudice, which he views as a prejudice of the Enlightenment (Gadamer 2013, p. 283). The epistemological task in the hermeneutics *of* validation is thus to reveal the hidden prejudice in validating. As Gadamer (2013, p. 310) has argued, this is a question for effective history (*Wirkungsgeschichte*)—that particular relation between past and present in which past tradition is constitutive of present orientation. The fundamental epistemological task is to distinguish between legitimate prejudice and prejudice that has to be overcome. Gadamer rehabilitates authority and tradition because they can be a source of legitimate prejudice (which has led his critics to argue that he is a conservative).

The foreknowledge of simulation validation includes diverse kinds of foreknowledge and prejudice: in particular, foreknowledge of the theoretical concepts of the phenomenon that is modelled, foreknowledge of the implemented theory and hypotheses, foreknowledge of the validation methods and techniques applied, foreknowledge of the domain of applicability, foreknowledge of the intended application of the model and foreknowledge of scientific values that are considered relevant, e.g. the value of accuracy.

A short illustrative example comes from climate science. As Rood explains (in Chap. 30 in this volume), 'an influential paper' by Oreskes, Shrader-Frechette, and Belitz (1994) established the formal argument that, in general, numerical models of geophysical phenomena cannot be validated. He summarizes that the argument was twofold. (1) 'The climate' cannot be observed in its entirety and (2) models are non-unique estimates of possible climate states. As Rood notes, 'the echoing of the statement that weather and climate models 'cannot be validated' does not serve the discipline well'. According to Rood, it has also contributed to a stable foundation of political argumentation that model-based predictions are too uncertain on which to base policy.

Let C-Not denote the claim that weather and climate models cannot be validated. From our hermeneutic perspective, claim C-Not shall be a starting point for a hermeneutic analysis. C-Not will be considered as foreknowledge, it has served as a prejudice in climate simulation validation. The hermeneutic analysis would have to reconstruct in detail how, on the one hand, C-Not led to great caution among climate scientists—who tended to distrust the term 'validation' and prefer to use expressions such as 'evaluation' (see Flato et al. 2013, as well as the study by Guillemont, 2010); on the other hand, the hermeneutic analysis would have to reveal how tremendous efforts in testing and validation (see the chapter by Rood) were forced by C-Not. Rood states that a 'culture of verification and validation' has been developed by climate scientists and software engineers. Empirical studies have identified different 'epistemic lifestyles' (Shackley 2001) that include verification and validation. We will finally not only understand the present orientation in climate science simulation evaluation, but also question claim C-Not. Some of the arguments that were used to support C-Not may be revealed as prejudice that can be overcome. Finally, even C-Not may be overcome.

I have also used this short illustrative example to indicate that Gadamer's concept of effective history (*Wirkungsgeschichte*) need not refer to the distant past. It may also apply to significant events of past decades.

9.4.4.3 The Hermeneutic Tasks of Three Groups of Interpreters

Philosophers of science, and—to a lesser degree—methodologists of their disciplines, can be expected to deal with the historical and epistemological hermeneutic tasks. However, working scientists are needed, not only to explore the tacit knowledge but also to interpret and change validation practices in the light of the new insights obtained. Philosophers of science and methodologists open up the validation tradition or traditions vis-à-vis the simulating scientists and support the establishment of a hermeneutic 'self-consciousness' among the practitioners. In this way, the hermeneutics *of* validation addresses the first-order art and second-order theory.

9.5 Discussion

What are the limitations to a hermeneutics *of* validation based on Gadamer's philosophical hermeneutics? These limitations become visible if we apply the criticism of Gadamer's approach with representatives of other philosophical perspectives.

Some limitations follow from the Gadamer–Habermas debate on the issues of rational reflection and material reality (for a recent review on that debate and its outcomes see Smith 2014). Habermas argued that Gadamer's hermeneutics leaves no room for genuinely rational reflection, since it is constitutively blind to potential sources of domination. These sources are embedded in hermeneutic reflection: tradition, authority and prejudice. As a consequence, argues Habermas, hermeneutic reflection must fall short as a model of critical reflection. Tradition, authority and prejudice are accountable to standards that lie beyond them—to rational standards. Some results of that debate demarcate limitations of the hermeneutics *of* validation, in particular for interdisciplinary dialogue. There is a tension between rational reflection and understanding, and this tension is important for the growth of knowledge, in particular for the role of traditions in that area. As Gadamer (1990) has pointed out, hermeneutic reflection at its best has a self-transformative character. Traditions advance through self-correction. Ultimately, hermeneutic reflection rests on practical insight. I contend that disciplinary traditions are challenged less by a Gadamerian dialogue than by a Habermasian discourse. Habermas (1984, 1996) discourse model establishes a stringent set of rules known as ideal speech situation to support the deliberation on, and the analysis and justification of, validity claims. Communicative rationality is more confrontative towards traditions, enhancing the growth of knowledge in simulation validation.

Markus' claim (1987, see above) that terminology and methods in the sciences are subject to an accelerated rate of obsolescence, rendering the use of Gadamer's concept of tradition shallow, raises the question as to whether there are traditions in simulation validation at all. This volume demonstrates that such traditions exist, providing the opportunity for further studies to reconstruct and describe them in more detail. From psychology (see e.g. Newton and Shaw 2014), we know thorough investigations into a discipline's validation traditions.

What is the significance of the hermeneutics of simulation validation with respect to the further development of computer simulation? The question of validation is an urgent one since computer simulations are developed in more and more disciplines. The hermeneutic perspective is oriented towards past and *present*. Interdisciplinary dialogue can advance the spread of validation methods and techniques across disciplines. It is no coincidence that the author of this chapter—a simulating sociologist—who is also one of two editors of this *Volume* wanted to edit this compendium. This *Volume* is a major step towards an interdisciplinary dialogue on simulation validation. Interdisciplinary dialogue may be evaluated as suitable for the late adopters, but not for the leading disciplines such as meteorology and climate science. However, these latter disciplines also profit from encountering their past.

9.6 Conclusions

In this chapter, a hermeneutics *in* simulation validation has been shown to be rather fruitless. Instead, I have proposed a hermeneutics *of* simulation validation based on Gadamer's philosophical hermeneutics.

The goal of a hermeneutics *of* validation is to understand simulation validation. Its contribution to the validation of computer simulation models is on two levels: first-order art and second-order theory. The challenge that has to be mastered is to set up a hermeneutic situation in the first place. As Ramberg and Gjesdal (2005) emphasize, appreciating hermeneutics is fundamentally a matter of perceiving a moving horizon, engaging a strand of dialogue. This *Volume* establishes such a hermeneutic situation.

Finally, I want to suggest one issue for the interdisciplinary dialogue. We should understand the preference for particular scientific values in different validation simulation traditions. As is shown by Schlesinger's definition, accuracy is presently the dominating scientific value in simulation validation. However, there is also, for instance, the value of comprehensiveness (see Chap. 40 by Hirsch Hadorn and Baumberger in this volume), apparently most often inferior. Let us discover and test our prejudice concerning accuracy and comprehensiveness and encounter the past and understand the validation traditions from which we come. Doing so will allow us to expand and enrich our horizons in relation to the prioritization of scientific values in simulation validation.

Acknowledgements The author thanks Claus Beisbart for helpful discussions concerning this manuscript.

References

Arendt, H. (1958). *The human condition*. Chicago: University of Chicago Press.

Arendt, H. (1969). *Crisis of the republic*. New York: Harcourt Brace Jovanovich.

Barlas, Y., & Carpenter, S. (1990). Philosophical roots of model validation. Two paradigms. *System Dynamics Review, 6*, 148–166.

Bernstein, R. J. (1983). *Beyond objectivism and relativism: Science, hermeneutics, and praxis*. Philadelphia, PA: University of Pennsylvania Press.

Caldwell, S., & Morrison, R. J. (2000). Validation of longitudinal dynamic microsimulation models. Experience with CORSIM and DYNACAN. In: L. Mitton, H. Sutherland, & M. J. Weeks (Eds.), *Microsimulation modelling for policy analysis. Challenges and innovations* (pp. 200–225). Cambridge: Cambridge University Press.

Carson, J. S. (1989). Verification and validation. A consultant's perspective. In: E. A. MacNair, K. J. Musselman, & P. Heidelberger (Eds.), *Proceedings 1989 Winter Simulation Conference* (pp. 552–557).

Crease, R. P. (1997). Hermeneutics and the natural science: Introduction. *Man and World, 30,* 259–270.

D'Agostino, F. (2014). Hermeneutics, epistemology, and science. In: J. Malpras & H. -H. Gander (Eds.), *The Routledge companion to hermeneutics* (pp. 417–428). London: Routledge.

Doublet, D. R. (2003). Der Hermeneutische Zirkel: Über Grenzen für die Interpretation und Bedingungen für das Verstehen. In S. U. Larsen & E. Zimmermann (Eds.), *Theorien und Methoden in den Sozialwissenschaften* (pp. 61–75). Wiesbaden: Springer VS.

Eger, M. (1997). Achievement of the hermeneutic-phenomenological approach to natural science. A comparison with constructivist sociology. *Man and World, 30,* 343–367.

Feher, M., Kiss, O., & Ropolyi, L. (Eds.). (1999). *Hermeneutics and science*. Dordrecht: Reidel.

Feinstein, A. H., & Cannon, H. M. (2003). A hermeneutical approach to external validation of simulation models. *Simulation & Gaming, 34,* 186–197.

Feyerabend, P. (1975). *Against method. Outline of an anarchistic theory of knowledge*. London: New Left Books.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., & Rummukainen, M. (2013). Evaluation of climate models. In: Stocker, T. F., D. Qin, G. -K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, & P. M. Midgley (Eds.), *Climate change 2013: The physical science basis*. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Forrester, J. W. (1961). *Industrial dynamics*. Cambridge, MA: MIT Press.

Gadamer, H. -G. (1990). Reply to my critics. In: D. Ormiston & A. Schrift (Eds.), *The hermeneutic tradition* (pp. 273–297). Albany, NY: Suny Press.

Gadamer, H. -G. (2013). *Truth and method*. Rev. 2nd edn. Trans. by J. Weinsheimer & D. G. Marshall. London: Bloomsbury. [1st German edition: 1960. Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik. Tübingen: Mohr].

Guillemot, H. (2010). Connections between simulations and observation in climate computer modeling. Scientist's practices and "bottom-up epistemology" lessons. *Studies in History and Philosophy of Modern Physics, 41,* 242–252.

Habermas, J. (1984). *The theory of communicative action. Vol. 1. Reason and the rationalization of society*. Trans. and introduced by T. McCarthy. Boston, MA: Beacon Press.

Habermas, J. (1996). *Between facts and norms. Contributions to a discourse theory of law and democracy*. Trans. and introduced by W. Rehg. Cambridge, UK: Polity Press.

Heelan, P. A. (1998). The scope of hermeneutics in natural science. *Studies in History and Philosophy of Science, 29,* 273–298.

Humphreys, P. (2004). *Extending ourselves. Computational science, empiricism, and scientific method*. Oxford: Oxford University Press.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169,* 615–626.

Kisiel, T. (1997). A hermeneutics of the natural sciences? The debate updated. *Man and World, 30,* 329–341.

Kleindorfer, G. B., & Geneshan, R. (1993). The philosophy of science and validation in simulation. In: *Proceedings of the 25th Conference on Winter Simulation (WSC 1993)*, New York, NY, USA (pp. 50–57).

Kleindorfer, G. B., O'Neill, L., & Ganeshan, R. (1998). Validation in simulation: Various positions in the philosophy of science. *Management Science, 44,* 1087–1099.

Kuhn, T. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Malpas, J., & Gander, H.-H. (Eds.). (2014). *The Routledge companion to hermeneutics*. London and New York: Routledge.

Markus, G. (1987). Why is there no hermeneutics of the natural sciences? Some preliminary theses. *Science in Context, 1,* 5–51.

Naylor, T. H., & Finger, J. M. (1967). Verification of computer simulation models. *Management Science, 14,* B92–B101.

Newton, P. E., & S. D. Shaw (2014). *Validity in educational and psychological assessment*. Sage Publications Ltd.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science, 263,* 641–646.

Poscher, R. (2014). Hermeneutics, jurisprudence and law. In J. Malpas & H.-H. Gander (Eds.), *The Routledge companion to hermeneutics* (pp. 451–465). London and New York: Routledge.

Reutlinger, A., Hagleiter, D., & Hartmann, S. (2018). Understanding (with) toy models. *The British Journal for the Philosophy of Science, 69*, 1069–1099.

Ramberg, B., & Gjesdal, K. (2005). Hermeneutics. In: E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (2005 ed.). https://plato.stanford.edu/archives/win2014/entries/hermeneutics/#Pragmatism.

Saam, N. J. (2017). Understanding social science simulations. Distinguishing two categories of simulations. In: M. Resch, A. Kaminski, & P. Gehring (Eds.), *The science and art of simulation I. Exploring - Understanding - Knowing* (pp. 67–84). Cham: Springer.

Saam, N. J., & Schmidl, A. (2018). 'A Distinct Element of Play.' Scientific computer simulation as playful investigating. In: A. Friedrich, P. Gehring, C. Hubig, A. Kaminski, & A. Nordmann (Eds.), *Arbeit und Spiel. Jahrbuch Technikphilosophie 2018* (pp. 99–118). Baden-Baden: Nomos.

Schlesinger, S. (1979). Terminology for model credibility. *Simulation, 32,* 103–104.

Shackley, S. (2001). Epistemic lifestyles in climate change modeling. In C. A. Miller & P. N. Edwards (Eds.), *Changing the atmosphere: Expert knowledge and environmental governance* (pp. 107–133). Cambridge, MA, USA: The MIT Press.

Smith, N. H. (2014). Hermeneutics and critical theory. In: J. Malpas & H. -H. Gander (Eds.), *The Routledge companion to hermeneutics* (pp. 600–611). London and New York: Routledge.

Steinmann, M. (2007). Auf dem Weg zu einer modernen Epistemologie. In G. Figal (Ed.), *Hans-Georg Gadamer: Wahrheit und Methode* (pp. 87–103). Berlin: Akademie Verlag.

Part III
Methodology—Preparatory Steps

Chapter 10
Assessing the Credibility of Conceptual Models

Axel Gelfert

Abstract Whether or not the results of a computer simulation are credible depends to a large extent on the credibility (or lack thereof) of the underlying conceptual model. If a model has been developed explicitly with the goal of running a computer simulation in mind, the two types of credibility may seem deeply intertwined. Yet, often enough, conceptual models predate the subsequent development of simulation techniques, or were first developed outside the context of computer simulation. In such a situation, the specific contribution that a conceptual model makes to the credibility of a simulation requires considerable analysis. How, then, should we assess the credibility of a conceptual model, and which factors ought to play a role in judging whether simulation results derived on its basis are trustworthy? In order to answer these questions, the present chapter begins with the premise that models are never by themselves credible *simpliciter*, but acquire credibility within a given context of inquiry, which itself depends on the cognitive interests of the inquirer. Judgments concerning the credibility of a conceptual model thus need to be based partly on a characterization of the intrinsic features of the model, partly on the cognitive goals and interests of its users. This realization helps explain why credible models have been variously understood as (pragmatically and empirically) adequate representations of real-world target systems, as constructions of "credible worlds" that display internal coherence, and as exploratory tools that may aid our understanding even before a well-developed underlying theory takes shape.

Keywords Credibility · Scientific modeling · Conceptual model · Empirical fit · Causal understanding

A. Gelfert (✉)
Technische Universität Berlin, Straße des 17. Juni 135,
FG Theoretische Philosophie H72, 10623 Berlin, Germany
e-mail: a.gelfert@tu-berlin.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_10

10.1 Introduction

Contemporary science would not be what it is, were it not for the emergence of computer simulation techniques from the mid-twentieth century onwards. First pioneered in such disciplines as nuclear physics and meteorology, numerical methods for implementing computer simulations have since spread to a wide range of scientific disciplines, including astrophysics, high-energy physics, materials science, quantum chemistry, biochemistry, molecular biology, ecology, climate science, economics, sociodynamics, and many others. This applies equally to fundamental and applied research, and also extends to neighboring disciplines such as design and engineering. As things stand, much of our best—and instrumentally most important—scientific knowledge is best thought of as being *simulation-based*.

Typically, discussions of computer simulation go hand in hand with discussions of scientific models, and for good reason: modeling and computer simulation both are modes of inquiry that scientists engage in when the behavior of a target system cannot easily be derived from an underlying theory—either because this task is too complex (and, for example, does not allow for an analytical solution) or because no underlying theory can be unequivocally specified (e.g., since the phenomenon in question is the result of a heterogeneous mix of factors). In order to construct a model and implement any simulation, decisions need to be made—however implicitly—about how to represent the target system in the real world that is the subject of the proposed simulation study. Such decisions concern, amongst others, the level of detail, overall structure, relevant factors, and purported causal mechanisms of the target system that is to be represented and simulated. Constructing a model of this sort will not only guide future inquiry, but will also shape how we should interpret any subsequent results; in this sense, specifying a model determines—at least in broad, qualitative terms—the very content of a simulation study.

This suggests that whether or not the results of a simulation study are credible also hinges on the credibility, or lack thereof, of the underlying model, which is often called the conceptual model. Crudely speaking, and notwithstanding the specific insights that only computer simulation methods can afford, a simulation can only ever be as good as the conceptual model on which it is based. It seems legitimate, then, to spend some time reflecting on what makes models credible in the first place, and it is precisely this question that the present chapter tackles. Yet models are never by themselves credible *simpliciter*, but acquire credibility within a given context of inquiry, which itself depends on the cognitive interests on the inquirer. To the extent that models are credible, then, they are credible *for an inquirer* in a particular problem situation, or so it will be argued. Combining these two perspectives—a focus on the particular model at hand, and a recognition of their significance to the model user—thus leads to the thesis that judgments concerning the credibility of a conceptual model need to be based on a characterization of the intrinsic features of the model as well as of the cognitive goals and interests of its users. This also helps explain why credible models have been variously understood as (pragmatically and empirically) adequate representations of real-world targets, as constructions of "credible worlds" that display internal coherence (and may also serve as an "intuitionpump"), or as exploratory tools that may aid our understanding even before a well-developed theoretical account of a phenomenon has become available.

The remainder of this chapter is organized as follows: Sect. 10.2 reviews the twin notions of verification and validation in connection with computer simulation and relates them to the multiplicity of functions of scientific models more generally. Section 10.3 takes our commonsense understanding of "credibility" (in connection with human interlocutors) as the starting point for a discussion of the qualities required for attributions of credibility to nonhuman agents and entities; Sect. 10.4 then applies these ideas to examples from scientific practice. Sections 10.5 and 10.6, respectively, reflect on specific—sometimes competing—criteria for model credibility: empirical fit and furthering causal understanding on the one hand (Sect. 10.5), and the construction of "credible worlds" along with exploratory uses of models on the other hand (Sect. 10.6). The chapter concludes with a brief reflection on how credibility is jointly constituted by features of the model itself and the overall goals and cognitive interests of its users (Sect. 10.7).

10.2 Simulation-Based Knowledge, Verification, Validity, and the Function of Models

Whether a particular scientific investigation begins with data collection or theoretical analysis, once we turn to computational methods to derive simulation-based knowledge, conceptual models are never very far off. If studying a problem using computer simulation is the explicit goal of a given process of inquiry, then deriving a conceptual model may be understood as an important first step in preparing the problem for the deployment of computer simulation methods. Yet, arguably, many conceptual models in science were developed independently of concerns with computer simulation—either because they predate the development of powerful simulation techniques or because they initially served independent illustrative, descriptive, or explanatory purposes and were only later found to lend themselves to computer simulation. Tackling a scientific problem using computer modeling is, rightly, often characterized as a multi-step process, with decisions concerning the problem domain preceding the process of designing a simulation model. This model can then be implemented on a given type of computer system. Thus, even before a conceptual model is proposed, "a description of the problem situation and the system in which the problem situation resides" (Robinson 2011, p. 1432) must be given. Where an underlying fundamental theory can reasonably be assumed to exist (e.g., in the case of planetary motion), this may involve specifying the relevant background assumptions (e.g., considering only objects on closed orbits and looking into their relative position to one another); where a fundamental theory is absent (e.g., in the case of modeling vehicular traffic flow), this may involve selecting well-delineated research questions ("How do spontaneous traffic jams occur?") and specifying the variables that are thought to best describe the phenomenon in question (say, measurable changes in average vehicular speed).

In recent years there has been a growing recognition of the heterogeneous character of model building in science, where models are constructed as representations to investigate a target system, not merely conceived of as realizations of theoretical relationships that are posited as true. On this view, not all models are "derived" from fundamental theory—not only because many research questions require models to be "made up from a mixture of elements, including those from outside the domain of investigation" (Morrison and Morgan 1999, p. 23), but also because models often represent phenomena that have yet to be subsumed under anything resembling a self-contained underlying theory. In the absence of a theory of the domain of investigation, models may thus serve an exploratory function (Gelfert 2016): it is by constructing models that scientists attempt to find out whether a purported phenomenon really does survive closer scrutiny and try to devise "proto-theories" whose relation to more fundamental theories then needs to be analyzed further. It is only once a decision has been made that a particular problem situation can, at least tentatively, be characterized using a certain set of representational tools that a *conceptual model* can be developed. Or, to put it another way, any conceptual model that is being advanced—be it in the form of a set of mathematical equations with corresponding physical interpretations, or as a less formalized set of assumptions, simplifications, and explanatory mechanisms—will already implicitly include *some* assumptions about which sorts of problem situations and approaches are appropriate. It is important to keep this in mind since, as we shall see, a conceptual model's *credibility* partly depends on it.

If we were only interested in constructing a model to represent a target system or phenomenon, arriving at a viable conceptual model might be considered a fitting conclusion to the process of scientific modeling. And indeed, a fair amount of philosophical writing on scientific models takes the construction of models to be the goal and outcome of scientific inquiry—even as it acknowledges that scientific models serve a variety of epistemic and non-epistemic functions. In this sense, what simulationists call the "conceptual model" is generally referred to by philosophers of science as "scientific model" *simpliciter*. Yet, the work of the simulationist does not stop with devising a conceptual model; instead, it is taken as a starting point for the further processes of designing a numerical model in computer code, and implementing the numerical model on a computer model. From the simulationist's viewpoint, the conceptual model is "a non-software specific description of the computer simulation model (that will be, is or has been developed), describing the objectives, inputs, outputs, content, assumptions and simplifications of the model" (Robinson 2008, p. 283). Seen in this light, the construction of a conceptual model by a simulationist is teleologically oriented toward the creation of a piece of software, the computer simulation model, in order to answer specific hypotheses about the target system by running a simulation. It would be hasty, however, to infer from this that the conceptual model play s a merely auxiliary role. Simulationists are keenly aware that many conceptual models enjoy independent support and justification, and that the success of implementing a simulation model partly depends on adequately translating the conceptual model into computer code.

The relationship between a (numerically implemented) simulation model and the (underlying) conceptual model is a nontrivial one, as becomes evident once we consider what we can infer—about the simulation and the conceptual model, respectively—from the empirical success (or lack thereof) of the numerical output thus generated. Even before we compare the numerical output against our observations or measurements, we may ask whether the simulation model does, indeed, adequately reflect the conceptual model; that is, we may engage in *verification* (see Chap. 11 by Rider and Chap. 12 by Roache in this volume). On the one hand, this involves "determining that a simulation computer program performs as intended, i.e., debugging the computer program" (Law and Kelton 1991, p. 299); in a departure from established usage in philosophy of science, the term "verification," thus understood, does not refer to the process of generating observable predictions and testing them empirically. The guiding idea behind verification in this sense is not *empirical testing*, but *formal demonstration* in the spirit of logic and mathematics: "Purely formal structures are verifiable because they can be proved by symbolic manipulation, and the meaning of these symbols is fixed and not contingent on empirically based input parameters" (Oreskes et al. 1994, p. 641). What is being verified, then, is not that the model is a successful representation of the target system, nor that the numerical output matches empirical observations, but rather that an already constructed conceptual model is correctly solved in the software code that constitutes the (implemented) simulation model. In this sense, verification (also called "technical validation") aims to demonstrate internal consistency, sometimes by way of benchmarking numerical results against analytical solutions (where available). On the other hand, this usage of "verification" involves the problematic assumption that formally verifying that a computer simulation approximately solves a set of underlying mathematical equations answers all relevant questions that may arise in the course of computationally implementing a conceptual model. For example, there is considerable latitude in how one should discretize the underlying equations in a simulation model, and each such choice may have advantages and drawbacks, yet these are not a matter of meeting (or failing) certain formal standards. It is, therefore, important to realize that verification is itself typically part of an iterative process of implementing and subsequently tweaking a simulation model and its implementation. As Eric Winsberg puts it, "there can be no justification of the final [conceptual] model that is independent of its discretized implementation, and there can be no justification of the implementation that is independent of the model" (Winsberg 2018, p. 158).

Turning to the second element in the often jointly used phrase "verification and validation," *validation* aims to ascertain the simulation model's performance across a range of empirical contexts, for example by simulating a real data source and comparing the calculated outcomes with real-world observations. (See Chap. 4 by Murray-Smith in this volume.) This can be a formidable task, given that "directly making the validity assessments requires technical expertise and full access to the model and external data" (Caro et al. 2014, p. 178). The ability to predict, or otherwise reproduce, empirical aspects of the behavior of the target system is key to a simulation model's *external validity*, where this refers to the generalizability of the findings of a simulation to the intended class of real-world cases. Different strategies can be pursued: at minimum, an implemented computer simulation should be able to adequately reproduce data sources that went into the creation of the model in the first place ("dependent validation"), though generally, it will be preferable to test a simulation's performance with respect to independent data sets—that is, data that is of a type that the simulation should be able to account for, but which was not actively utilized in the process of simulation design. As in the case of testing theories, the predictive capacity of simulation models, too—that is, its ability to predict empirical results before they have been measured or observed—is often seen as carrying great weight ("predictive validation"). At the same time, drawing too close a parallel between validation of a simulation model and theory-testing, can be misleading. Scientists themselves have occasionally emphasized that "[v]alidation is not a procedure for testing scientific theory or for certifying the "truth" of current scientific understanding, nor is it a required activity of every modelling project" (Rykiel 1996, p. 299), and have lamented the widespread belief "that validation establishes the veracity of the model" (Oreskes et al. 1994, p. 642).

To be sure, textbook definitions of "validation" in computer simulation studies often equate it with "determining whether a simulation model (as opposed to the computer program) is an accurate representation of the system" (Law and Kelton 2000, p. 265), sometimes with the caveat that a model should be "an accurate representation of the real world *from the perspective of the intended uses of the model*" (ITT, cited after Zacharias et al. 2008, p. 302). Critics of such definitions have argued that "[t]he implication is that validated models tell us how the world really is", when we should always keep in mind that any agreement between observed measurements and simulation results "in no way demonstrates that the model that produced the output is an accurate representation of the real system" (Oreskes et al. 1994, p. 642). It seems fair to assume, however, that the authors quoted above intended their textbook definitions to be understood elliptically, in that they would not deny that further assumptions and inferences are required to warrant moving from empirical validation of the simulation results to representational success of the underlying model. Perhaps, then, validation is best understood as both relative to the context of inquiry and the goal of the inquirer:

> Validation is a demonstration that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the *intended application of the model.* (Rykiel 1996, p. 233; italics original.)

What constitutes a "satisfactory range" depends, at least in part, on the applicable standards of empirical performance which, again, vary between users, depending on their goals, and across different contexts: "That is, a model is declared validated within a specific context which is an integral part of the certification. If the context changes, the model must be re-validated." (ibid.)

The recognition that context matters stems from the realization that models and simulation may serve a wide range of purposes, from promoting epistemic goals (e.g., affording insight into the causal basis of a particular phenomenon) to non-epistemic objectives (e.g., serving as the basis for policy decisions). It also reflects the fact—well-known to scientific practitioners, but perhaps underappreciated in philosophy of science—that the process of modeling and simulation does not, in practice, divide up neatly into distinct "phases" of (1) constructing a conceptual model, (2) translating it into computer code and verifying that the code correctly implements it, and (3) validating the numerical output by comparison with empirical observations. Instead, what one finds—often, not always—is that the various phases overlap and are deeply intertwined. For example, it is not uncommon for approximations that were initially employed during the stage of implementation and verification to become an essential part of the simulation model as a whole, and for them to be credited with ensuring its overall empirical success (and, thus, its validity).[1]

Few practitioners of computer simulation studies would consider any given validation of a simulation model to be sacrosanct. Not only is it the case that, as mentioned earlier, models that have been validated within one context for a given purpose may need to be re-validated for a different use in another context. Like models and simulations in general, the practice of validation, too, serves different (if often complementary) purposes. Sometimes, when the mutual relationships between the numerical output, the simulation model, the conceptual model, and the underlying theory are well-understood, the successful validation of a computer simulation may indeed be interpreted as (good, but defeasible) evidence that the underlying model adequately represents the way the world is. Often, however, the route from a scientific problem to a computer simulation model is less clear-cut, and the various relationships between models, theories, and simulations are contested. In such a situation, validation may also function as a tool "for building model credibility in the user community" (Rykiel 1996, p. 230). This is not to say that validation is being carried out strategically, let alone with manipulative intent to persuade others; rather, validation ensures that a computer simulation adheres to shared standards of accuracy in the given context of inquiry.

An excessive focus on verification and validation may, on occasion, obscure other functions of models and simulations, beyond their ability to reproduce observed phenomena. As the ecologist Edward Rykiel puts it, "modelling and the benefits to be gained from it can also be stifled by an overemphasis on model validation" (1996, p. 240). Such benefits include, but are not limited to, exploratory uses of models in the absence of a fully formed theoretical framework, which have recently begun to receive philosophical attention.[2] Indeed, "[e]xploration of model behavior without validation testing is a legitimate, reportable activity" (Rykiel 1996, p. 241). This—along with the realization that, even where validation of a simulation model is possible, not much can be inferred with certainty about whether or not "a model accurately represents the 'actual processes occurring in a real system'" (Oreskes et al. 1994, p. 642)—strongly suggests that, in relying on conceptual models in our simulation design, we implicitly presuppose that those models enjoy independent justification. For this reason, and because models and simulations enjoy considerable autonomy from one another, the rest of this chapter will focus on the diverse sources of credibility of the conceptual model.

[1] For a fascinating case study, see Lenhard (2007).

[2] For a detailed argument that exploration is a core function of scientific modeling, see Gelfert (2016).

10.3 Taking the Notion of "Credibility" Seriously

If we are to gain a deeper understanding—beyond technical measures of fit, distance, fractional variance, etc., (which, in any case, can only be determined post hoc)—of what leads scientists to trust some conceptual models more than others, then it may be best to begin by taking the notion of *credibility* seriously. At the risk of appearing overly literal-minded, in the present section I wish to discuss some of the relevant connotations of "credibility" as a general concept; in the next two sections, we shall then encounter a range of examples of how scientists tend to arrive at judgments of credibility in relation to scientific models.

On the face of it, it may seem puzzling why, in the seemingly neutral context of model evaluation, one should invoke the term "credibility" at all, given its ethical and interpersonal overtones. Why not stick with more objective criteria such as "reliability" or "verisimilitude"? The concept of *credibility* has its natural place in human communication and is of a piece with—albeit slightly less emotionally tinged than—the concept of *trustworthiness*. (Cf. Chap. 17 by Saam in this volume for the role of humans in validation.) Yet it is not by chance that scientists should turn to the notion of credibility in their interactions with scientific models and simulations, or so I wish to suggest. In interpersonal communication, credibility is usually considered to be a function of both the trustworthiness and competence of an agent. To be sure, we also speak of isolated claims as being "credible" if we think they merit belief. However, when it comes to the credibility of models, the closer analogy is with epistemic agents, not with the level of individual propositions. To put it another way, when we believe a model-based prediction, we typically do so because we consider the model to be credible, not because we have independent reason to think that the specific prediction in question is somehow *ex ante* more likely to be true than any of its close competitors. A credible model, then, is one that we can turn to, with some confidence, for answers on a suitably wide range of relevant research questions. Once we consider a model credible, and resolve to work with it for the purposes of inquiry, we begin to trust its results—not blindly, of course, but in a way that grants its results some measure of (defeasible) default justification.

In human interactions, once we trust someone, we depend, at least in part, on their goodwill. When I ask you for directions to the train station and trust your answer, the success of my subsequent actions depends, among other things, on your having chosen not to play a prank on me and send me in the wrong direction. What I end up believing depends, in part, on your mental processes. This is why, for someone to be deemed credible, they must not only be deemed to be competent with respect to the subject matter in question, but must also be trustworthy (that is, honest and sincere). (See e.g., McGinnies and Ward 1980.) In the case of models, while there is no analogy to the involvement of another mind in our own process of belief formation, we likewise depend on factors internal to the model which cannot readily be inspected. Competence and trustworthiness are not categories that can be directly applied to scientific models, but it is not difficult to identify desiderata that are structurally similar. Like a competent interlocutor, a good model should be able to provide reliable information regarding a broad range of thematically related questions; similarly, in much the same way that a trustworthy interlocutor would not suddenly start offering wildly misleading claims, a good model should not exhibit sudden discontinuities in the quality of information we can extract from it.

It is perhaps no surprise that the very term "inquiry" is ambiguous, inasmuch as it refers both to objective empirical investigation and to interpersonal requests for information. Indeed, in the Baconian tradition of experimental natural philosophy that was at the heart of the Scientific Revolution, scientific inquiry was often likened to the process of *interrogation*—sometimes by violent means, as critics of the metaphor have pointed out. (See Merchant 1980.) Scientific experimentation itself was seen as a method of bringing about conditions that allowed for the extraction of truth—a way of "putting Nature to the question" (where this, of course, was a common euphemism for torturing someone at the rack). What matters for present purposes is not the problematic character of Bacon's imagery, but rather the transactional conception of inquiry as *interrogation*. But surely, one might wonder, there is a difference between experimenting on nature by bringing about material conditions that may elicit novel observable phenomena, and "interrogating" a conceptual model by various means of analysis? Indeed, there is; yet, as Joseph Pitt notes (in a passage concerned with scientific theories, but in a way that naturally extends to models), there are also striking parallels:

> When a scientist works with a theory to derive some results, she is in some sort of communication with it. She knows that if she does this she will get, or at least, ought to get this result. It is in her being able to anticipate the response of the theory to her manipulations that she is communicating with it. (Pitt 2007, p. 55)

Pitt intends this to be more than just a useful metaphor for understanding how we engage with theories. The key notion is "manipulation"—which seems even more apt in the case of models which, as mentioned earlier, are often made up of a heterogeneous mixture of elements, arranged precisely in a way to enable inferences about the target system (and, thereby, make it possible to extract information about the world). As Pitt notes, "to the extent that we manipulate theories [*read:* models—A.G.] we communicate with them": "The key here is in knowing how to communicate and with what kinds of things we communicate." (Pitt 2007, p. 55).

There is clearly a metaphorical element to likening scientific inquiry to verbal interrogation—but this is no more problematic in the case of scientific models than with respect to scientific experimentation, or so I wish to suggest. Scientists continuously labor with scientific models—often the same ones, with only minor variations of the same underlying equations or formalisms—and, over time, come to see them as "mediators, contributors, and enablers of scientific knowledge" (Gelfert 2016, p. 127). When they judge a conceptual model to be *credible*, this is more than merely an interim assessment of a model's utility "here and now," but expresses a commitment to its future use and expected fruitfulness. Judgments of credibility, then, play an important part in the evaluation of conceptual models, and it will be insightful, in the next section, to discuss the standards and criteria deployed by scientists in their assessment of whether a conceptual model deserves our trust.

10.4 The Credibility of Models: Lessons from Scientific Practice

Scientists, in assessing the models they are using, are typically less concerned with reporting overall levels of credibility concerning specific models, let alone subjective judgments of their trustworthiness, but instead—rightly so—tend to acknowledge the hybrid nature of model credibility:

> The credibility of a modeling analysis should be assessed at several levels: validation, design, data, analyses, reporting, interpretation, and conflicts of interests. Validation assesses how well the model accords with reality. The design should follow accepted standards for conceptualizing and framing the model. The data used in building model should be suitable for the purpose, properly analyzed, and incorporated in the model. Analyses should provide the information required to support decision maker. (Caro et al. 2014, p. 178)

"Reporting" and "interpretation", which are being acknowledged by the authors as "not specifically pertaining to a model's credibility" (ibid.), nonetheless are central to the credibility of modeling *as a process* and its application in practical contexts. The final point—conflicts of interest—demonstrates how the overall credibility of a modeling analysis depends both on the credibility of the model (according to the criteria specified in the previous section) and on the trustworthiness of the modeler who, after all, has decided to deploy one (type of) model rather than another for a particular purpose.

If, for the time being, we take actual usage at face value, we find that "credibility" for scientists has an—in the eyes of philosophers perhaps surprisingly—instrumentalist character, with strong social connotations. Earlier, we quoted the ecologist Rykiel as arguing that "validation is not an essential activity for evaluating research models, but is important for building model credibility *in the user community*" (Rykiel 1996, p. 230; italics added). Credibility, in turn, is best defined in terms of the demand that, as Stewart Robinson puts it, a model "[b]e believed by the clients" (Robinson 2011, p. 1433). And Caro et al. (2014), the same group of health scientists who gave the pithy characterization of the hybrid character of model credibility quoted above, have also drawn up a list of questions that may guide assessments of "the relevance and credibility of a modeling study". In addition to obvious concerns regarding verification and validation, these include questions such as the following:

> Does the model have sufficient face validity to make its results credible for your decision? Is the design of the model adequate for your decision problem? Are the data used in populating the model suitable for your decision problem? (Caro et al. 2014, p. 176; format adapted from table)

"Face validity," in particular, is thought to be a first criterion by which to screen out straightforwardly implausible proposals, inasmuch as a model should not contain unrealistic and implausible assumption about core elements that a model is intended to get right.[3] Whether a model meets this desideratum is thought to be "the easiest

[3] This should not be understood as contradicting the frequent—and entirely correct—observation that, as William Wimsatt puts it, false models may function "as means to truer theories" (Wimsatt 2007, p. 94).

aspect of credibility for a user to check because it does not require in-depth technical knowhow"; at the same time, if "parts of the model fail face validity, the effect on credibility depends on the user's judgment about whether the questionable parts are so unrealistic or inappropriate that they affect the accuracy of the results" (Caro et al. 2014, p. 179).

As scientists themselves are keen to point out, mere success at reproducing empirical results does not suffice to render a model credible:

> Agreement between model and data does not imply that the modeling assumptions accurately describe the processes producing the observed climate system behavior; it merely indicates that the model is one (of maybe several) that is plausible, meaning that it is empirically adequate. (Knutti 2008, p. 4652)

Just as we do not consider an interlocutor credible merely in virtue of having made a series of truthful assertions, we do not place trust in a model just because it happens to have successfully reproduced some amount of data. Attributions of credibility derive from the warranted presupposition that a source of information is systematically reliable across a range of relevant contexts and questions. Background knowledge, thus, is a key to assessments of the credibility of models: "The model results we trust most are those that we can understand the best, and relate them to simpler models, conceptual or theoretical frameworks". (Knutti 2008, p. 4656)

If there is one near-universal feature of how scientists talk about the credibility of their models, then it would have to be their recognition that models serve a variety of purposes, such that attributions of credibility depend on the goals of the modeling process. This is not to suggest that attributions of model credibility are subjective, or cannot be challenged, but simply to acknowledge that, for such an attribution to be meaningful (and potentially intersubjectively compelling), the goals and contexts of the modeling process need to be specified. This is precisely why Caro et al. (2014), in their proposed list of diagnostic questions concerning the relevance or credibility of a modeling study, ask researchers to consider what external validation, internal validation, face validity, design aspects, etc., of a model-based study have to contribute toward "mak[ing] its results credible *for your decision*?" (178, italics added) Other scientists, in a similar spirit, note that models are to be assessed by their ability to "[p]roduce sufficiently accurate results for the purpose" that a modeler has chosen (Robinson 2011, p. 1433), and in view of how "acceptable for pragmatic purposes" (Rykiel 1996, p. 230) its results are. This reflects, once more, the largely pragmatic-instrumentalist attitude of practicing scientists regarding the utility of models in specific problem-oriented contexts of inquiry.

Wendy Parker, in the context of discussing precisely what is being confirmed when models—climate models in particular—are found to fit with observations and past data, resists the thought (to be discussed in the next section) that "as we accumulate instances of fit between observational data and output from a climate model, we are accumulating evidence of the truth of the hypothesis embodied by the model" (Parker 2009, p. 234). Such a view regards attributions of credibility as "divorced from any particular use or application of the models"; instead, Parker argues, we should recognize a growing need, especially in the case of complex modeling analyses such as in climate science, “to try to discern, in a principled and careful way, what a […] model’s performance” in specific contexts “indicates about its adequacy (or inadequacy) for various predictive and explanatory purposes.” (Parker 2009, p. 243) Like the scientists quoted above, Parker believes that such *adequacy-for-purpose*, for many modeling contexts, is a more important desideratum than truth or wholesale empirical adequacy. As Parker puts it,

> adequacy-for-purpose does not work like truth and empirical adequacy[…:] from the assumption that a model is adequate for an explanatory or predictive purpose, information about how the model is likely to perform in various other respects, or information about what other properties the model is likely to possess, does not simply follow as a matter of course. (Parker 2009, p. 238)

This suggests that assessments of the credibility of models will, by necessity, always have to be tentative and context-dependent—even if, on rare occasions, a model may turn out to be successful and credible across a wide range of questions and applications.

None of this should come as a surprise: after all, scientists use models in situations of incomplete knowledge—for example, because an underlying fundamental theory cannot be directly applied in any straightforward way to the case at hand, or because it is not even available in principle, or because the available data suggests, but does not entail, a particular interpretation of an empirical phenomenon. Models are also being employed in contexts where a “full” description or derivation may simply be too costly, perhaps because it would require too much time, computational resources, or the like. In all of these situations, it is natural to expect that modelers will face trade-offs—e.g., between completeness (of a model description and derivation) and timeliness (of results and predictions), or between the generality, realism, and precision of one’s models (Levins 1966). Given that trade-offs prevent us from maximizing all desiderata simultaneously, and given that different purposes call for different desiderata to be maximized, we simply should not expect to find that the same model is the most adequate across all contexts. For the same reason, it would be largely uninformative to simply extend the same set of “standard criteria for evaluating the adequacy of a theory” (Kuhn 1977, p. 322) —that is, accuracy, consistency, scope, simplicity, fruitfulness, and other theoretical virtues—to the case of models, without specifying in more detail how these—no doubt worthy—desiderata can be achieved and ascertained in the case of modeling. (See also Chap. 40 by Hirsch Hadorn and Baumberger in this volume.) In short, if one’s goal is to achieve adequacy-for-purpose, a more fine-grained approach to assessing the credibility of one’s conceptual model should be favored. Achieving model credibility—especially as the basis for future simulations—is a complex process and, given “the extent to which this process focuses on elements *external* to anything we would reasonably include as part of theory, it would be unrealistic to interpret this warranting process as being about the relationship of the results to some formal model” of a theory (Winsberg 2001, p. S450).

10.5 Empirical Fit and Causal Understanding

A major tension in the notion of model credibility arises from the question of whether models should always strive to fit the actual world, or whether they can serve explanatory (and other legitimate) functions by imagining plausible (yet counterfactual) scenarios. In this section and the next, we will discuss each of these aspects in detail, keeping in mind that they point to a difference in emphasis, rather than to any fundamental inconsistency. Often, a model that is deemed a successful representation of real-world findings will also successfully predict what would happen if conditions were different, for example, because it has correctly identified an underlying causal mechanism. If we are lucky enough to have a well-confirmed fundamental theory at our disposal, and if the particular problem situation at hand poses no special obstacles to the theory's application, we may even be able to derive a model that "inherits," so to speak, the underlying theory's strengths. This is how philosophers of science used to think about scientific models *in general*, before realizing the significantly greater heterogeneity, diversity, and tentativeness of models in actual scientific practice. Yet, wherever it is, in fact, possible to embed a model in a theory, the credibility of the underlying theory can legitimately rub off on the model as well. At the same time, a highly simplified model—perhaps even one that could not, in principle, be realized—may also be a source of insight about why things are the way they are, for example, because it showcases why certain alternative scenarios could not play out in the world as we know it. Still, there exists more than just a difference in emphasis between, on the one hand, treating models primarily as accurate representations of real-world phenomena, and on the other hand, treating them as "credible worlds" (Sugden 2000) in their own right, which allow us to explore relationships which may, or may not, obtain in (or shed light on) the actual world.

The tension between those who regard models as a way of accounting for empirical data and those who are willing to grant models greater independence from empirical phenomena, is largely due to the fact that models occupy a middle ground between theory and data. Some philosophers of scientific models have even gone so far as to claim that the primary function of models is to serve as "mediators" between the empirical world and the realm of theoretical hypotheses: models, on this account, "are *not* situated in the middle of an hierarchical structure between theory and the world", but operate outside the hierarchical "theory-world axis" (Morrison and Morgan 1999, pp. 17–18). On rare occasions, we may be able to describe the modeling process as an instance of *applying* a fundamental theory to a specific case at hand; but more often than not, such a description would be wildly inaccurate, since modeling often involves interpolating between different realms, making multiple (sometimes inconsistent) assumptions, engaging in different rounds of idealization and de-idealization—all of which render scientific models typically "a *mixture* of elements, including those from outside the domain of investigation" (Morrison and Morgan 1999, p. 23). This echoes a sentiment by Nancy Cartwright, who has long held that theories "do not generally represent what happens in the world—only models represent in this way" (Cartwright 1999, p. 180). While such a view of scientific models opens up a wider range of considerations that modelers can hope to draw on in their quest for model credibility, and while it is generally agreed that models are “inherently intended for specific phenomena” (Suárez 1999, p. 75), when the rubber hits the road—viz., in any given actual instance of model-based inquiry—this view is no clearer than its predecessors on when a model should be deemed credible.

Scientists often place great store by a model’s ability to reproduce empirical data; yet, given the inevitable simplifications that go into the design of serviceable scientific models, they are at the same time well aware of the fact that perfect empirical fit, even across a range of situations, may be a fluke—or simply the lucky result of errors from different sources canceling out. One way to guard against mistaking such accidental successes of a model for a sign of its overall credibility is to systematically broaden the range of situations being considered and test a model’s performance against the corresponding data sets. If a model performs well with respect to a wide range of independent empirical situations and data sets, it becomes progressively unlikely that its successful performance is entirely a matter of chance. This is why, in addition to predictive successes, retrodiction is likewise valued, since it affords an alternative way of comparing a model against empirical reality—provided the past data in question was not itself used in the construction of the model: “A model demonstrates empirical fit to the extent that its logical implications are observed in data; the data may be historical or not yet observed” (Yuengert 2006, p. 87). When viewed from this angle, ascertaining that model *A* has a better empirical fit than model *B* becomes a matter of demonstrating that *A* entails more empirical consequences found in the data than *B*.

While continued empirical success is a good, if fallible, indicator of a model’s “being on to something”, it is clear that it cannot be the final word on what makes a model credible. As an example from the special sciences, consider economic models of addictive behavior, in particular, rational addiction models and time inconsistency addiction models. (The discussion in this paragraph follows Yuengert 2006.) Rational addiction models depict consumers as forward-looking and seeking to maximize utility over their life cycle, all the while taking into account the future consequences of their (current) choices. Addictive reinforcement, on this model, merely reflects the assumption that an increase in the addictive stock increases the marginal utility of current consumption. This contrasts sharply with time inconsistency models, according to which consumers have self-control problems and cannot trust themselves to enact their consumption plans, even when in possession of full information. Time-inconsistent consumers, at any moment, pursue immediate gratification more than they would have professed to prefer at any previous point in time. One might expect such radically different assumptions to make an observable difference, once the two types of models are put to the test. Yet, interestingly, both types of models “are nearly indistinguishable by conventional econometric methods” (Yuengert 2006, p. 78), and the case rational addiction and time inconsistency models is sometimes regarded as a case of underdetermination by data (see Goldfarb et al. 2001).

In situations where two models are empirically equivalent, yet a decision needs to be made as to which model should be adopted (e.g., because, due to lack of time and resources, it is not possible to pursue both modeling strategies simultaneously), one evidently needs to appeal to selection criteria other than empirical fit. This may take the form of privileging certain theoretical desiderata—notably, simplicity and parsimony—or may be guided by background assumptions about the causal basis of the phenomenon in question. Thus, in the case of addiction models, it has been argued that ensuring that models of habit information be formulated in terms of rational decision-making leads to models that are "formally equivalent to models without habit formation" (Spinnewyn 1981, p. 92), but only by redefining wealth and the cost of current consumption in unwieldy ways; in other words, the rationality assumption "leads to unnecessary complications" (Chaloupka et al. 2000, p. 115). Yet such a conclusion may be unacceptable to proponents of rational choice theory, for whom the rationality assumption is a nonnegotiable core element of their paradigm. In addition to background commitments and general theoretical virtues, however, there is a third set of considerations that stem from realism about the purported causal basis of the phenomenon being modeled. Thus, in the case at hand, those who take independent findings, including on issues unrelated to addiction, to establish beyond doubt the realism and causal significance of time inconsistency for behavior in general, will likely consider an explanation of addiction as being caused by time-inconsistent preferences to be superior.

In recent years, predictive modeling, not least on the basis of AI algorithms and machine learning, has taken hold across a wide range of activities, in the data-centric sciences as well as in business and the corporate world. To be sure, the goal of prediction has been an integral part of science (and, by extension, of scientific modeling and simulation) from early on; only in recent years, however, has it become a realistic prospect to sift through vast amounts of data in the search for correlations and to "train" neural networks on training data in the hope that they will successfully predict new samples (or recognize relevant features in incoming data). Such data-driven approaches bring their own challenges. Without a firm (and independently justifiable) set of prior assumptions, the only justification of such models consists in their continued predictive successes. Common dangers include committing so-called "Type III" errors (i.e., developing a model that answers the wrong question), over-fitting of models to the data (e.g., when a model reflects the structure of a given data set—including its noise—so well that its predictions do not generalize to new data sets *of the same kind*), and ignoring systematic changes in the environment (such that past data fails to be a guide to the future). In such a situation, model-immanent approaches can only do so much to alleviate any shortcomings a model has acquired by way of how it was developed—even when a model appears to be empirically successful, e.g., in relation to a given data set (as in the case of overfitting). Great care must, therefore, go into the very construction of data-driven models, e.g., by deploying more sophisticated sampling techniques.

Empirical fit, then, is only one consideration among several that researchers draw on when they attempt to determine how much credibility a model merits. Theoretical virtues such as parsimony and unification may aid in resolving situations where multiple models are empirically equivalent; appeals to the realism of assumptions likewise have a role to play, in particular in the following two cases: "if realistic assumptions can be expected to result in better empirical fit eventually, or if realistic assumptions promote worthy goals other than empirical fit" (Yuengert 2006, p. 87). To the latter possibility—that empirical fit may be outweighed by, or may at least trade off against, other worthy goals of inquiry—we shall now turn.

10.6 Models and the Exploration of Credible Worlds

Empirical fit and numerical accuracy may indicate that a model stands in the right sort of relation to its target system, and to the world at large, but, for the reasons already outlined, they can at best constitute defeasible evidence that a model satisfies the kind of "world-linking conditions" (Grüne-Yanoff 2009, p. 81) that would merit trust in its future performance and overall credibility. Turning from one set of desiderata (relating to a model's accuracy and empirical fit) to the other cluster of desiderata identified earlier, viz. a model's aptness for exploring possible scenarios and generating "modal knowledge about what might be *possible* about the target system" (Massimi 2018, p. 339), as well as its fruitfulness in generating truth-conducive lines of inquiry, an analogous point can still be made. After all, though it may be difficult to quantify and compare fit and accuracy—let alone infer on their basis how credible a model is *overall*—we may expect assessments of explanatory success and exploratory potential to be even more controversial. Nonetheless, in what follows, I shall sketch two approaches that tackle these more qualitative criteria of how much insight into the world a given model affords us. The first such approach conceives of models as a way of constructing *credible worlds*; the second regards them as *exploratory tools*.

The term "credible world" as a characterization of the way models operate is due to the economist Sugden (2000). It emphasizes that, for a model to afford insight to its user, it need not always be derived from an idealization of an actual target system. Sometimes, such idealizations may be possible: when modeling mechanical motion, friction is often treated as negligible, and the resulting mechanical models may be considered as idealized representations of actual bodies in motion (where friction is inevitable), which nonetheless display wide applicability. Yet, in many areas of science—not least those that deal with complex systems (such as the social sciences, including economics)—it is often difficult, if not impossible, to determine in advance which factors can, or cannot, be neglected. If idealization is understood as the process of starting from real-world target systems and then proceeding by isolating causally important factors from those of minor significance, then this methodology may face serious limitations when it comes to complex systems.[4] By contrast, Sugden's notion of "credible world," in line with other recent accounts of model-based science, acknowledges the central role of model *construction*. Modeling, in essence, amounts to the construction of credible worlds using representational tools (such as mathematics); whether these successfully "latch on" to the actual world is a question

[4] This, it should be noted, is not the only that one may interpret the procedure of theoretically "isolating" relevant factors; for an alternative view, see (Mäki 2009).

which can only be successfully tackled once a model has been specified. Whereas in the case of gradual idealization from real target system the satisfaction of world-linking conditions can be assumed (if, perhaps, only initially), in the case of models as credible worlds, such a linkage needs to be established subsequently—e.g., by relying on such criteria as similarity, induction, and explanatory success. As Chao (2014, p. 591) puts it rather succinctly,

> if a prototype theoretical model can be applied to a set of particular models across time, space, and context, and each particular model is regarded as satisfactorily explaining a particular […] real-world phenomen[on], then it can be inductively concluded that this prototype theoretical model is credible.

Naturally, on this account, an important criterion of the credibility of models is their *coherence*—both internally and with known external (e.g., causal) constraints: "If a model lacks coherence, its results cannot be seen to follow naturally from a clear conception of how the world might be" (Sugden 2000, p. 26). Yet, coherence among the model's assumptions does not itself suffice: "For a model to have credibility, it is not enough that its assumptions cohere with one another; they must also cohere with what is known about causal processes in the real world". (ibid.) However, in order to conclude, with confidence, that a given credible-world model represents the way the world *actually* is, a further step will typically be required: viz., the abductive inference that, "[i]f a result R is caused by a set of causal factors F in the model world M, and R occurs in the real world W, then we have reason to believe that F operates in W" (Chao 2014, p. 592). Failure to establish a model's real-world connection need not, however, disqualify the model from further study: while such a model could not function as a *surrogate* for its real-world target, it may nonetheless *substitute for* real-world inquiry.[5]

A similar conclusion regarding the sources of a model's credibility may be reached from a perspective that acknowledges that models may serve a variety of functions. In addition to representing actual target systems and deriving specific results, predictions, and explanations about them, models also help *explore* further avenues of inquiry. From this perspective, whether or not a model is credible may not be solely a matter of how faithfully it represents a given target system, and how closely its results mirror the latter, but may also depend on how fruitful it is—for example, in the generation of potential explanations, or when it comes to establishing in-principle possibilities (or, as the case may be, impossibility theorems—which may play an important role in guiding future inquiry). On the one hand, this acknowledges that, in order to make headway in our attempts to model reality, we must sometimes introduce falsehoods—not merely as an unavoidable side effect of idealization and abstraction, but as a direct and deliberate consequence of making (sometimes heavy-handed) model assumptions; on the other hand, it broadens the range of legitimate uses of models to also include those instances of modeling that precede the full theoretical articulation of a phenomenon (or class of phenomena). Not all legitimate uses of a model should, of course, be thought of as bolstering its *credibility*. Yet, if

[5] I am here drawing on Uskali Mäki's distinction between "surrogate models" and "substitute models" (Mäki 2009, pp. 35–37).

we take seriously the earlier idea that one can draw a parallel between the credibility of models and the trustworthiness of interlocutors (see Sect. 10.2), it is not at all far-fetched to insist that the credibility of a model goes beyond the brute empirical reliability of the individual claims it makes about the world. Just as one person may be deemed a "more credible choice" (as a candidate for political office, say) than another, some models may be considered more credible than others—not because of any decisive difference in their past track record, but because background considerations suggest that they hold more promise than their competitors. Judgments of credibility, we might say, are forward-looking—if positive, they engender trust in future performance—in a way that is not captured by looking at brute empirical track record alone. This line of argument is entirely compatible with, but does not presuppose, the idea that some models may be best thought of as "credible worlds." Yet it acknowledges more explicitly that "exploratory modeling often serves the purpose of developing a grasp of (as yet theoretically inaccessible) phenomena" (Gelfert 2016, p. 95)—a situation that scientists encounter all the time. It also serves as a reminder that assessments of a model's credibility depend importantly on their role and function: treating a model as a credible "proof of principle" (Gelfert 2016, pp. 85–86), say, is a quite different matter from relying on it as a credible source of precise numerical predictions.

As an example of such exploratory models, consider certain types of "toy models"—viz., models that are strongly idealized and simplified, so much so that they may border on being minimal, "stylized" accounts of a single aspect of a target phenomenon. Some such models may be derived from empirically well-confirmed theories, e.g., when planetary motion is modeled as two point masses orbiting one another. Such cases may be considered "embedded toy models," since they are at the same time models of an underlying well-developed theory *and* extremely simplified and idealized models of phenomena. This contrasts with what has been called "autonomous toy models" (Reutlinger et al. 2017): that is, extremely stylized models that are not models of a theory (and which, in some cases, "seem to bear no relevant relation to a well-confirmed framework theory" at all; ibid.: 11). When the lack of such a relevant relation to an underlying theory is due to the absence of well-developed theoretical resources that one might otherwise draw on, we may properly deem such autonomous toy models "exploratory" in the sense discussed in the previous paragraph. Whereas, in the case of embedded toy models, the underlying theory usually contains some of the resources required for successfully "de-idealizing" and applying a model to a given target situation for predictive purposes, in the case of autonomous toy models—and of exploratory models, more generally—empirical prediction, let alone numerical accuracy, is rarely the explicit goal. Instead, researchers often deploy exploratory models with the aim at generating intelligible explanations of certain types of phenomena. The more we succeed in cultivating our ability to "recognise qualitatively characteristic consequences [. . .] without performing exact calculations" (De Regt and Dieks 2005, p. 151), the more trust we place in exploratory models, all the while recognizing their intrinsic limitations. While we must guard against mistaking any subjective "aha experiences" or fleeting feelings of familiarity for signs of the truth of our models, we should not downplay the importance of understanding to successful scientific practice. When researchers "refer to the results of their simulations by saying 'we trust our results' or 'we trust our computer simulations'" they not only claim that the results are true (or approximately true), but also that they "understand why they are correct (or approximately correct)" (Durán 2018, p. 98). Recognizing exploratory fruitfulness as contributing to the credibility of a model does not entail that such fruitfulness can somehow compensate for other deficiencies (e.g., representational failure, or lack of realism); rather it amounts to yet another acknowledgment that the credibility of a model is the joint result of features of the model, aspects of the world, and the cognitive goals and interests of its users.

10.7 Summary

Simulation-based methods in science are deeply intertwined with the development of credible conceptual models, where the latter may be variously thought of as (pragmatically adequate) representations of real-world targets, as constructions of "credible worlds" that display internal and external coherence, or as exploratory tools that may aid our understanding even before a well-developed theoretical account of a phenomenon, or class of phenomena, becomes available. While it would be wrong to think that computer simulation methods are a "mere application" of the conceptual model to a particular problem situation, it would likewise be misleading to assume that the credibility of a simulation can somehow be divorced from that of the model. To be sure, verifying that a computer simulation performs as intended and ascertaining furthermore that its results are valid across a range of empirical contexts are important steps in establishing a simulation's credibility. Yet, whether or not the results of a simulation are credible also hinges on the credibility, or lack thereof, of the underlying model. Yet such credibility, as this chapter has aimed to show, is not up to the model alone, but is jointly constituted by features of the model and the overall goals and cognitive interests of its users.

In much the same way that we demand of credible human interlocutors that they be competent and trustworthy—that is, able and willing to give reliable information across a range of relevant questions, without sudden unexpected failures—credible conceptual models should make reliable and relevant information available to those that depend on them. While this will often be a matter of how faithfully a model represents an actual target system, it would be hasty to think that this exhausts the concerns of scientists who depend on models and simulations. Sometimes, creating a model that constitutes a "credible world" in its own right—even if does not map on to the actual world—can help advance our understanding, and on yet other occasions, the most credible model may simply be the one that shows the most promise in generating fruitful new lines of research. Whether a conceptual model enjoys credibility, then, is as much a matter of its intrinsic structure and its relation to the world at large as it is a reflection of the goals and cognitive interests of its users.

References

Caro, J. J., Eddy, D. M., Kan, H., Kaltz, C., Patel, B., Eldessouki, R., & Briggs, A. H. (2014). Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: An ISPOR-AMCP-NPC good practice task force report. *Value in Health*, *17*(2), 174–182.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Chaloupka, F. J., Tauras, J. A., & Grossman, M. (2000). The economics of addiction. In P. Jha & F. J. Chaloupka (Eds.), *Tobacco control in developing countries* (pp. 107–129). Oxford: Oxford University Press.

Chao, K.-K. (2014). Models and credibility. *Philosophy of the Social Sciences, 44*(5), 588–605.

De Regt, H., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, *144*(1), 137–170.

Durán, J. M. (2018). *Computer simulations in science and engineering: Concepts, practices, perspectives*. New York: Springer.

Gelfert, A. (2016). *How to do science with models: A philosophical primer*. Cham: Springer.

Goldfarb, R., Leonard, T., & Suranovic, S. (2001). Are rival theories of smoking underdetermined? *Journal of Economic Methodology*, *8*(2), 229–251.

Grüne-Yanoff, T. (2009). Learning from minimal economic models. *Erkenntnis*, *70*(1), 81–99.

Knutti, R. (2008). Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society A*, 4647–4664.

Kuhn, T. S. (1977) Objectivity, value judgment, and theory choice. In *The essential tension*. Chicago: The University of Chicago Press.

Law, A. M., & Kelton, W. D. (1991/2000). *Simulation modeling and analysis*. New York: McGraw-Hill.

Lenhard, J. (2007). Computer simulation: The cooperation between experimenting and modeling. *Philosophy of Science*, *74*(2), 176–194.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, *54*(4), 421–431.

Mäki, U. (2009). MISSing the world: Models are isolations and credible surrogate systems. *Erkenntnis, 70*(1), 29–43.

Massimi, M. (2018). Perspectival modeling. *Philosophy of Science, 85*(3), 335–359.

McGinnies, E., & Ward, C. D. (1980). Better liked than right: Trustworthiness and expertise as factors in credibility. *Personality and Social Psychology Bulletin, 6*(3), 467–472.

Merchant, C. (1980). *The death of nature: Women, ecology, and the scientific revolution*. San Francisco: Harper and Row.

Morrison, M., & Morgan, M. S. (1999). Models as mediating instruments. In M. Morrison & M. S. Morgan (Eds.), *Models as mediators: Perspectives on natural and social science* (pp. 10–37). Cambridge: Cambridge University Press.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the Earth sciences. *Science, 263*(5147), 641–646.

Parker, W. (2009). Confirmation and adequacy-for-purpose in climate modelling. *Proceedings of the Aristotelian Society, 83*(1), 233–249.

Pitt, J. (2007). Speak to me. *Metascience, 16*(1), 51–59.

Reutlinger, A., Hangleiter, D., & Hartmann, S. (2017). Understanding (with) toy models. *British Journal for the Philosophy of Science,* 1–31. Advance article https://doi.org/10.1093/bjps/axx005.

Robinson, S. (2008). Conceptual modeling for simulation, Part I: Definition and requirements. *Journal of the Operational Research Society, 59*(3), 278–290.

Robinson, S. (2011). Choosing the right model: Conceptual modeling for simulation. In S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, & M. Fu (Eds.), *Proceedings of the 2011 Winter Simulation Conference* (pp. 1428–1440). IEEE.

Rykiel, E. J. (1996). Testing ecological models: The meaning of validation. *Ecological Modelling, 90*(3), 229–244.

Spinnewyn, F. (1981). Rational habit formation. *European Economic Review, 15*(1), 91–109.

Suárez, M. (1999). Theories, models, and representations. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 75–83). New York: Plenum Publishers.

Sugden, R. (2000). Credible worlds: The status of theoretical models in economics. *Journal of Economic Methodology, 7*(1), 1–31.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, Mass.: Harvard University Press.

Winsberg, E. (2001). Simulations, models, and theories: Complex physical systems and their representations. *Philosophy of Science*, *68*(Proceedings), S442–S454.

Winsberg, E. (2018). *Philosophy and climate science*. Cambridge: Cambridge University Press.

Yuengert, A. M. (2006). Model selection and multiple research goals: The case of rational addiction. *Journal of Economic Methodology, 13*(1), 77–96.

Zacharias, G. L., MacMillan, J., & Van Hemel, S. B. (2008). *Behavioral modeling and simulation: From individuals to societies*. Washington, DC: The National Academies Press.

Chapter 11
The Foundations of Verification in Modeling and Simulation

William J. Rider

Abstract The practice of verification is grounded in mathematics highlighting the fundamental nature of its practice. Models of reality are fundamentally mathematical and verification assures the connection between the modeling intended and achieved in code. Code verification is a process where the correctness of a computer code for simulation and modeling is proven. This “proof” is defined by the collection of evidence that the numerical approximations are congruent with the model for the physical phenomena. **The key metric in code verification is the order of accuracy of the approximation that should match theoretical expectations. In contrast, solution verification is an aspect of uncertainty estimation associated with numerical error in simulations.** Solution verification uses many of the same approaches as code verification, but its principal outcome is an estimate of the numerical error. The order of convergence is a secondary outcome. Together these two practices form an important part of the foundation of quality and credibility in modeling and simulation.

Keywords Verification · Error estimate · Convergence · Order-of-accuracy · Solution verification · Robust statistics

11.1 Verification in Modeling and Simulation

> A very great deal more truth can become known than can be proven. – Richard Feynman, “The Development of the Space-Time View of Quantum Electrodynamics,” Nobel Lecture (11 December 1965, (Brown and Feynman 2000), p. 29)

Many models are expressed using the language of mathematics most often in the form of differential equations. For simple or idealized cases, the equations for the model can be solved though analytical means. For a variety of reasons, however, the analytical solutions are limited and inadequate for modeling most circumstances. For this reason models are solved via computers and via numerical approximations. In

W. J. Rider (✉)
Sandia National Laboratories, Center for Computing Research, Albuquerque, NM 87185, USA
e-mail: wjrider@sandia.gov

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation, Simulation Foundations, Methods and Applications*,
https://doi.org/10.1007/978-3-319-70766-2_11

the process of numerical approximation the propriety of computer modeling to the intended model should not be assumed, but rather proven through careful checks and tests. How do we do this in practice? How can we provide proof of the correctness of the intended model as well as the accuracy of the model?

The answer to this question is the processes and procedures of *verification*. This chapter is devoted to describing the foundational aspects of verification.

Theoretically, the conditions for the numerical approximation to solve a model correctly are well known. The process of verification harnesses this theoretical knowledge to provide confidence in the numerical solution. In addition, the numerical solutions are approximate thus (almost) always deviating from the exact solution.[1] Verification also provides means to quantify this error.

In modeling and simulation verification is a set of activities broadly supporting the quality. Verification is a dominantly mathematical exercise that assures the quality of models. In this chapter the utility of verification is focused on models described by differential equations. The practice of verification for models based on other approaches is not discussed here. **Verification is paired with validation as a means to assure modeling quality.** Validation is the comparison of simulated results with physical reality either from experiments or observations. Validation is an exercise in physics and engineering.

Verification consists of two modes of practice: **code verification** where the mathematical correctness of the computer code is assessed, and **solution (calculation) verification** where the numerical error (uncertainty) is estimated (for verification, these errors include discretization error, linear, and nonlinear equation solution tolerance/discrepancy, round-off error, and parallel processing consistency; see Chap. 3 by Roy in this volume for a classification of errors). Solution verification is a core practice in estimating numerical error in simulations subjected to validation, or simply as part of a modeled prediction. Both activities are closely linked to each other and they are utterly complementary in nature. To a large extent the methodology used for both types of verification is similar. Much of the mathematics and flow of work are shared in all verification, but details, pitfalls and key tips differ (Rider et al. 2016), and the differences between the two are important to maintain (Roache 1998, 2009; Stern et al. 2001; Oberkampf and Roy 2010).

The remainder of the chapter will provide basic overviews of code verification (Sect. 11.2) and solution verification (Sect. 11.4) including simple examples of each of these key practices.

[1] There are some simple cases where numerical solutions are exact. These cases are often simple or highly contrived such as the finite element “patch” test.

11.2 Code Verification

> Science replaces private prejudice with public, verifiable evidence. – Richard Dawkins, The Enemies of Reason, The Irrational Health Service[2]

Computer modeling and simulation is often an activity where **continuous** mathematics is converted to **discrete** computable quantities. This process involves approximation of the continuous mathematics and in almost every non-pathological circumstance is inexact. The core of modeling and simulation is the solution of (partial) differential equations using approximation methods. **Code verification is a means of assuring that the approximations used to make the discrete solution of differential equations tractable on a computer are correct.** A key aspect of code verification is determining that the discrete approximation of the differential equation is consistent with the continuous version of the differential equation.

Analytical **consistency** demands that the so-called order of accuracy of the differential equation be at least one. In other words, the discrete equations produce solutions that are the original continuous equations plus terms that are proportional to the size of the discretization or to a higher power of it. In verification practice, consistency is the observation of a positive convergence rate indicating convergence. Numerical methods can properly exhibit sublinear convergence rates while operating completely correctly (Banks et al. 2008). This character may be examined by solving problems with an exact analytical solution (or a problem with very well controlled and characterized errors) using several discretization sizes allowing the computation of errors, and determining the order of accuracy.

It has been shown that the combination of consistency with **stability** of the approximation means that the approximation converges to the correct solution of the continuous differential equation (Lax and Richtmyer 1956). A simple definition of stability is the nature of a solution where small changes in the conditions result in small changes in solution. If the solution changes greatly due to small changes, the solution is unstable. The stability is associated with the appropriate concept for the approximation in question, e.g., 0-stability for ODEs or Von Neumann stability for PDEs. For steady-state solutions require a certain attention to stability of the approximation when approached iteratively. In these cases the stability can mimic that associated with time-dependent solutions.

I will examine both the nature of different types of problems to determine code verification and the methods of determining the order of accuracy. **One of the key aspects of code verification is the congruence of the theoretical order of accuracy for a method, and the observed order of accuracy.** The latter is obtained from a comparison of simulation outputs with a reference solution. The former is usually determined via a numerical analysis exercise using a series expansion method such as Taylor series or Fourier series. It can be done via pen and paper, chalkboard or more recently via symbolic arithmetic. The discrete system of variables is approximated

[2] Documentary, dir. Russel Barnes; writer Richard Dawkins; 2007, here quoted from https://www.goodreads.com.

and the approximations are then replaced by one of these series expansions. Such methods are described in numerical analysis texts where solvers for ODEs are the simplest rudimentary examples.

It is important to note that the theoretical **order of convergence** also depends upon the problem being solved. The order of convergence is a product of both this order of accuracy, and the problem being solved and its degree of mathematical smoothness, the existence of derivatives of the function. The problem must possess enough regularity to support the convergence rate expected. For some problems the correct observed rate of convergence does not match the asymptotic rate of convergence because of the nature of the solution or the lack of an asymptotic range itself (i.e., problems with singularities) (Banks et al. 2008). **At this point it is important to point out that code verification produces both an order of accuracy and an observed error in solution. Both of these quantities are important. For code verification, the order of accuracy is the primary quantity of interest. It depends on both the nature of the approximation method and the problem being solved.** If the problem being solved is insufficiently regular and smooth (sufficient existence of derivatives of the functional solution), the order of accuracy will not match the theoretical expectations of the method.

As mentioned above, the key idea behind code verification is determining that the discrete approximation of the differential equation is consistent with the continuous version of the differential equation. To ensure this, the error is determined by comparing the output of the computer simulation with an otherwise available solution. Ideally the available solution is produced by a precise analytical method. One can then produce a sequence of numerical approximations and check to see if the theory available explains or predicts the dependence of the error on the parameter defining the sequence. If the theory and the observed behavior is consistent, the sequence provides evidence of correctness.

When one conducts a code verification study there is a basic flow of activities and practices to conduct. First, one looks at a code to target and a problem to solve. Several key bits of information should be immediately focused upon before the problem is solved. What is the order of accuracy for the method in the code being examined, and what is the order of accuracy that the problem being solved can expose? In addition the nature of the analytical solution to the problem should be carefully considered. For example what is the nature of the solution? Closed form? Series expansion? Numerical evaluation? Some of these forms analytical solution have errors that must be controlled and assessed before the codes method may be assessed. By the same token are there auxiliary aspects of the codes solution that might pollute results? Solution of linear systems of equations? Stability issues? Computer roundoff or parallel computing issues? In each case these details could pollute results if not carefully excluded from consideration.

Next, one needs to produce a solution on a sequence of meshes by running the simulation program. For simple verification using a single discretization parameter only two discretizations are needed for verification (two equations to solve for two unknowns). **For code verification the standard model for error is simple, generally a power law, $E_k = Ah_k^a$ where the error E_k in the kth solution is proportional**

to the discretization parameter h_k to the power (order) a. The constant of proportionality is A. The order, a, is the target of the study and one looks at its congruence with the expected theoretical order for the method on the problem being solved. It is almost always advisable to use more than the minimum number of meshes to assure that one simply is not examining anomalous behavior from the code.

One of the problems with code verification is the rarity of the observed order of convergence to exactly match the expected order of convergence (for numerical work what does "exact" mean?). The question of how close is close enough haunts investigations. Invariably the observed order will deviate from the expected order by some amount. The question for the practitioner is how close is acceptable? Generally this question is given little attention. **There are more advanced verification techniques that can put this issue to rest by producing uncertainties on the observed order, but the standard techniques simply produce a single result.** Usually this results in rules of thumb that apply in broad brushes, but undermine the credibility of the whole enterprise. Often the criterion is that the observed order should be within a tenth of the theoretically expected result. More generally, common sense should be applied to the decision-making as long as it can be justified.

Another key caveat comes up when the problem is discontinuous, meaning the solution of the equations contains a jump in variables. In this case the observed order is either set to one for nonlinear solutions, or weakly tied to the theoretical order of convergence. **For the wave equation this result was studied by Banks, Aslam and Rider and admits an analytical and firmly determined result.** Banks et al. (2008) **In this case the issue of inexact congruence with the expected rate of convergence remains.** In addition for problems involving systems of equations will have multiple features each having a separate order of convergence, and the rates will combine within a solution. Ultimately in an asymptotic sense the lowest order of convergence will dominate as $h \rightarrow 0$. This is quite difficult to achieve practically.

The last major issue that comes up in code verification (and solution verification too) is the nature of the discrete mesh and its connection to the asymptotic range of convergence. All of the theoretical results apply when the discretization parameter h is small in a broad mathematical sense. This is quite problem specific and generally ill defined. Examining the congruence of the numerical derivatives of the analytical solution with the analytical derivatives can generally assess this. When these quantities are in close agreement, the solution can be considered to be asymptotic. Again these definitions are loose and generally applied with a large degree of professional or expert judgment.

It is useful to examine these issues through a concrete problem in code verification. The example I will use is a simple ordinary differential equation integrator for a linear equation $du(t)/dt = -au(t)$ coded up in Mathematica. We could solve this problem in a spreadsheet (like MS Excel), python, or a standard programming language. The example will look at two first-order methods, forwards $u(t_{n+1}) + ahu(t_n) = u(t_n)$ and backwards $u(t_{n+1}) + ahu(t_{n+1}) = u(t_n)$ Euler methods. Both of these methods produce leading first-order errors in an asymptotic sense, $E_k = Ch_k + O(h_k^2)$. If h_k is large enough, the high-order terms will pollute the error and produce deviations from the pure first-order error. **Let us look at this example and the concrete analysis**

from verification. This will be instructive in getting to similar problems encountered in general code verification.

Here is the Mathematica code we are going to verify

```
ForwardEuler[h_, T_, a_] :=
(
uo = 1;
t = 0.0;
While[t < T,
(* integration *)
t = t + h;
un = uo + a h uo;
Print["t= ", t, " u(t) = ", un, " err = ", Abs[un - Exp[a t]]];
uo = un
];
)

BackwardEuler[h_, T_, a_] :=
(
uo = 1;
t = 0.0;
While[t < T,
(* integration *)
t = t + h;
un = uo/(1 + a h);
Print["t= ", t, " u(t) = ", un, " err = ", Abs[un - Exp[a t]]];
uo = un
];
)
```

This is coded in Mathematica for both a forward and backward Euler method for a linear ODE. The code is similar to what one might code in C or Pascal. It is a coded version of the difference equations expressed in the previous paragraph. This is probably about as simple as a numerical method can be.

Let us look at the forward Euler integrator for several different choices of h, different end times for the solution and a number of discrete solutions using the method. I will do the same thing for the backwards Euler method, which is different because it is unconditionally stable with respect to step size. Both the forward and backwards Euler methods are first-order accurate as a analyzed via Taylor series expansions. For this simple ODE, the method is stable to a stepsize of $h = 2$ and I can solve the problem to three stopping times of $T = 1.0$, $T = 10.0$ and $T = 100.0$. The analytical solution is always, $u(T) = \exp(-aT)$. I then solve this problem using a set of step sizes, $h = 1.0$, $h = 0.5$, $h = 0.25$, $h = 0.125$ to demonstrate different convergence behaviors.

Table 11.1 Computed order of convergence for forward Euler (FE) and backward Euler (BE) methods for various stopping times and step sizes. The rates of convergence deviate from theory under many circumstances

h	FE T = 1	FE T = 10	FE T = 100	BE T = 1	BE T = 10	BE T = 100
1	1.64	0.03	0	0.79	1.87	16.99
0.5	1.20	0.33	4e-07	0.88	1.54	11.78
0.25	1.08	0.65	0.002	0.93	1.30	7.17
0.125	1.04	0.83	0.05	0.96	1.16	4.07
0.0625	1.02	0.92	0.27	0.98	1.08	2.40
0.03125	1.01	0.96	0.55	0.99	1.04	1.63

I can give results for various pairs of step sizes with both integrators, and see some common pathologies that we must deal with. Even solving such a simple problem with simple methods can prove difficult and prone to heavy interpretation (arguably the simplest problem with the simplest methods). Much different results are achieved when the problem is run until different stopping times. We see the impact of accumulated error (since I am using Mathematica so aspects of round-off error are pushed aside). In cases where round-off error is significant, it would be another complication. Furthermore the backward Euler method for multiple equations would involve a linear (or nonlinear) solution that itself has an error tolerance that may significantly impact verification results.

These results are compiled in Table 11.1 showing significant variation in the method's performance. We see good results for $T = 1.0$ and a systematic deviation for longer ending times. The good results are associated with the observed convergence rate being close to the theoretically expected rate. This problem has been acknowledged in the ODE community for long integration times and is expected. At the core are differences between the local errors associated with truncation error and global error associated with full calculations (Ascher and Petzold 1998; Hairer et al. 1993). We would not expect this to occur for boundary value problems. To get acceptable verification results would require much smaller step sizes (for longer calculations!). **This shows how easy it is to scratch the surface of really complex behavior in verification that might mask correctly implemented methods.** What is not so well appreciated is that this behavior is expected and amenable to analysis through standard methods extended to look for it.

11.3 Types of Code Verification Problems and Associated Benchmarks

Dont give people what they want, give them what they need. – Joss Whedon, Conversations (University Press of Mississippi, 2011, (Lavery and Burkhead 2011), p. 31)

The problem types are categorized by the difficulty of providing a solution coupled with the quality of the solution that can be obtained. These two concepts go hand in hand. A simple closed form solution is easy to obtain and evaluate. Conversely, a numerical solution of partial differential equations is difficult and carries a number of serious issues regarding its quality and trustworthiness. **These issues are addressed by an increased level of scrutiny on evidence provided by associated benchmarks. Each benchmark is not necessarily analytical in nature, and the solutions are each constructed in different means with different expected levels of quality and accompanying benchmarks. This necessitates the differences in level of required documentation and accompanying supporting material to assure the user of its quality.** These recommendations are the direct result of executing code verification for high consequence programs for nearly twenty years. Invariably when code verification produces concerns for an important code, the credibility of the verification problem itself is questioned.

Next, I provide a list of types of benchmarks along with an archetypical example of each. This is intended to be instructive to the experienced reader, who may recognize the example. The list is roughly ordered in increasing level of difficulty and need for greater supporting material (Kamm et al. 2008).

- Closed form analytical solution (usually algebraic in nature). Example: Incompressible, unsteady, 2-D, laminar flow given by the Taylor and Green known as the Taylor-Green vortices (note the 3-D version of this problem does not have an analytical solution, but is a common benchmark problem) Taylor, G. I. and Green, A. E., Mechanism of the Production of Small Eddies from Large Ones, Proc. R. Soc. Lond. A, 158, 499521 (1937) (Taylor and Green 1937; Kim and Moin 1985; Drikakis and Rider 2006).
- Analytical solution with significantly complex numerical evaluation.
- Series solution. Example: Numerous classical problems, in H. Lambs book, Hydrodynamics, Cambridge University Press (or Dover), 1932 (Lamb 1932). Classical separation of variables solution to heat conduction. Example: Incompressible, unsteady, axisymmetric 2-D, laminar flow in a circular tube impulsively started (Szymanski flow), given in White, F. M. (1991). Viscous Fluid Flow, New York, McGraw Hill, pp. 133–134 (White 1991).
- Nonlinear algebraic solution. Example: The Riemann shock tube problem, J. Gottlieb, C. Groth, Assessment of Riemann solvers for unsteady one-dimensional inviscid flows of perfect gases, Journal of Computational Physics, 78(2), pp. 437–458, 1988 (Gottlieb and Groth 1988).
- A similarity solution requiring a numerical solution of nonlinear ordinary differential equations.
- Manufactured Solution. Example: Incompressible, steady, 2-D, turbulent, wall-bounded flow with two turbulence models (makes no difference to me), given in Ea, L., M. Hoekstra, A. Hay and D. Pelletier (2007). “On the construction of manufactured solutions for one and two-equation eddy-viscosity models.” International Journal for Numerical Methods in Fluids. 54(2), 119–154 (Eça et al. 2007).

- Highly accurate numerical solution (not analytical). Example: Incompressible, steady, 2-D, laminar stagnation flow on a flat plate (Hiemenz flow), given in White, F. M. (1991). Viscous Fluid Flow, New York, McGraw Hill. pp. 152–157 (White 1991).
- Numerical benchmark with an accurate numerical solution. Example: Incompressible, steady, 2-D, laminar flow in a driven cavity (with the singularities removed), given in Prabhakar, V. and J. N. Reddy (2006). "Spectral/hp Penalty Least-Squares Finite Element Formulation for the Steady Incompressible Navier–Stokes Equations." Journal of Computational Physics. 215(1), 274–297 (Prabhakar and Reddy 2006).
- Code-to-code comparison data. Example: Incompressible, steady, 2-D, laminar flow over a back-step, given in Gartling, D. K. (1990). "A Test Problem for Outflow Boundary Conditions-Flow Over a Backward-Facing Step." International Journal for Numerical Methods in Fluids. 11, 953–967 (Gartling 1990).

Below is a list of the different types of data associated with verification problems defined above. Here, "data" refers to information that needs to be given about the benchmarks, and not empirical data from the target system of the simulation.

Depending on the nature of the test problem only a subset of these data are advisable. This will be provided below in the following list of data types. As noted above, benchmarks with well-defined closed form analytical solutions require relatively less d than a benchmark associated with the approximate numerical solution of PDEs.

- Detailed technical description of the problem (report or paper)
- Analysis of the mathematics of the problem (report or paper)
- Computer analysis of solution (input file)
- Computer solution of the mathematical solution
- Computer implementation of the numerical solution
- Error analysis of the exact numerical solution
- Derivation of the source term and software implementation or input
- Computer implementation of the source term (manufactured solution)
- Grids for numerical solution
- Convergence and error estimation of approximate numerical solution
- Uncertainty and sensitivity study of numerical solution
- Description and analysis of computational methods
- Numerical analysis theory associated with convergence
- Code description/manuals
- Input files for problems and auxiliary software
- Patch test description, Derivation, input and analysis
- Unusual boundary conditions (inflow, piston, etc.)
- Physics restrictions (boundary layer theory, inviscid,)
- Software quality documents
- Scripts and auxiliary software for verification
- Source code
- Metric descriptions

- Verification results including code version, date, etc.
- Numerical sensitivity studies
- Feature coverage in verification

The mapping of these forms of documentation to the different types of verification problems is given in the appendix at the end of the chapter. The wealth of potential documentation for verification work highlights the complexity of the professional practice of verification. The work is far more involved and technical than commonly appreciated.

The use of direct numerical simulation (DNS) (Moin and Mahesh 1998; Moser et al. 1999; Le et al. 1997) requires a similar or even higher level of documentation than analytical solutions. This topic is being addressed because of increasing interest in using comparison to DNS as a means of assessing the quality and correctness of simulation in many fields most prominently fluid dynamics. DNS is commonly used in fluid dynamics research and other fields also being called "first principles" or "fully resolved" and used in a similar vein. Here, I will adopt the term DNS to describe this broader class of simulations discussing their appropriate use in code verification (Galli and Pasquarello 1993; Kotschenreuther et al. 1995). This coincides with the discussion of the last type of verification benchmark where a complex numerical method with significant approximations is utilized to produce the solution. **As a numerically computed benchmark, the burden of proof is much larger.** Code verification is best served by exact analytical solutions because of the relative ease in assuring benchmark solution accuracy. Nonetheless, it remains a common practice due to its inherent simplicity. It also appeals to those who have a vested interest in the solutions produced by a certain computer code. The credibility of the comparison is predicated on the credibility of the code producing the benchmark used as the surrogate for truth. Therefore the documentation of the benchmark must provide the basis for the credibility.

The use of DNS as a surrogate for experimental data has received significant attention. This practice violates the fundamental definition of validation we have adopted because no observation of the physical world is used to define the data. This practice also raises other difficulties, which I will elaborate upon. First the DNS code itself requires that the verification basis be further augmented by a validation basis for its application. This includes all the activities that would define a validation study including experimental uncertainty analysis numerical and physical equation based error analysis. Most commonly, the DNS is promoted to provide validation, but the DNS contains approximation errors that must be estimated as part of the error bars for the data. For most DNS simulations the results are only significant statistically. As such the data must be processed to produce various statistical measures with necessary attention being paid to statistical convergence as well. **Furthermore, the code must have documented credibility beyond the details of the calculation used as data. This level of documentation again takes the form of the last form of verification benchmark introduced above because of the nature of DNS codes. For this reason I include DNS as a member of this family of benchmarks.**

11.4 Solution Verification

> There are two ways to do great mathematics. The first is to be smarter than everybody else. The second way is to be stupider than everybody else – but persistent. – Raoul Bott[3]

The second form of verification is solution verification. This is quite similar to code verification, but its aim is the estimation of approximation errors in a calculation. When one runs into a problem without an analytical solution, the estimation of errors is more intricate because errors are not known and must be estimated from an approximate solution. One examines a series of solutions and estimates the solution that is indicated by the sequence. Essentially the question of what solution is the approximation appearing to converge toward is being asked. If the sequence of solutions converges, the error in the solution can be inferred. **As with code verification the order of convergence and the error is a product of the analysis. Conversely to the code verification, the error estimate is the primary quantity of interest, and the order of convergence is secondary.**

Most often, solution verification involves examining error and results without the knowledge the exact solution. **This makes it a more difficult task than code verification where an exact solution is known removing a major uncertainty. A secondary issue associated with not knowing the exact solution is the implications on the nature of the solution itself.** With an exact solution, a mathematical structure exists allowing the solution to be achievable analytically. Furthermore, exact solutions are limited to relatively simple models that often cannot model reality. Thus, the modeling approach to which solution verification is applied is necessarily more complex. All of these factors are confounding and produce a more perilous environment to conduct verification.

The way to cope with this generally more hostile analysis environment involves improved analysis methods. One of the key elements in the analysis is contending with the lack of certainty about the solution, its nature and character mathematically. **For this reason the knowledge and guarantees about the results is missing. For instance we do not know what order of convergence to reasonably expect from the analysis and cannot use this to screen our results.** Generally speaking, if the verification result shows convergence at the theoretical rate for the method we can be sure we are solving problems that have smooth regular solutions either by character or sufficient mesh resolution. Usually the applied problems that modeling and simulation are attacking are mathematically difficult and may not be practically amenable to computational resolution sufficient to assure regularity. Philosophically, the whole reason for modeling and simulation is solving problems that are beyond our analytical grasp. In a deep sense the complex and difficult character to problems is unavoidable for the practical a use of modeling with computers. When we have successfully attacked the problem of verification for a problem without an exact solution, the same analysis methodology can improve our code verification practice.

[3] Quoted from http://www-history.mcs.st-and.ac.uk/Quotations/Bott.html.

It is important to understand solution verification within the broader context of computational modeling. **Solution verification contributes to the overall enterprise of analysis uncertainty quantification.** The most classical investigation will involve comparing the modeled results with observations in the real World (ideally an experiment). There are many elements to the uncertainty in this case including the model parameters, the constitutive properties, the experimental measurements and the numerical solution. Solution verification is the process for examining and estimating the numerical error and specifying its uncertainty. Sometimes this is applied in the use of computational modeling for purposes of decision-making or scenario testing where no real World data exists. In this case the numerical error is an important element in the overall lack of certainty about the results. If the numerical error is well behaved it will be a bias from the exact continuum solution to the model. This bias is important to understand in how it might skew the results and any advise.

When one lays out the mathematical framework for solution verification, the immediate impression is an added difficulty compared to code verification. This is due to the lack of direct knowledge of the precise solution. The full solution to the problem is inferred from the inaccurate numerical solutions. The equation to solve is the following $S_0 = S_k + Ch_k^a$ where the new unknown is the ostensible estimate of the exact solution S_0 that is the solution where $h = 0$. The solutions used to determine this estimate are S_k the solutions found with h_k. **We notice that we have three unknowns, S_0, C, a meaning that the well-determined solution requires three pieces of determined data, S_k.** As we will discuss, this problem can be solved in a variety of ways including under-, fully and overdetermined forms. (Here the problem, is over- (or under-)determined if we have more (or less) S_k than for a well-determined problem.) We also note that another option when more data is available would be to expand the error ansatz to include more degrees of freedom.

One of the key issues to recognize with solving this problem is an aspect of complexity because of the general nonlinearity of the determination of the model. The solution to this coupled system of nonlinear equations is generally subtle, and necessarily solved numerically. As such, the solution can have its own errors requiring some care and verification. The system of equations admits a simple analytical solution in special cases where the discrete solutions use a sequence of meshes where $r = h_k/h_{k-1}$ is constant. In this case we can write the solution in closed form $\log(E_{1,2}/E_{2,3})/\log(r)$, where $E_{k,k-1} = S_k - S_{k-1}$. More generally, we need to attack this with a coupled nonlinear solve. If we deal with an overdetermined version of the problem we will use a nonlinear least squares solver (or this is the knee-jerk response at least). One could also seek to modify or enrich the error ansatz to utilize the extra degrees of freedom. As I discuss next, following the nonlinear least squares path opens the door to some more interesting and robust choices (Huber 1996).

The general overdetermined version of the solution verification equation (i.e., more than three grids) would be amenable to solution via nonlinear least-squares method. **This is not the only choice, and consideration of this opens the door to other choices. The solution to the overdetermined problem is not unique, and the solution has the imprint of the method of solution.** As such the choice of least squares implies a number of explicit assumptions that the typical practitioner does not even know they are making. For example, one may choose to solve the overdetermined problem in a different norm than the two norm (i.e., least squares) (Bjork 1996). One may choose to solve a constrained problem instead of an unconstrained problem. In addition, one could consider solving an under-determined problem adding either constraints or regularizing the solution. A classical example of regularization is the Tikhonov method where a penalty is added to make the problem well determined. A popular recent approach focuses on a similar regularization, but in the one norm (compressed sensing, LASSO, ...) (Rider et al. 2016; Tibshirani 1996).

There are several practical issues related to this whole thread of discussion. One often encountered and extremely problematic issue is insanely high convergence rates. After one has been doing verification or seeing others do verification for a while, the analysis will sometimes provide an extremely high convergence rate. For example a second-order method used to solve a problem will produce a sequence that produces a seeming 15th order solution (this example is given later). This is a ridiculous and results in woeful estimates of numerical error. **A result like this usually indicates a solution on a tremendously unresolved mesh, and a generally unreliable simulation. This is one of those things that analysts should be mindful of. Constrained solution of the nonlinear equations can mitigate this possibility and exclude it *a priori*.** This general approach including the solution with other norms, constraints and other aspects is explored in (Rider et al. 2016). The key concept is that the solution to the error estimation problem is not unique and highly dependent upon many unwritten assumptions. Different assumptions lead to different results to the problem and can be harnessed to make the analysis more robust and impervious to issues that might derail it.

The techniques discussed in that paper were originally devised to deal with the all too common case where only one or two different grids are used and the error estimation problem is under-determined. The approach taken to solve this problem involves adding constraints to the solution based on expert knowledge and judgment. This problem was then approached when it was realized that the under- fully- and overdetermined cases should all be treated consistently. **The verification problem is solved repeatedly using different assumptions resulting in a natural variation in the results providing uncertainty in the error estimation and the rate of convergence. If the data is self consistent with a well-defined solution the uncertainty in the error will itself be small and the convergence rate will also be certain.** Conversely if the data is conflicting or opposes expert expectations, the uncertainty will be large. This entire methodology produces a more robust numerical uncertainty that adapts to the data, and avoids using fixed size safety factors. It turns out that this expert judgment is usually called into action with verification, but in an *ad hoc* manner and only when the issues are serious. So-called robust verification adds the expert judgment from the outset so that more subtle issues are subject to the same treatment.

Instead of solving the verification equation once using a nonlinear least-squares approach, robust verification solves the problem in a multitude of ways. This involves solving the verification problem using other error norms in a constrained minimiza-

Table 11.2 Computed solution verification error estimates for forward Euler (FE) for various step sizes. The same simple linear ODE is solved as in the code verification section

h	Solution, t = 1	Error, t = 1
0.20	0.3277	0.0402
0.10	0.3487	0.0192
0.05	0.3585	0.0094
0.02	0.3642	0.0037
0.01	0.3660	0.0018
Estimate	0.3678	±0.0002

tion framework. The data from a verification study is also used over. One standard assumption is that the solutions on the finer grids (smaller h) are closer to the exact solution, and this data is more prominent in the solution. The end result of the analysis is a multitude of estimates of the numerical error and convergence. **These results are then subjected to robust statistical examination using median statistics. We report the median of the estimates as the error and convergence rate. The median deviation is used to place an uncertainty on this estimate. One of the key benefits of this estimation is its lack of susceptibility to corruption by outliers in the analysis.** Outliers are further suppressed in the analysis by the use of expert judgment as constraints. For example, the absurdly large convergence rates are removed by the constraints if the rate of convergence is constrained to be below a given value.

Before moving to examples of solution verification we will show how robust verification can be used for code verification work. Since the error is known, the only uncertainty in the analysis is the rate of convergence. As we can immediately notice, this technique will get rid of a crucial ambiguity in the analysis. In standard code verification analysis, the rate of convergence is never the exact formal order, and expert judgment is used to determine if the results is close enough. **With robust verification, the convergence rate has an uncertainty and the question of whether the exact value is included in the uncertainty band can be asked.** Before showing the results for this application of robust verification, we need to note that the exact rate of verification is only the asymptotic rate in the limit of $h = 0$. For a finite step size the rate of convergence should deviate from this value and for simple cases the value can be derived using a modified version of classical numerical analysis.

Our first example of solution verification will repeat our examination of simple ODE integrators, but disregard our knowledge of the exact solution. The results of this study are given in Table 11.2 showing the solution and true error for different step sizes. I**t is a useful example because we can examine the efficacy of solution verification with a precise knowledge of the true errors.** We can use the data from our code verification study to good effect here. Here is the raw data used for the forward Euler study.

For the code verification part of the example, the estimated truncation error is $E = 0.2030h^{1.0245\pm0.0124}$ (the constant also has uncertainty, but its value has to be matched to the solution and convergence rate). The error bars do not take us to the theoretical convergence rate of one. The data is consistent with the rate being above one (and this is analytically expected). Using this same data for solution verification yields the following model, $S(h) = 0.3678 \pm 0.0002 - 0.2080h^{1.0386\pm0.0207}$. Close examination shows that this solution is quite close to the exact solution 0.0001 and within the error bars. If we use the standard techniques of simply least square fitting the data we get the following model, $S(h) = 0.3677 - 0.2239h^{1.0717}$. The error estimate here is 0.0017, which ends up being rather over generous when the standard safety factor of 1.25 is applied. **Using the robust verification technique we get a better estimate of the exact solution, the actual convergence rate and a tighter error bound.**

11.5 Solution Verification for Complex Problems

> Supposing is good, but finding out is better. – Mark Twain, Autobiography of Mark Twain, Vol. 3 (2015, (Twain 2015), p. 99)

It is also useful to look at a pathological case where the rate of convergence is absurd and standard analysis would be prone to missing it. The case we have at our fingertips involved very coarse grid solutions to large eddy simulation in a complex geometry relevant to heat transfer and fluid flow in nuclear reactors (Sagaut 2006). Early calculations were used to estimate the mesh required for well-resolved calculations. **As we found out, this is a perilous enterprise. A couple of codes (one production and one research) we enlisted in this study used some initial grids that were known to be inadequate. One of the codes was relatively well trusted for this class of applications and produced three solutions that for all appearances appeared reasonable.** An example solution is shown in Fig. 11.1 (Rider et al. 2010). One of the key parameters is the pressure drop through the test section. Using grids 664 K (664,000), 1224 K and 1934 K elements we got pressure drops of 31.8 kPa, 24.6 kPa and 24.4 kPa respectively. Using a standard curve fitting for the effective mesh resolution gave an estimate of 24.3 kPa ± 0.0080 kPa for the resolved pressure drop and a convergence rate of 15.84. This is an absurd result and needs to simply be rejected immediately. Using the robust verification methodology on the same data set, gives a pressure drop of 16.1 kPa ± 13.5 kPa with a convergence rate of 1.23, which is reasonable. **Subsequent calculations on refined grids produced results that were remarkably close to this estimate confirming the power of the technique even when given data that was substantially corrupted.**

Our final example is a simple case of validation using the classical phenomena of vortex shedding over a cylinder at a relatively small Reynolds number. Solution verification should be an integral part of any proper validation exercise as part of a systematic identification of modeling uncertainty. This is part of a reasonable

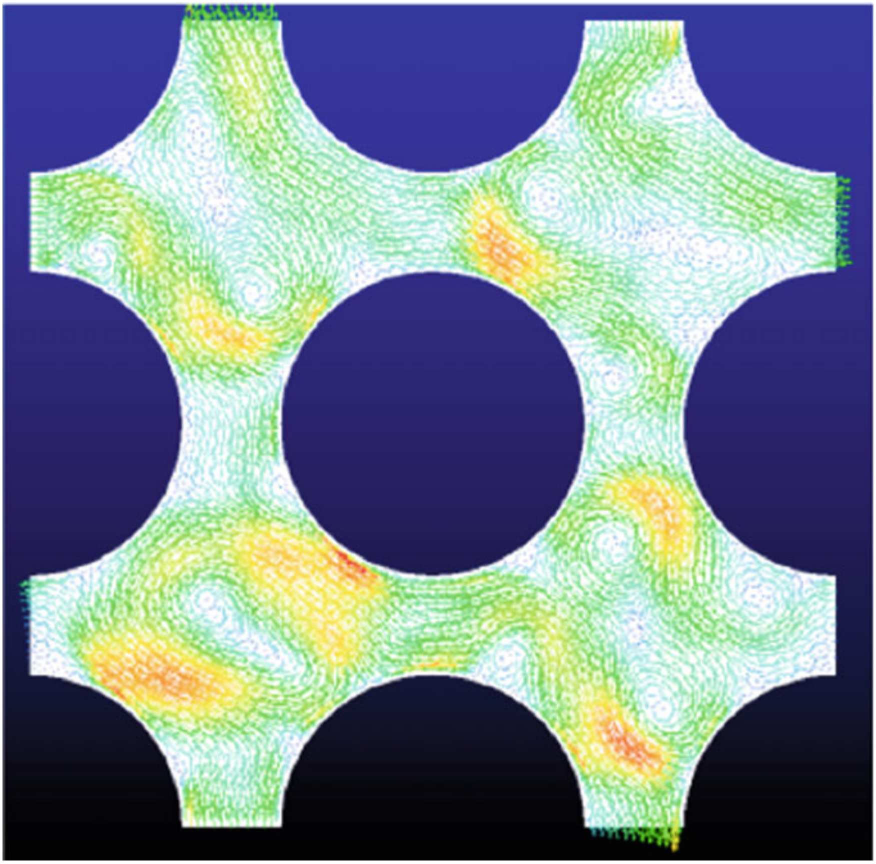


Fig. 11.1 A cross section of the flow in a unit cell for the large eddy simulation considered here

effort to validate a research code before using in on more serious problems. The key experimental value to examine is the Stouhal number defined, $St = f\ell/U$ the shedding frequency f normalized by the size ℓ of the cylinder and the velocity U, which has the value experimentally of 0.164 ± 0.005 for a flow of Reynolds number 100 (the Reynolds number is the non-dimensional ratio of inertial to viscous force in a flow). In Fig. 11.2 I show the calculations and the raw data for one of the calculations used (Pawlowski et al. 2006; Lin et al. 2006). Some of the raw data for shedding frequency is provided in Table 11.3.

When we apply the robust verification methodology to this data I find that the code produces a Strouhal number that is slightly larger than the experimental value $St(h) = 0.1657 \pm 0.0013 + Ch^{1.8486 \pm 0.1476}$. Including error bars recovers the experimental value within the bounds. This can be regarded as a modest success for the codes ability to be considered for more complex flows.

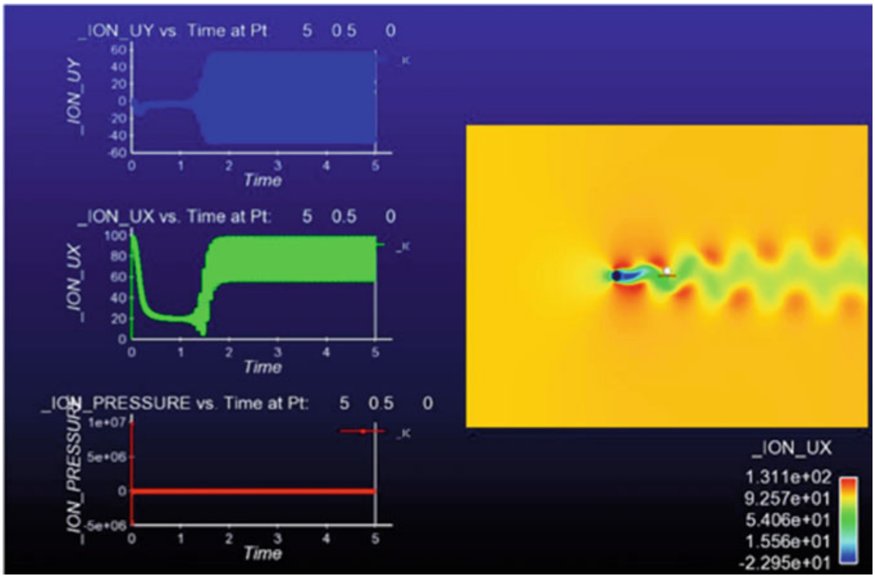


Fig. 11.2 A frame of the solution quantities for the vortex shedding example including the solution and shedding data generated

Table 11.3 Computed Strouhal number estimates for vortex shedding for mesh sizes

Δt	RMS h	St
0.002	0.054111988	0.110474853
0.002	0.023801688	0.152492294
0.002	0.010786082	0.164777976
0.002	0.005264375	0.165127187

11.6 Conclusion and Prospectus

> The foundation of data gathering is built on asking questions. Never limit the number of hows, whats, wheres, whens, whys, and whos, as you are conducting an investigation. A good researcher knows that there will always be more questions than answers. – Karl Pippart III, Operation Mexico! (Pippart 2015)[4]

In this chapter I have examined the foundational aspects of verification. I have provided the common core of mathematical expressions of numerical error used to define both the process of code and solution verification. In each case we have given examples of the practice. Code verification is reliant upon exact solutions whereas solution verification does not have this luxury. I have discussed the numerous ways code verification can be accomplished. Code verification acts to confirm that a numerical

[4] Here quoted after https://www.goodreads.com/author/quotes/14129212.Karl_Pippart_III.

approximation is implemented correctly by compiling mathematical evidence of correctness. Solution verification produces evidence of the magnitude of the numerical error, and the well-behaved nature of the systematic approximations present in the modeling. **Both practices are grounded on the same model for numerical error and involve solving problems on a sequence of numerical grids to determine the coefficients in the error model. Together these two practices provide a cornerstone for credibility of a modeling and simulation activity.** Ultimately the foundation of verification is grounded on asking hard questions about numerical methods used in simulation and collecting evidence about the answers.

Acknowledgements The author would like to thank Tim Trucano, Vince Mousseau, Patrick Knupp, Bill Oberkampf, Chris Roy, Greg Weirs, and Patrick Roache for numerous discussions and informative suggestions. The DOE ASC V&V program has provided support for the development of the verification methodology. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, National Technology and Engineering Solutions of Sandia, LLC. Operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLCDE-AC04-94AL85000.

Appendix

Details of Code Verification Documentation

Below, I briefly describe the characteristics of each type of benchmark documentation (could be called artifacts or meta-data) associated with a code verification benchmarks. These artifacts take a number of concrete forms such as a written document, computer code, mathematical solution in document or software form, input files for executable codes, input to automatic computer analysis, output from software quality systems, among others. While all of these will not typically exist, a subset of these documents should be available to support any questions or concerns about the verification problem or its implementation.

- Detailed technical description of the benchmark (report or paper): This can include a technical paper in a journal or conference proceeding describing the benchmark and its solution. Another form would be a report informal or formal from an institution providing the same information.
- Analysis of the mathematics (report or paper): For any solution that is closed form, or requiring a semi-analytical solution, the mathematics must be described in detail. This can be included in the paper (report) discussed previously or in a separate document.
- Computer analysis of solution (input file): If the mathematics or solution is accomplished using a computerized analysis, the program used and the input to the program should be included. Some sort of written documentation such as a manual for the software ideally accompanies this artifact.

- Computer solution of the mathematical solution: The actual computerized solution of the mathematical problem should be included in whatever form the computerized solution takes. This should include any error analysis completed with this solution.
- Computer implementation of the numerical solution: The analytical solution should be implemented in a computational form to allow the comparison with the numerical solution. This should include some sort of error analysis in the form of a report.
- Derivation of the source term and software implementation or input: In the case of the method of manufactured solutions, the source term used to drive the numerical method must be derived through a well-defined numerical procedure. This should be documented through a document, and numerical tools used for the derivation and implementation.
- Computer implementation of the source term (manufactured solution): The source term should be included in a form amenable to direct use in a computer code. The language for the computer code should be clearly defined as well as the compiler and computer system used.
- Grids for numerical solution: If a solution is computed using another simulation code all relevant details on the numerical grid(s) used must be included. This could be direct grid files, or input files to well-defined grid generation software.
- Convergence and error estimation of numerical solution: The numerical solution must include a convergence study and error estimate. These should be detailed in an appropriately peer-reviewed document.
- Uncertainty and sensitivity study of numerical solution: The various modeling options in the code used to provide the numerical solution must be examined vis-a-vis the uncertainty and sensitivity of the solution to these choices. This study should be used to justify the methodology used for the baseline solution.
- Description and analysis of computational methods: The methods used by the code used for the baseline solution must be completely described and analyzed. This can take the form of a complete bibliography of readily available literature
- Numerical analysis theory associated with convergence: The nature of the convergence and the magnitude of error in the numerical solution must be described and demonstrated. This can take the form of a complete bibliography of readily available literature.
- Code description/manuals: The code manual and complete description must be included with the analysis and description.
- Input files for benchmarks and auxiliary software: The input file used to produce the solution must be included. Any auxiliary software used to produce or analyze the solution must be full described or included.
- Unusual boundary conditions (inflow, piston, outflow, Robin, symmetry,): Should the benchmark require unusual or involved boundary or initial conditions, these must be described in additional detail including the nature of implementation.
- Physics restrictions (boundary layer theory, inviscid, parabolized Navier–Stokes,): If the solution requires the solution of a reduced or restricted set of equations, this

must be fully described. Examples are boundary layer theory, truly inviscid flow, or various asymptotic limits.

- Software quality documents: Of non-commercial software used to produce solutions, the software quality pedigree should be clearly established by documenting the software quality and steps taken to assure the maintenance of the quality.
- Scripts and auxiliary software for verification: Auxiliary software or scripts used to determine the verification or compute error estimates for a software used to produce solution should be included.
- Source code: If possible the actual source code for the software along with instructions for producing an executable (makefile, scripts) should be included with all other documentation.
- A full mathematical or computational description of metrics used in error analysis and evaluation of solution implementation or numerical solution.
- Verification results including code version, date, and other identifying characteristics: The verification basis for the code used to produce the baseline solution must be included. This includes any documentation of verification, peer-review, code version, date completed and error estimates.
- Feature coverage in verification: The code features covered by verification benchmarks must be documented. Any gaps where the feature used for the baseline solution are not verified must be explicitly documented.

Here are the necessary data requirements for each category of benchmark, again arranged in order of increasing level of documentation required. For completeness each data type would expected to be available to describe a benchmark of a given type.

- Common elements for all types of benchmarks (it is notable that the use of proper verification using an analytical solution results in the most compact set of requirements for data, manufactured solutions also).

  1. Paper or report
  2. Mathematical analysis
  3. Computerized solution and input
  4. Error and uncertainty analysis
  5. Computer implementation of the evaluation of the solution
  6. Restrictions
  7. Boundary or initial conditions

- Closed form analytical solution

  1. Paper or report
  2. Mathematical analysis
  3. Computerized solution and input
  4. Error and uncertainty analysis
  5. Computer implementation of the evaluation of the solution
  6. Restrictions
  7. Boundary or initial conditions

- Manufactured Solution

1. Paper or report
2. Mathematical analysis
3. Computational solution and input
4. Error and uncertainty analysis
5. Computer implementation of the evaluation of the solution
6. Derivation and implementation of the source term
7. Restrictions
8. Boundary or initial conditions

- Numerical solution with analytical solution
- Series solution, Nonlinear algebraic solution, Nonlinear ODE solution

1. Paper or report
2. Mathematical analysis
3. Computerized solution and input
4. Error and uncertainty analysis
5. Computer implementation of the evaluation of the solution
6. Input files
7. Source code
8. Source code SQA
9. Method description and manual
10. Restrictions
11. Boundary or initial conditions

- Highly accurate numerical solution (not analytical), numerical benchmarks or code-to-code comparisons.

1. Paper or report
2. Mathematical analysis
3. Computational solution and input
4. Error and uncertainty analysis for the solution
5. Computer implementation of the evaluation of the solution
6. Input files
7. Grids
8. Source code
9. Source code SQA
10. Method description and manual
11. Method analysis
12. Method verification analysis and coverage
13. Restrictions
14. Boundary or initial conditions.

References

Ascher, U. M., & Petzold, L. R. (1998). *Computer methods for ordinary differential equations and differential-algebraic equations*, vol. 61. Siam.

Banks, J. W., Aslam, T., & Rider, W. (2008). On sub-linear convergence for linearly degenerate waves in capturing schemes. *Journal of Computational Physics*, *227*(14), 6985–7002.

Bjork, A. (1996). *Numerical Methods for Least Squares Problems*. SIAM.

Brown, L. M., & Feynman, R. P. (2000). *Selected papers of Richard Feynman: With commentary* (Vol. 27), World scientific series in 20th century physics Singapore: World Scientific.

Drikakis, D., & Rider, W. (2006). *High-resolution methods for incompressible and low-speed flows*. Springer Science & Business Media.

Eça, L., Hoekstra, M., Hay, A., & Pelletier, D. (2007). On the construction of manufactured solutions for one and two-equation eddy-viscosity models. *International Journal for Numerical Methods in Fluids*, *54*(2), 119–154.

Galli, G., & Pasquarello, A. (1993). First-principles molecular dynamics. In *Computer simulation in chemical physics* (Springer, pp. 261–313).

Gartling, D. K. (1990). A test problem for outflow boundary conditions-flow over a backward-facing step. *International Journal for Numerical Methods in Fluids*, *11*(7), 953–967.

Gottlieb, J., & Groth, C. P. (1988). Assessment of Riemann solvers for unsteady one-dimensional inviscid flows of perfect gases. *Journal of Computational Physics*, *78*(2), 437–458.

Hairer, E., Nørsett, S. P., & Wanner, G. (1993). *Solving ordinary differential equations*. I, volume 8 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin.

Huber, P. (1996). *Robust Statistical Procedures*. SIAM.

Kamm, J. R., Brock, J. S., Brandon, S. T., Cotrell, D. L., Johnson, B., Knupp, P., et al. (2008). *Enhanced verification test suite for physics simulation codes*. Los Alamos National Laboratory (LANL), Los Alamos, NM: Tech. rep.

Kim, J., & Moin, P. (1985). Application of a fractional-step method to incompressible navier-stokes equations. *Journal of Computational Physics*, *59*(2), 308–323.

Kotschenreuther, M., Dorland, W., Beer, M., & Hammett, G. (1995). Quantitative predictions of tokamak energy confinement from first-principles simulations with kinetic effects. *Physics of Plasmas*, *2*(6), 2381–2389.

Lamb, H. (1932). *Hydrodynamics*. Cambridge: Cambridge University Press.

Lavery, D., & Burkhead, C. (Eds.). (2011). *Joss Whedon: Conversations*. Jackson: University Press of Mississippi.

Lax, P. D., & Richtmyer, R. D. (1956). Survey of the stability of linear finite difference equations. *Communications on Pure and Applied Mathematics*, *9*(2), 267–293.

Le, H., Moin, P., & Kim, J. (1997). Direct numerical simulation of turbulent flow over a backward-facing step. *Journal of Fluid Mechanics*, *330*, 349–374.

Lin, P. T., Sala, M., Shadid, J. N., & Tuminaro, R. S. (2006). Performance of fully-coupled algebraic multilevel domain decomposition preconditioners for incompressible flow and transport. *International Journal for Numerical Methods in Engineering*, *67*, 208–225.

Moin, P., & Mahesh, K. (1998). Direct numerical simulation: a tool in turbulence research. *Annual Review of Fluid Mechanics*, *30*(1), 539–578.

Moser, R. D., Kim, J., & Mansour, N. N. (1999). Direct numerical simulation of turbulent channel flow up to re τ= 590. *Physics of Fluids*, *11*(4), 943–945.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge: Cambridge University Press.

Pawlowski, R. P., Shadid, J. N., Simonis, J. P., & Walker, H. F. (2006). Globalization techniques for Newton-Krylov methods and applications to the fully coupled solution of the Navier-Stokes equations. *SIAM Review*, *48*(4), 700–721.

Pippart, K. (2015). *Operation Mexico!: Carl Kiekhaefer vs the 1951–1953 Pan American Road Race*. Minneapolis: Mill City Press.

Prabhakar, V., & Reddy, J. (2006). Spectral/hp penalty least-squares finite element formulation for the steady incompressible Navier-Stokes equations. *Journal of Computational Physics*, *215*(1), 274–297.

Rider, W., Witkowski, W., Kamm, J. R., & Wildey, T. (2016). Robust verification analysis. *Journal of Computational Physics*, *307*, 146–163.

Rider, W. J., Kamm, J. R., & Weirs, V. G. (2016). Procedures for calculation verification. *Simulation Credibility, 31*.

Rider, W. J., Kamm, J. R., Weirs, V. G., & Cacui, D. G. (2010). *Verification, validation and uncertainty quantification workflow in CASL.* Albuquerque, NM: Sandia National Laboratories.

Roache, P. (1998). *Verification and validation in computational science and engineering*. Hermosa Publishers.

Roache, P. (2009). *Fundamentals of verification and validation*. Hermosa Publishers.

Sagaut, P. *Large eddy simulation for incompressible flows: An introduction*. Springer Science & Business Media, 2006.

Stern, F., Wilson, R. V., Coleman, H. W., & Paterson, E. G. (2001). Comprehensive approach to verification and validation of CFD simulations-part 1: Methodology and procedures. *Journal Fluids Engrng.*, *123*, 793802.

Taylor, G. I., & Green, A. E. (1937). Mechanism of the production of small eddies from large ones. *Proceedings of the Royal Society of London*, *158*(895), 499–521.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Twain, M. (2015). *Autobiography of Mark Twain. Vol. 3*. The Mark Twain papers. University of California Press, Berkeley. A publication of the Mark Twain Project of the Bancroft Library.

White, F. (1991). *Viscous fluid flow*. New York: McGraw-Hill Inc.

Chapter 12
The Method of Manufactured Solutions for Code Verification

Patrick J. Roache

Abstract Verification of codes that numerically approximate solutions of partial differential equations consists in demonstrating that the code is free of coding errors and is capable, given sufficient discretization, of approaching exact mathematical solutions. This requires the evaluation of discretization errors using known benchmark solutions. The best benchmarks are exact analytical solutions with a sufficiently complex solution structure; they need not be physically realistic since verification is a purely mathematical exercise. The Method of Manufactured Solutions (MMS) provides a straightforward and general procedure for generating such solutions. For complex codes, the method utilizes symbolic manipulation, but here it is illustrated with simple examples. When used with systematic grid refinement studies, which are remarkably sensitive, MMS can produce robust code verifications with a strong completion point.

Keywords Manufactured solutions · Simulation · Benchmark · Verification · Turbulence · Convergence · Symbolic manipulation

12.1 Introduction

We are concerned in this chapter only with simulation models that are based on discretization of partial differential equations (PDEs). This covers most of classical physics, broadly defined, as well as some models in economics, ecological systems, and other disciplines of basic and applied science and engineering.

General code verification was defined rather tersely by the IEEE three decades ago (IEEE 1991) as “Formal proof of program correctness.” This definition has stood the test of time, but arguably benefits from expanded description; e.g., see Roache (1998a, b, p. 26 ff.) A definition specific to PDE codes in the context of computational solid mechanics was given in ASME (2006, p. 23): “the process of

P. J. Roache (✉)
Consultant, 1215 Apache Dr., Socorro, NM 87801, USA
e-mail: hermosa@sdc.org

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_12

determining that the numerical algorithms are correctly implemented in the computer code and of identifying errors in the software." See also ASME (2009), Oberkampf and Roy (2010), and Chaps. 2, 3, 5, and 11 in this volume. Generally, we find that such legalistic definitions tend to be sterile and/or inadequate, and that expanded descriptions are more useful. Basically, a code should do what the code manual says it does. See discussions throughout Chap. 2 of Roache (1998a, b).

For verification of PDE codes, we use a restricted definition of code verification, being concerned only with the ability of the code to produce mathematically accurate answers when sufficient discretization resolution is used. (This can contrast with computer science concepts of code verification that might include Quality Assurance issues that may have no effect on accuracy.) Determining this restricted sense of code correctness can only be accomplished by systematic discretization convergence tests using a "benchmark" solution which is preferably exact but at least reliable.

Once the verification of the code has been established, one can solve a specific problem which, if it is nontrivial, does not have an available exact solution. Verification of the computational *solution* then involves error *estimation*, since the benchmark solution is not known, whereas verification of the *code* involved error *evaluation* from a known benchmark solution. Both verifications (of code and solution) are purely mathematical activities, with no concern whatever for the accuracy of physical laws. That is the concern of validation, i.e., the agreement of the mathematics with observational science. In this view (see Chap. 27 by Roache in this volume), validation of computer simulations requires the three distinct activities referred to collectively as Verification and Validation (V&V): verification of the code, verification of the solution, and validation. For reasons both logical and practical, these activities should be performed in this order (Roache 2009; see also Chap. 42 by Beisbart in this volume).

The best benchmark solution is an exact analytical solution, i.e., a solution expressed in simple primitive functions like *sin*, *exp, tanh*, etc. Benchmark solutions involving infinite series are not desirable, typically being more numerical trouble to evaluate accurately than the PDE code itself (Roache 2009). It is not sufficient that the benchmark solution be exact; it is also necessary that the solution structure be sufficiently complex that all terms in the governing equation being tested are well exercised. Some early and misleading claims of accuracy of commercial codes for computational fluid dynamics (CFD) were based on comparisons with Poiseuille, Couette or Rayleigh problems, which do not even activate the advection terms. Many papers and reports have approached verification of codes in a haphazard and piecemeal way, comparing single-grid results for a few exact solutions on problems of greatly reduced complexity.

The Method of Manufactured Solutions (MMS) provides a systematic and general procedure for generating analytical solutions for code verification. The methodology provides for convincing, robust verification of a code via systematic discretization convergence testing. This procedure is straightforward though tedious to apply, and verifies all accuracy aspects of the code: formulation of the discrete equations (interior and boundary conditions) and their order of accuracy, the accuracy of the solution procedure, and the user instructions.

For early history and references, see (Roache 2002, 2009). The first systematic exposition of the method with application to multidimensional nonlinear problems appears to be (Roache and Steinberg 1984, Steinberg and Roache 1985) with expanded applications in Roache (2002, 2009). Acceptance was slow and misunderstanding was not uncommon, even by senior researchers. Now, MMS is regarded by many V&V specialists as the "gold standard" for PDE code verification, but still, it can be difficult to understand on first exposure. Based on experience of colleagues and myself, including teaching short courses, the misunderstanding seems due to the deceptive simplicity (elegance?) of the concept. Although applicable to high dimensional complex problems, the MMS concept is best described with simple examples in one space dimension (1D).

In what follows, Sect. 12.2 will first present the basic idea of MMS for generating exact benchmark solutions. Section 12.3 will illustrate this process using three simple examples. Then Sect. 12.4 will detail the application of benchmark MMS solutions to code verifications. The remaining sections present features and further examples of MMS.

12.2 Broad Description of MMS

The basic idea of the MMS procedure is to simply manufacture an exact solution, without being concerned about its physical realism. The "realism" or lack thereof has nothing to do with the mathematics, and verification is a purely mathematical exercise. In the original, most straightforward and most universally applicable version of the method, one simply includes in the code a general source term and uses it to generate a nontrivial but known solution structure. We follow the classic counsel of Polya (1957): *Only a fool starts at the beginning; the wise one starts at the end.*

We first pick a continuum solution. Interestingly enough, we can pick a solution almost independent of the code and the hosted equations (using a little prudence). That is, we can pick a solution, then use it to verify an incompressible Navier–Stokes code, a Darcy flow in porous media code, a heat conduction code, an electrode design code, a materials code, etc.

We want a solution that is nontrivial but analytic, and that exercises all ordered derivatives in the error expansion and all terms, e.g., cross-derivative terms. MMS can handle discontinuities (see below) but for this broad description, we consider smooth solutions. For example, chose a solution involving *tanh*. This solution also defines boundary conditions, to be applied in any (all) forms, i.e., Dirichlet, Neumann, Robin, etc. Then the solution is passed through the governing PDEs to give a source term that produces this solution. (This description sounds circular, which relates to difficulties with acceptance.)

12.3 Three Example Problems in MMS

To emphasize the generality of the concept, we pick the first example solution *before we specify the governing equations*. Then, we will use this same solution for two different problems, i.e., sets of governing PDEs and boundary conditions. The chosen solution *U(t,x)* is the following.

$$U(t, x) = A + \sin(B), \; B = x - Ct \tag{12.1}$$

12.3.1 Example 1

First, we apply this 1D transient solution to the nonlinear Burgers equation. (This equation is often taken as a model for CFD algorithm development but it is adequate to describe MMS for a wide range of PDE problems.)

$$u_t = -uu_x + \alpha u_{xx} \tag{12.2}$$

Incidentally, the specified solution *U(t,x)* is the exact solution for the constant velocity advection equation $u_t = -Cu_x$ with boundary condition $u(t,0) = A + sin(-Ct)$, so for high Reynolds number problems (small α) it may look "realistic" in some sense, but it is not a solution to our governing Eq. (12.2), and its "realism" or lack thereof is irrelevant to the task of code verification.

We determine the source term *Q(t,x)* which, when added to the Burgers equation for *u(t,x)*, produces the solution *u(t,x)* = *U(t,x)*. We write the Burgers equation as an operator (nonlinear) of *u*,

$$L(u) \equiv u_t + uu_x - \alpha \; u_{xx} = 0 \tag{12.3}$$

Then, we evaluate the Q that produces U by operating on U with L.

$$\begin{aligned} Q(t, x) &= L(U(t, x)) \\ &= \partial U/\partial t + U \partial U/\partial x - \alpha \; \partial^2 U/\partial x^2 \end{aligned} \tag{12.4}$$

By elementary operations on the manufactured solution *U(t,x)* stated in Eq. (12.1), we obtain

$$Q(t, x) = -Ccos(B) + [A + sin(B)]\, cos(B) + \alpha sin(B) \tag{12.5}$$

If we now solve the modified equation

$$L(u) \equiv u_t + uu_x - \alpha \; u_{xx} = Q(t, x) \tag{12.6}$$

$$u_t = -uu_x + \alpha\ u_{xx} + Q(t, x) \tag{12.7}$$

with compatible initial and boundary conditions, the exact solution will be *U(t,x)* given by Eq. (12.1).

The initial conditions are obviously just *u(0,x)* = *U(0,x)* everywhere. The boundary conditions are determined from the manufactured solution *U(t,x)* of Eq. (12.1). Note that, we have not yet even specified the domain of the solution. If we want to consider the usual model $0 \leq x \leq 1$ or something like it, the same solution Eq. (12.1) applies, but of course, the boundary values are determined at the corresponding locations in x. Note also that we have not yet even specified the *type* of boundary condition. This aspect of the methodology has often caused confusion. Everyone knows that different boundary conditions on a PDE produce different solutions; not everyone recognizes immediately that the same solution *U(t,x)* can be produced by more than one set of boundary condition *types*. The following combinations of inflow (left boundary, e.g., $x = 0$) or outflow (e.g., $x = 1$) boundary conditions will produce the same solution *U(t,x)* over the domain.

Dirichlet—Dirichlet:

$$u(t, 0) = U(t, 0) = A + sin(-Ct),\ u(t, 1) = U(t, 1) = A + sin(1 - Ct) \tag{12.8}$$

Dirichlet—Outflow Gradient (Neumann):

$$u(t, 0) = A + sin(-Ct),\ \partial u \big/ \partial x(t, 1) = cos(1 - Ct) \tag{12.9}$$

Robin (mixed)—Outflow Gradient (Neumann):

The Robin boundary condition, $F = au + bu_x = c$ where a, b, and c are constants, is to be applied as a time-dependent condition at the left boundary, so $F(t, 0) = c$.

$$\begin{aligned} &a\,u + b\,u_x = c\,applied\,at\ (t, 0) \rightarrow \\ &\qquad\qquad \text{given}\,a\,\text{and}\,b,\ \text{select}\,c\ = a[A + sin(-Ct)] + b\,cos(-Ct) \\ &\partial u \big/ \partial x(t, 1) = cos(1 - Ct) \end{aligned} \tag{12.10}$$

For this time-dependent solution, the boundary values are time-dependent. It also will be possible to manufacture time-dependent solutions with steady boundary values, if required by the code.

12.3.2 Example 2

To further clarify the concept, we now apply the same solution to a different problem, choosing as the new governing PDE a Burgers-like equation that might be a candidate for a 1D turbulence formulation based on the mixing-length concept.

$$u_t = -uu_x + \alpha u_{xx} + 2\lambda\left[x(u_x)^2 + x^2 u_{xx}\right] \tag{12.11}$$

Writing the mixing-length model equation as a nonlinear operator of u,

$$L(u) \equiv u_t + uu_x - \alpha\ u_{xx} - 2\lambda[x(u_x)^2 + x^2 u_{xx}] = 0 \tag{12.12}$$

we evaluate the Q_m that produces U by operating on U with L_m= L from (12).

$$\begin{aligned} Qm(t,x) &= Lm(U(t,x)) \\ &= \partial U/\partial t + U\partial U/\partial x - \alpha\partial^2 U/\partial x^2 - 2\lambda\ [x(\partial U/\partial x)^2 + x^2\partial^2 U/\partial x^2] \end{aligned} \tag{12.13}$$

By elementary operations on the same manufactured solution *U(t,x)* stated in Eq. (12.1),we obtain

$$\begin{aligned} Q_m(t,x) = &- C\,cos(B) + [A + sin(B)]cos(B) + \alpha\,sin(B) \\ &- 2\lambda\left[x\,cos^2(B)\ -\ x^2 sin(B)\right] \end{aligned} \tag{12.14}$$

If we now solve the modified model equation

$$Lm(u) \equiv u_t + uu_x - \alpha\ u_{xx} - 2\lambda[x(u_x)^2 + x^2 u_{xx}] = Qm(t,x) \tag{12.15}$$

$$u_t = -uu_x + \alpha\ u_{xx} + 2\lambda[x(u_x)^2 + x^2 u_{xx}] + Qm(t,x) \tag{12.16}$$

with compatible initial and boundary conditions, the exact solution for this “turbulent” problem again will be *U(t,x)* given by Eq. (12.1), as it was for the previous “laminar” problem.

The same initial and boundary conditions and boundary values from the previous problem can apply, since these are determined from the solution, not from the governing PDE nor Q.

12.3.3 Example 3

We have shown how the same solution can be used as the exact solution to verify two different codes with different governing equations, with different source terms being created to manufacture the same solution. A third example will demonstrate the arbitrariness of the solution form. Rather than the somewhat realistic solution to a constant velocity advection equation given by Eq. (12.1), we consider the “unrealistic” but equally valuable solution as follows.

$$U_e(t,x) = \sin(t)\,e^x \tag{12.17}$$

Following the same procedure for the Burgers Eq. (12.2), we evaluate the terms in Eq. (12.4) from the solution U_e of Eq. (12.17) and obtain

$$Q_e(t,x) = \cos(t)e^x + [\sin(t)e^x]^2 - \alpha \sin(t)e^x \tag{12.18}$$

(arranged for readability rather than compactness). This, when added to Eq. (12.2), produces the manufactured solution Eq. (12.17) when compatible initial and boundary conditions are evaluated from Eq. (12.17).

12.3.4 Complex Problems

MMS is applicable to complex nonlinear systems of equations, such as full Navier Stokes in general non-orthogonal coordinates, provided that the code is capable (or modifiable) to treat source terms in each PDE. MMS has been used in finite element codes both at the global solution level and at the element level (basis functions). To test periodic boundary conditions, one simply chooses a periodic function for the MMS solution.

12.4 Application to Code Verification

Once a nontrivial exact analytic solution has been generated, by MMS or perhaps another method, the solution is now used to verify a code by performing systematic discretization convergence tests (usually, grid convergence tests) and monitoring the convergence as $\Delta \to 0$, where Δ is a measure of discretization: Δx, Δt in a finite difference (FDM) or finite volume (FVM) code, element size in a finite element (FEM) code, etc. The procedure has been described in Chap. 11 by Rider in this volume; also Roache (1998a, b, 2009), Oberkampf and Roy (2010).

The fundamental concept "order of convergence" is based on behavior of the error of the discrete solution. There are various measures of error, but in some sense, we are always referring to the difference between the discrete solution *f(Δ)* (or a functional of the solution, such as drag coefficient) and the exact continuum solution,

$$E = f(\Delta) - f_{exact} \tag{12.19}$$

The most fundamental requirement for *code verification* is that E → 0 as Δ → 0. In addition, we like to verify not only the *fact* of convergence but the *order* of convergence, ideally estimated a *priori* by analysis of the discretization methods used. By definition, for an order p method and for a well-behaved problem (exceptions are discussed in Roache 2009, Chaps. 6 and 8), the error in the solution E asymptotically as $\Delta \to 0$ will be proportional to Δ^p. This terminology applies to every mathematically consistent methodology: FDM, FVM, FEM, block spectral, pseudo-spectral, vortex-in-cell, etc., regardless of solution smoothness. Thus,

$$E = f(\Delta) - f_{exact} = C\,\Delta^p + H.O.T. \tag{12.20}$$

where HOT are higher order terms. We then monitor the numerical error as the grid is systematically refined. Thorough iterative convergence is required (see below). Successive grid halving is not required, just refinement. Theoretically (from Eq. 12.20), values of C= $E/\Delta\ ^{\mathrm{p}}$ should become constant as the grid is refined for a uniformly *p-th* order method ("uniformly" implying at all points for all derivatives). Formulaic details of the calculation of observed *p* from grid convergence testing and many examples are given in Roache (2009), Oberkampf and Roy (2010). If observed *p* is not ~ theoretical *p*, this may indicate a coding error, or it may indicate a limitation of the approximations in the analysis for theoretical *p*. In either case, the code is still useable and would be claimed as "verified" at the observed *p*. Confidence is greatly enhanced if observed *p* ~ theoretical *p*.

Roy (2001), Roy et al. (2000) showed how to treat mixed-order convergence, a long-standing and practical difficulty in grid convergence studies. Mixed-order behavior can arise from the use of first-order discretization for advection and second order for diffusion, or from the first-order convergence rate of nominally second-order methods applied to discontinuities. The procedure involves another grid level to evaluate *two* leading coefficients in the error expansion. Especially important, the papers demonstrate how non-monotonic convergence occurs from mixed-order methods in the non-asymptotic range without blaming nonlinearity. MMS can verify such mixed-order convergence.

Inadequate iterative convergence produces false-negative evaluations of observed p. The extrapolation implicit in the order calculation amplifies machine round-off errors, so the iteration error control is more demanding for evaluation of *p* than for the PDE solution itself. Unfortunately, a priori specifications of iterative convergence criteria (e.g., maximum allowable change of some solution metric over one iteration divided by the iteration relaxation parameter) are not reliable. The recommended procedure is to test the sensitivity of the code verification results (notably observed *p*) to the iterative convergence stopping criteria. Note that this difficulty is not specific to MMS but occurs with any calculation of observed *p*; in fact, widely chosen MMS solutions are less vulnerable than most classical solutions, as noted above. Also, note that (as many V&V specialists have warned), the default iteration stopping criteria used in commercial CFD codes are often highly inadequate.

This verification procedure detects all ordered errors *E,* i.e., $E \to 0$ asymptotically as $\Delta \to 0$. It will not detect coding mistakes that do not affect the answer obtained, e.g., mistakes in an iterative solution routine which affect only the iterative convergence rate. In the present view, these mistakes are not considered as code verification issues, since they affect only code efficiency, not accuracy. Note that such efficiency issues should not be a concern to regulatory agencies. Other esoteric mistakes that are difficult to detect are described in (Roache 2009, Chap. 8; Knupp and Salari 2003).

The procedure does not evaluate the adequacy of non-ordered approximations, e.g., distance to an outflow boundary, distance to an outer (wind tunnel wall-like) boundary, etc. The errors of such approximations (which, I claim, are not inherently "numerical") do not vanish as $\Delta \rightarrow 0$, hence are "non-ordered modeling approximations." The adequacy of these approximations must be assessed by sensitivity tests which may be described as "justification" exercises (Roache 2009).

When this systematic grid convergence test is verified for all point-by-point values, we have verified

- input routines
- any equation transformations (e.g., boundary fitted coordinates),
- the order of the discretization,
- the encoding of the discretization, and
- the accuracy (but not efficiency) of the matrix solution procedure.

This MMS technique was originally applied in Roache and Steinberg (1984), Steinberg and Roache (1985) to long Fortran code produced by Symbolic Manipulation methods. The original 3D non-orthogonal coordinate code contained about 1800 lines of dense Fortran. It would be impossible to check this by reading the source code, yet the MMS procedure verified the code convincingly. Round-off error was not a problem.

The technique of code verification by monitoring grid convergence is extremely powerful. Upon initial exposure to the technique, analysts are often negative about the method because they intuit that it cannot be sensitive enough to pick up subtle errors. After exposure to numerous examples, if they remain negative it is usually because the method is *excessively* sensitive, revealing minor inconsistencies such as first-order discretizations at a single boundary point in an elliptic problem that effects the *size* of the error very little (as correctly intuited) but still reduces the asymptotic rate of convergence to first *order* for the entire solution. For examples, see Roache (2009).

The fact that the MMS solution may bear no relation to any physical problem does not affect the rigor of the accuracy verification of codes. The only important point is that the solution (manufactured or otherwise) be nontrivial: it should exercise all the terms in the error expansion. The algebraic complexity may be something of a difficulty, but it is not insurmountable, and in practice has been easily handled using Symbolic Manipulation (SM) software packages. Using the source code writing capability of SM software, it is not even necessary for the analyst to look at the form of Q. Rather, the specification of the solution (e.g., *tanh* function) to the SM software results in some complicated analytical expression that can be directly converted by the SM software to a source code segment, which is then readily emplaced in a source code module (subroutine, function, etc.) that then is called in the code verification procedure. This "emplacement" can be performed by hand by the analyst without actually reading the complicated source code expressions, or can itself be automated in the SM software.

MMS has been applied successfully to nonlinear systems of equations, with separate Q's generated for each equation. Both steady (stationary) and unsteady manufactured solutions may be formulated. Nonlinearity is an issue only because of uniqueness questions; the source term complexity may be worse because of nonlinearity, but managing that is the job of the SM software. Nonuniqueness conceivably could be an issue because the code might converge to another legitimate solution other than the MMS solution, producing a false-negative code verification. In much experience, nonuniqueness has never been an issue.

In Steinberg and Roache (1985), we applied the procedure to the coupled nonlinear (quasi-linear) PDEs of an elliptic grid generation method for non-orthogonal coordinates; the MMS solution was a 3D analytical coordinate transformation or parametrization. All operations for source code were performed by SM, including development of Euler–Lagrange equations for variational grid generation and all discretizations (Steinberg and Roache 1986a, b, 1992).

Note that the MMS solution should be generated in the original ("physical space") coordinates *(x,y,z,t)*. Then the same solution can be used directly with various non-orthogonal grids or coordinate transformations.

MMS (in this basic Q form) requires that the code being verified must include accurate treatment of source terms. Many codes, including the most popular modern commercial and open-source PDE codes, are built with source terms included, and many algorithms allow trivial extension to include Q's. However, directionally split algorithms (e.g., Roache 1998b) involve complexities at boundaries, especially for non-orthogonal coordinates.

Also see Roache (2009) for the following topics: early applications of MMS concepts, discussions and examples of mixed first- and second-order differencing, small parameter (high Reynolds number) difficulties, economics of dimensionality, applications of MMS to 3D grid generation codes, effects of strong and inappropriate coordinate stretching, debugging with MMS, examples of many manufactured or otherwise contrived analytical solutions in the literature, approximate but highly accurate solutions (often obtained by perturbation methods) that can also be utilized in code verification, special considerations required for turbulence modeling and other problems with multiple scales, example of MMS code verification with a 3D grid-tracked moving free surface, code robustness, examples of the remarkable sensitivity of code verification via systematic grid convergence testing, and several methodologies for verification of solutions including the Grid Convergence Index (see also Chap. 11 by Rider in this volume).

12.5 Features and Examples of MMS Code Verification

12.5.1 Radiation Transport Codes

Pautz (2001) presented his experience applying MMS to a radiation transport code that uses 3D tetrahedral elements in space and discrete ordinates in the angular discretization. The author discovered coding mistakes in input routines and in discretization of certain boundary data. Second order convergence for norms and third-order convergence for average scalar flux were verified. A subtle aspect revealed is the requirement for consistent finite element weighting on the MMS source term, which is now a recognized issue. Based on the earlier 1D analysis in the literature, it was expected that all the examined quantities would exhibit third-order convergence, but the results of the MMS procedure demonstrated only second-order convergence for the norms in multidimensions.

Blackwell et al. (2009) applied MMS to enclosure radiation, verifying their non-rigorous theoretical analysis that indicated $p = 2$ in contrast to another analysis that indicated $p = 3$.

12.5.2 Nonhomogeneous and Nonlinear Boundary Conditions

An arbitrary MMS solution may have nonhomogeneous boundary conditions, e.g., $\partial u/\partial x \neq 0$. To use such manufactured solutions, the code would require the capability of treating boundaries with $\partial u/\partial x \neq 0$. This might be inconvenient, e.g., some codes have hard-wired no-slip conditions at a wall with $u = 0$, or $\partial u/\partial x = 0$. Rather than modify the code, some thought will produce MMS solutions with homogeneous boundary values. Fortunately, modern commercial and open-source PDE codes have this capability for general treatment of boundary conditions, which is also the feature that facilitates validation; see Roache (2004, 2009) and Chap. 27 by Roache in this volume.

The so-called "radiation" outflow conditions are usually linear and are already covered by the previous discussion. Nonlinear boundary conditions, e.g., simple vortex conditions at outflow, or true (physical) heat transfer radiation boundary conditions, are possible. It may be possible to select an MMS solution that meets the nonlinear boundary condition; otherwise, a source term would need to be used in the nonlinear boundary equations.

12.5.3 Shocks, Partitioning, and "Glass-Box" Verification

Shock solutions are treatable by the MMS, with additional considerations. The simplest approach is to verify the shock-capturing algorithms separately on inviscid benchmark problems such as oblique shock solutions, provided that shock curvature is not viewed as a major question. If it is, one may use attached curved shock solutions obtained by the method of characteristics and/or detached bow shock solutions obtained by the classical inverse method. Any shock-capturing algorithm based purely on geometric limiters will be oblivious to the source terms and should work without modification.

The assumption involved in this approach is that the option matrix of the code can be *partitioned* (Roache 2009). That is, the verification of the shock-capturing algorithm and coding will not be affected by later inclusion of viscous terms, boundary conditions, etc. Other option-partitioning assumptions will occur to the reader, such as: separated verification of a direct banded Gaussian elimination routine in a FEM code; verification of shock-capturing algorithms separate from nonideal gas effects; radioactive decay option (which is dimensionless) verified separately from the spatial discretization of flow equations. This partitioning approach requires the "black-box" verification philosophy to be modified to a "glass-box" (Oberkampf and Trucano 2002) in which some knowledge of code structure is required to justify the approach. Thus it will be more difficult to convince reviewers, editors, contract monitors, regulators, stakeholders, etc., that the approach is justified. The work savings can be enormous, of course, avoiding the factorial increase of complexity inherent in option combinations.

12.5.4 Shocks, Multiphase Flows, and Discontinuous Properties

Without using code partitioning, J. Powers and colleagues (Grismer and Powers 1996; see also Roache 2009 for additional references) pioneered convincing code verification for flows with shock waves. The benchmark solutions may involve asymptotic approximations in geometry and/or Mach number M, e.g., an analysis neglecting terms of order $\varepsilon = 1/M^2$. This approximation can be made very accurate by choosing high M, say $M \sim 20$. Note again the distinction of mathematics versus science; it is not a concern that the code being tested might be built on perfect gas assumptions that are not valid at such high M. This does not affect the mathematics of code verification; the code would not be applied at such high M when accuracy of the physics becomes important, during validation.

Woods and Starkey (2015) applied MMS to shocks and other discontinuities using an "integrative MMS approach" (contrasted to "differential MMS" herein) based on "intelligent subdivision of the integration domains" to obtain a rigorous, one-step verification procedure for shock-capturing codes.

Brady et al. (2012) applied MMS to multiphase flows which necessitate discontinuous properties at the interface, where careful evaluation of source terms is required. They also offer additional guidelines to help locate coding mistakes. MMS for multiphase flows were also considered by Choudhary et al. (2014).

Grier et al. (2014, 2015) treated discontinuous MMS solutions, focusing on numerical integration techniques to address the problem of evaluating source terms consistently in finite volume methods. FVM do not store solution values at the center of the cell but rather integrated average values, which will converge more slowly than expected to the MMS point values unless special care is taken in the integration; the discrepancy is aggravated by discontinuous MMS solutions. (Alternately, one might consider post-processing the MMS solution to produce cell integrated average values for direct comparison, using methods consistent with the FVM code. However, this would add another layer of processing, the details of which would depend on the FVM solution code algorithm.)

Appendix A of ASME (2009) contains an MMS heat conduction problem with discontinuous step change in conductivity and contact resistance.

12.5.5 Verification of Boundary Conditions

Bond et al. (2007) presented an exemplary study applying MMS to CFD code verification of boundary conditions, including insightful observations. The FEM code being verified solves Euler, Navier–Stokes, and RANS equations on skewed, nonuniform, unstructured 3D meshes. Particular emphasis was placed on verification of numerical boundary conditions: slip, no-slip (adiabatic and isothermal), and outflow (subsonic, supersonic, and mixed), and on code segments that calculate solution gradients, a nontrivial issue in hexahedral grids with high aspect ratios near boundaries. The more demanding $L\infty$ norm was used and recommended, as well as the usual L_1 and L_2 norms. Among many interesting results, one provided a particular caution regarding precision issues. The symbolic manipulation software used to generate source functions writes source code in double precision but with only single precision constants, which later corrupted the initial verification exercise. The authors recommended an additional criterion for claiming verification of double-precision accuracy; the relative errors should be smaller than the single precision limit. Another caution involves orientation of the outflow boundary in supersonic flow along a constant pressure surface, which might permit certain coding errors to go undetected. This difficulty arose due to an ambitious approach of building boundary condition values into the MMS solution, rather than treating them crudely with the source term. Especially noteworthy was the success of MMS is disclosing a weakness of the solution algorithm in regard to the partitioning of multiprocessors. The paper is also valuable for presenting anecdotal debugging history, rather than a simple "pass" evaluation.

Choudhary et al. (2016) also focused on MMS verification of various important boundary conditions for both compressible and incompressible CFD codes.

12.5.6 Unsteady Flows and Divergence-Free MMS

An illustration of MMS applied to unsteady flows was given by Eça and Hoekstra (2007b). For the 2D laminar flows, a general formulation was developed that allowed an analyst to specify an arbitrary continuous function that is incorporated into an analytical form for velocities which satisfy the incompressible continuity constraint (divergence-free) exactly. Likewise, nonslip and impermeability conditions are met exactly by the MMS. Two time dependencies were considered: an exponentially decaying solution and a periodic solution. The exercise verified the code, and additionally shed light iteration error.

Choudhary et al. (2016) also gave special attention to MMS solutions which identically satisfy the divergence-free velocity field for incompressible flows, and to curved boundaries.

12.5.7 Variable Density Flows; Combustion

Shunn et al. (2012a, b) used MMS for variable density PDE codes applicable to combustion problems. Issues included use of tabulated state properties and effects of sub-iterations in the time advancement, especially for problematical time-splitting methods.

12.6 Attributes of MMS Code Verification

12.6.1 Two Multidimensional Aspects

In the first 1D example problem (Sect. 12.3.1), we noted that the MMS solution, since it is analytic, can be applied over any range of the dependent spatial variable x. This feature extends to multidimensions, e.g., the same multidimensional analytic solution could be applied to flow problems of a rectangular cavity, a backstep, a wing, etc.

Also, multidimensional problems might require a little more thought to assure that all terms of the governing equations are exercised. For example, a manufactured solution of form *U(t,x,y) = F1(t) + F2(x) + F3(y)* will not be adequate to exercise governing equations containing cross-derivative terms such as $\partial^2 u/\partial x \partial y$ since these are identically zero no matter how complex are the F's.

12.6.2 Blind Study

Salari and Knupp (2000) exercised MMS in a blind study, in which one author (Knupp) deliberately introduced errors into a CFD code previously developed and verified by the other (Salari). Then the code author tested the sabotaged code with the MMS. This exercise was not performed on merely model problems, but on a full time-dependent, compressible and incompressible, Navier–Stokes code with plenty of options. In all, 21 cases were studied, including one "placebo" (no mistake introduced) and several that involved something other than the solution (e.g., wrong time step, post-processing errors). Several formal mistakes (not order-of-convergence errors) went undetected, as expected (Roache 2002, 2009). All ten of the code errors that would affect accuracy were successfully detected, as well as several less serious mistakes.

12.6.3 Burden of MMS and Option Combinations

An experienced reviewer (Rider 2018) has stated that MMS puts a rather large burden on the code development teams, and that the source terms for MMS are difficult, prone to error, and need a high degree of software quality work and extensive debugging to produce reliable results.

I acknowledge this burden. In my own experience, the burden first involves some up-front work of becoming adept at using Symbolic Manipulation software. Once achieved, this development and training time can be amortized over verifications of many codes. In my own experience, the indictment of "prone to error" applies more to traditional methods. And it is true that producing a reliable and general MMS solution that exercises all the relevant terms certainly involves more work than coding an already developed single traditional solution, but possibly not if the simplified solution must be developed. Also, if one considers the suite of traditional problems usually required, then the amount of work may be less using a single MMS solution.

This claim is especially justifiable when one considers the curse of large numbers of code option combinations, discussed in Roache (2009). Suppose the non-separable option combinations number 100. A single MMS solution could easily replace a suite of 10 highly simplified classical solutions, reducing the required number of expensive grid convergence calculations by an order of magnitude, from 1000 to 100.

It is my opinion that the MMS approach is especially useful, reducing bookkeeping and total workload, when applied to extensive option combinations during *regression verification* of code modifications. (In major computational research environments, routine regression code verification activities are sometimes performed daily.)

12.6.4 Code Verification for Commercial Codes

Since code verification should be accomplished by the code developer, a question arises. Should a user assume that a commercial, open-source, or government code is verified? Roy (2015) reassessed previous misgivings by several V&V specialists and reached the same pessimistic evaluation: generally, the answer is no. This should be surprising since, as Roy noted, code verification is arguably the most mature subtopic in V&V; the main code verification techniques have been around for decades. Even if the vendor has published code verification for option combinations of interest, the user is strongly advised to scrutinize the results carefully for details and well founded conclusions.

12.6.5 Code Verification with a Strong Completion Point

Simple problems (even trivial problems) often serve a purpose during code development, and the results are often considered as partial verification. But apparently, it is not widely recognized that, once a code (for a specific set of code option combinations) has been convincingly verified on a complex problem that exercises all terms in the governing equations, it is nearly pointless to continue verifying the code on simpler problems. I say "nearly" because the exercises still have some value as *confirmation* exercises (Roache 2009, Chap. 1). MMS provides a robust code verification and terminates. A code user who performs confirmations gains confidence in the code and in their ability to set up the code and interpret the results. Such code confirmation exercises are valuable as part of user training but should not be confused with robust code verification. Similarly, we recognize that simple classical problems (e.g., 1D linear wave propagation) are useful in algorithm development, in exploring algorithm and code characteristics, and in comparing the performance of different codes. In fact, these classical problems are *more useful than MMS* for these purposes, since the general MMS solutions are typically unrealistic and opaque. But these comparison exercises, though valuable, should not be confused with robust code verification. These simple problems are complementary to the MMS approach, but if the comparison is taken as "partial verification" this leads to unending activity and invites criticism of the basic concept and legitimacy of code verification.

"Code verification is *not* an ongoing exercise. Verification, as we have said, is an exercise in mathematics, not science. When one proves a theorem, the work is completed. Proving the formula for solution of a quadratic equation is not ongoing work. This is not to say that one could not have made an error in the proof of a theorem, nor that confirmation exercises… are not valuable in confidence-building. It is to say that code verification is a mathematical activity that in principle comes to a conclusion, e.g., a code is or is not 2nd-order accurate." (Roache 1998a, b, p. 28) For an alternative view on code verification without a strong completion point, see discussion in (Roache 2002, 2009).

12.6.6 Proof?

Does such thorough code verification deserve the term *proof*? This is another semantic question whose answer depends on the community context. Logicians, philosophers and pure mathematicians clearly view "proof" differently from scientists and engineers, with an often other-worldly standard. For example, Fermat's Last Theorem is easily demonstrable, but do such exercises constitute *proof*? Certainly not to a mathematician. Since some philosophers maintain that it is not possible even in principle to prove Newton's laws of gravity, they are not likely to accept the notion of proof of correctness of a complex code.

The notion of *proof* is at the heart of very important criticisms, not just of the subject MMS, but of the concepts of code verification and especially *certification* for controversial public policy projects (Roache 2009). One might agree with philosophers who maintain it is not possible to prove Newton's Laws, but would one be willing to cancel a public policy project (e.g., nuclear waste disposal) because the modeling used Newton's Laws? Presumably not, but stakeholders are willing to cancel such projects under the guise of unprovability of code correctness. Great harm is done when these standards for proof of philosophers, mathematicians or logicians are applied to down-to-earth science and engineering projects. If we accept such out-of-context standards then we cannot do anything, literally. For example, we have no proof of convergence for realistic systems because the Lax Equivalence theorem only holds for linear systems (Roache 1998b).

The word *proof* is itself a technical term, with different appropriate standards in logic, pure mathematics, applied mathematics, engineering, criminal law versus torts versus civil law (consider "beyond a reasonable doubt"), etc. The first definition in one dictionary for *proof* is "The evidence or argument that compels the mind to accept an assertion as true." In this sense, if not in a mathematical sense, one could claim that MMS can provide proof of code verification. I am unhesitating in claiming "convincing demonstration" and "robust verification" for the MMS approach.

For further discussion on the possibility of a useful theorem related to MMS, see (Roache 2002, 2009). For the extensive discussion of V&V issues related specifically to modeling of nuclear waste disposal, see Roache (1998a Appendix C). For extensive discussion on semantics of V&V in computational physics specifically related to Popper's philosophy, see Roache (2012) and Chap. 27 by Roache in this volume. For the discussion of some current issues in V&V, including climate modeling, see Roache (2016).

12.6.7 Mere Mathematics

Rider (2018) noted that "the truism that verification is a purely mathematical exercise often works against verification. This is often used as an excuse to diminish its priority in code development. For codes used for science and engineering saying that it's just math can be used to say it's not important. This is unfortunate, but needs to be acknowledged and dealt with head-on."

We might expect that model developers would want some assurance that the code actually solved their model correctly, maybe even before they compared results to validation experiments! Furthermore, many validation exercises do not compare point values of all solution variables but only solution functionals, e.g., total heat flux. Especially in such comparisons, it is possible to achieve satisfactory agreement at particular experimental set points (i.e., values of experimental parameters) even though the code may have nontrivial errors.

In such situations, model developers might claim successful validation of their model M1 but in fact the code may contain an error E1. The actual "model" that is "validated" is not M1 but some M2 = M1 + E1, where E1 is unknown to the developers. The result is a contradiction of a fundamental tenet of science: reproducibility. Other code developers who incorporate Model M1 correctly will not obtain the same results, for better or worse.

In recent history of fluid dynamics, it has been difficult to achieve the same results from different codes that ostensibly incorporate the same RANS turbulence model due to coding errors and to incomplete specification (documentation) of model details.

12.6.8 Irrelevance of Solution Realism to Code Verification

MMS generates solutions with no required concern for realism of the solution. Thus, acceptance requires that the judge recognize code verification as purely mathematical exercise. Physical realism and even realizability are irrelevant. Actually, there is no *requirement* that the MMS solution look unrealistic, and we can invent appealing solutions if necessary to satisfy managers, regulators, public stakeholders, etc. But it is worthwhile to understand that this "realism" is mere window dressing when we consider only the legitimacy of code verification per se. Solution realism is also risky in that it encourages a dangerous misconception, invites criticism and arguments about what constitutes "adequate realism" (surely a qualitative concept), and ostensibly justifies piecemeal and perpetual code verification exercises.

Furthermore, realistic solutions can actually be less desirable because often they only weakly exercise some terms, e.g., streamwise second derivatives in boundary layers. For the purpose of detecting ordered errors, it is best that the different solution terms in the governing equations be very roughly the same size. (An order of magnitude variation is not problematical.) As pointed out by Rider (2018), this is easier to control with wisely chosen unrealistic MMS solutions than with many classical solutions.

12.7 Reasons for Solution Realism in MMS

In spite of my claims above that MMS "solution realism" is irrelevant to legitimacy of code verification per se, it is also true that there are uses for realism, both inside and outside of code verification.

12.7.1 Realistic MMS in Code Verification of Glacial Ice Flow Modeling

Bueler et al. (2007) developed a realistic MMS solution to verify a code for solving glacial ice flows based on shallow (thin-film) ice approximations. Solution realism was important to gain acceptance at a time when the glaciology science community was skeptical of models and verifications. The 3D time-dependent model involves many difficult features: a free boundary, thermo-mechanical coupling between a highly nonlinear power law viscosity and the temperature distribution, and coupling between energy conservation and thin-layer mass conservation PDEs with integrals in the nonlinear PDE coefficients.

MMS was applied by starting with an exact solution to an isothermal ice model and then manufacturing a coupled exact solution from it (see Sect. 12.3.1). Solution realism aided interpretation of controversial temperature “spokes” in ice flows found by several investigators.

The paper contains highly detailed descriptions, unusual for an archive journal not devoted to V&V, of the implementation, advantages and disadvantages of the MMS procedures. The authors state that the glaciology community could substantially replace intercomparison of codes with true code verification using legitimate MMS exact solutions.

A subsequent model was described in Bueler and Brown (2009). Development of the University of Alaska—Fairbanks Parallel Ice Sheet Model continues and the open-source PISM code (www.pism-docs.org) has been widely used in climate modeling. The MMS verification procedure is built into the system and is used in daily regression code verifications.

12.7.2 Realistic MMS in Solution Verifications and Turbulence Models

MMS is applicable to code verification but not to solution verification per se. However, in devising *methods* for solution verification, MMS can play an important role in tuning empirical parameters for the classic Grid Convergence Index (GCI) method and variations (Roache 1993, 1998a, b, 2009). MMS has also contributed to estimation of errors due to incomplete iteration and outflow boundary conditions, and to evaluating solution adaptive grid generation methods (Eça and Hoekstra 2007b, 2009; Pelletier et al. 2004, Roache 2009). Many benchmark-quality solutions are required to achieve statistical significance, and each solution requires expensive brute-force discretization convergence computations. MMS solutions, if they are realistic, can be used to economically obviate the need for such expensive fine-grid solutions.

This approach is especially effective for evaluating turbulence models. However, it is far from a straightforward application of MMS. RANS turbulence models are especially difficult due to discontinuous switches, min/max functions, and strongly nonlinear terms. As noted by Eça et al. (2007a, b), in a typical RANS model, there are no linear terms!

Eça and Hoekstra (2007a) and Eça et al. (2007a, b) used realistic MMS to study wall-bounded turbulence in 2D separated flows using both 1 and 2 equation RANS models; not surprisingly, they showed that the RANS models were inadequate in the near-wall region. Eça et al. (2007a, b) published detailed MMS solutions for several RANS models in conjunction with the Lisbon V&V Workshops (Eça et al. 2009). The benchmark realistic solutions and MMS source codes for six RANS models are available at the University of Lisbon website (Eça 2006). See Roache (2009) for additional references on the Lisbon Workshops.

Pelletier et al. (2004) used realistic MMS to tackle two of these difficult problems at once, turbulence models and solution adaptive FEM mesh generation, in the simulation of impinging round jets. They used the k-ε turbulence model and manufactured solutions for turbulent kinetic energy, eddy viscosity, and velocity.

12.7.3 Realistic MMS in Singularity Studies

Sinclair et al. (2006) independently developed realistic MMS (termed Tuned Test Problems) to evaluate methods for treatment of singularities during grid convergence studies. The techniques developed automatically detect and distinguish between cases of TTP-specified power singularities, logarithmic singularities, or simply grids not yet in the asymptotic range. For a summary, see Roache (2009, Sect. 5.4.10.1).

12.7.4 Other Uses and Generation Methods for Realistic MMS

Oberkampf and Roy (2010, Sect. 6.4, p. 235) note other cases in which physically realistic exact MMS solutions are desired: assessing sensitivity of a numerical scheme to mesh quality, and evaluating the reliability of discretization error estimators, as well as judging the overall effectiveness of solution adaptation schemes (see above). They describe two main approaches to generating realistic MMS solutions: theory-based solutions (see Sects. 12.3.1 and 12.7.1), and the Method of Nearby Problems. (The latter does not produce a single global analytical solution and has not seen much use.)

12.8 Alternative Formulations and General References for MMS

The basic *inverse* concept of MMS is to complicate the original problem a little to manufacture an intended solution; source terms are most straightforward and universally applicable. Another approach to MMS developed by Knupp and Salari (2003) is applicable to variable coefficient problems, e.g., groundwater transport or heat conduction codes with variable properties. A solution is manufactured by solving inversely for the distribution of variable coefficients that produce it.

Doebling (2016) verified a Lagrangian hydrodynamics code using an old (Fickett and Rivard 1974) exact solution for detonation problems. It was not described as MMS, and does not use manufactured source terms. But by (p. 1) "judicious selection of the material specific heat ratio, the problem has an exact solution with linear characteristics."

Burg and Murali (2004, 2006) developed a "residual formulation of MMS". The manufactured exact solution sets the initial condition, and only one iteration is used to evaluate the residuals. The residuals contain information on p in the sense of a Taylor's series expansion. But this approach does not actually verify the observed accuracy of a code since no solution is produced. While somewhat helpful for identifying locations of coding errors, the approach is not convincing for robust code verification, in my opinion.

Other general expositions of MMS are given in Knupp and Salari (2003), Roy (2005), Pelletier and Roache (2006), Wang et al. (2009), Oberkampf and Roy (2010). Besides the library of MMS solutions for turbulence (Eça 2006) already cited, Malaya et al. (2013) have created a library of code verification solutions including MMS as well as analytical solutions.

12.9 Conclusion

The Method of Manufactured Solutions for code verification was often met early with skepticism, but is now widely accepted. MMS enables one to produce many exact analytical solutions for use as benchmarks in systematic discretization refinement tests, which tests are remarkably sensitive for code verification. The method is straightforward and, when applied to all option combinations in a code, can lead to robust code verification with a strong completion point. It eliminates the typical haphazard, piecemeal and never-ending approach of partial code verifications with various highly simplified traditional problems that still leave the user unconvinced. Although the method requires some up-front work to become adept at using Symbolic Manipulation software, once achieved, this training time can be amortized over verifications of many codes. Producing a reliable and general MMS solution that exercises all the relevant terms typically involves more work than a single traditional solution, but if one considers the suite of traditional problems often used, then the amount of work can be less using MMS. The MMS approach is especially useful and reduces the book-keeping and total workload when used for regression verification of code modifications affecting option combinations.

Acknowledgements I gratefully acknowledge help from C. Beisbart, L. Eça, P. Moin, W. L. Oberkampf, C. J. Roy, N. Saam, L. Shunn. and especially W. J. Rider.

References

ASME. (2006). *ASME V&V 10-2006. Guide for verification and validation in computational solid dynamics.*

ASME. (2009). *ASME V&V 20-2009. Standard for verification and validation in computational fluid dynamics and heat transfer*.

Blackwell, B., Dowding, K., & Modest, M. (2009). Cylindrical geometry verification problem for enclosure radiation. *Journal of Thermophysics and Heat Transfer, 23,* 711–715. https://doi.org/10.2514/1.39861.

Bond, R. B., Ober, C. C., Knupp, P. M., & Bova, S. W. (2007). Manufactured solution for computation fluid dynamics boundary condition verification. *AIAA Journal, 45*(9), 2224–2236.

Brady, P. T., Herrmann, M., & Lopez, J. M. (2012). Code verification for finite volume multiphase scalar equations using the method of manufactured solutions. *Journal of Computational Physics, 231,* 2924–2944.

Burg, C. O. E., & Murali, V. K. (2004). *Efficient code verification using the residual formulation of the method of manufactured solutions. AIAA Paper 2004-2628*, 34th AIAA Fluid Dynamics Conference, Portland, Oregon, June, 2004.

Burg, C. O. E., & Murali, V. K. (2006). The residual formulation of the method of manufactured solutions for computationally efficient code verification. *International Journal of Computational Fluid Dynamics, 20*(7), 2006.

Bueler, E., Brown, J., & Lingle, C. (2007). Exact solutions to the thermomechanically coupled shallow-ice approximation: effective tools for verification. *Journal of Glaciology, 53*(182), 499–516.

Bueler, E., Brown, J. (2009). Shallow shelf approximation as a "sliding law" in a thermomechanically coupled ice sheet model. *Journal of Geophysical Research*, *114*, F03008. https://doi.org/10.1029/2008jf001179.

Choudhary, A., Roy, C. J., Dietiker, J.-F., Shahnam, M. & Garg, R. (2014). Code verification for multiphase flows using the method of manufactured solutions, FEDSM2014-21608. In *Proceedings of the ASME 2014 4th Joint US-European Fluids Engineering Division Summer Meeting (FEDSM)*. Chicago, IL, August 3–7, 2014.16 M3.

Choudhary, A., Roy, C. J., Luke, E. A., & Veluri, S. P. (2016). Code verification of boundary conditions for compressible and incompressible computational fluid dynamics codes. *Computers & Fluids, 126,* 153–169.

Doebling, S. W. (2016). The escape of high explosive products: an exact-solution problem for verification of hydrodynamics codes. *Journal of Verification, Validation and Uncertainty Quantification, 1,* 041001–1–041001–13.

Eça, L. (2006). Workshop Website. http://maretec.ist.utl.pt/~maretec.daemon/html_files/CFD_workshops/Workshop_2006.htm.

Eça, L., & Hoekstra, M. (2007a). Evaluation of numerical error estimation based on grid refinement studies with the method of manufactured solutions. Report D72-42, MARIN, May 2007.

Eça, L., & Hoekstra, M. (2007b). *Code verification of unsteady flow solvers with the method of manufactured solutions*. Paper No. ISOPE-2007-565, International Society of Offshore and Polar Engineers.

Eça, L., & Hoekstra, M. (2009). Evaluation of numerical error estimation based on grid refinement studies with the method of manufactured solutions. *Computers and Fluids*, https://doi.org/10.1016/j.compfluid.2009.01.003.

Eça, L., Hoekstra, M., Hay, A., & Pelletier, D. (2007a). A manufactured solution for a two-dimensional steady wall-bounded incompressible turbulent flow. *International Journal of Computational Fluid Dynamics, 21,* 175–188.

Eça, L., Hoekstra, M., Hay, A., & Pelletier, D. (2007b). On the construction of manufactured solutions for one and two-equation eddy-viscosity models. *International Journal for Numerical Methods in Fluids, 54,* 119–154.

Eça, L., Hoekstra, M., Roache, P. J., & Coleman, H. (2009). *Code verification, solution verification and validation: An overview of the 3rd Lisbon Workshop*. AIAA Paper No. 2009-3647, 19th AIAA Computational Fluid Dynamics, San Antonio, Texas, June 2009.

Fickett, W., & Rivard, C. (1974). *Test problems for hydrocodes*. Los Alamos, New Mexico:Los Alamos Scientific Laboratory, Report No. LA-5479.

Grier, B., Alyanak, E., White, M., Camberos, J., & Figliola, R. (2014). Numerical integration techniques for discontinuous manufactured solutions. *Journal of Computational Physics, 278,* 193–203.

Grier, B., & Figliola, R. (2015). Discontinuous solutions using the method of manufactured solutions on finite volume solvers. *AIAA Journal, 53,* 2369–2378.

Grismer, M. J., & Powers, J. M. (1996). Numerical predictions of oblique detonation stability boundaries. *Shock Waves, 6,* 147–156.

IEEE. (1991). *IEEE standard glossary of software engineering terminology*, IEEE Std 610.12-1990, New York, IEEE.

Knupp, P., & Salari, K. (2003). *Verification of computer codes in computational science and engineering*. Boca Raon, FL: CRC Press.

Malaya, N., Estacio-Hiroms, K. C., Stogner, R. H., Schulz, K. W., Bauman, P. T., & Carey, G. F. (2013). MASA: A library for verification using manufactured and analytical solutions. *Engineering with Computers, 29,* 487–496.

Murali, V., Burg, C. O. E. (2002). *Verification of 2D navier-stokes codes by the method of manufactured solutions*. AIAA Paper 2002-3109, 32nd AIAA Fluid Dynamics Conference, St. Louis, June, 2002.

Oberkampf, W. L., & Trucano, T. G. (2002). Verification and validation in computational fluid dynamics. *AIAA Progress in Aerospace Sciences*.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and Validation in Scientific Computing*. Cambridge, UK: Cambridge University Press.

Pautz, S. D. (2001). Verification of transport codes by the method of manufactured solutions: The ATTILA experience. In *Proceedings of the ANS International Meeting on Mathematical Methods for Nuclear Applications, M&C 2001*. Salt Lake City, Utah, Sept 2001.

Pelletier, D., & Roache, P. J. (2006). Verification and validation of computational heat transfer. In W. J. Minkowycz, E. M. Sparrow, & J. Y. Murthy (Eds.), *Handbook of Numerical Heat Transfer* (2nd ed.). New York:Wiley.

Pelletier, D., Turgeon, E., & Tremblay, D. (2004). Verification and validation of impinging round jet simulations using an adaptive FEM. *International Journal for Numerical Methods in Fluids, 44,* 737–763.

Polya, G. (1957). *How to solve it, a new aspect of mathematical method*. Princeton, NJ: Princeton University Press.

Rider, W. J. (2018). Personal communication 5/5/2018.

Roache, P. J. (1993). A method for uniform reporting of grid refinement studies, ASME FED-Vol. 158. In I. Celik, C. J. Chen, P. J. Roache, & G. Scheurer (Eds.), *Quantification of uncertainty in computational fluid dynamics*. ASME Fluids Engineering Division Summer Meeting, Washington, DC, 20–24 June 1993, pp. 109–120.

Roache, P. J. (1998a). *Verification and validation in computational science and engineering*. Albuquerque, NM: Hermosa Publishers.

Roache, P. J. (1998b). *Fundamentals of computational fluid dynamics*. Albuquerque, NM: Hermosa Publishers.

Roache, P. J. (2002). Code verification by the method of manufactured solutions. *ASME Journal of Fluids Engineering, 114*(1), 4–10.

Roache, P. J. (2004). Building PDE codes to be verifiable and validatable. *Computing in science and engineering*. Special Issue on Verification and Validation, September/October 2004, 30–38.

Roache, P. J. (2009). *Fundamentals of verification and validation*, Hermosa Publishers, Albuquerque, NM, Ch. 3 and Appendix C.

Roache, P. J. (2012). *A defense of computational physics*. Albuquerque, NM: Hermosa Publishers.

Roache, P. J. (2016). Verification and validation in fluids engineering: some current issues. *ASME Journal of Fluids Engineering*. FE-16-1206. https://doi.org/10.1115/1.4033979.

Roache, P. J., & Steinberg, S. (1984). Symbolic manipulation and computational fluid dynamics. *AIAA Journal, 22*(10), 1390–1394.

Roy, C. J. (2001). *Grid convergence error analysis for mixed-order numerical schemes*. AIAA Paper 2001–2606, June 2001 (Anaheim).

Roy, C. J. (2005). Review of code and solution verification procedures for computational simulation. *Journal of Computational Physics, 205*(1), 131–136.

Roy, C. J. (2015). Code verification: past, present and future, keynote lecture. In *ASME V&V Symposium*, La Vegas, NV, 13 May 2015.

Roy, C. J., McWherter-Payne, M. A., & Oberkampf, W. L. (2000). Verification and validation for laminar hypersonic flowfields, AIAA 2000-2550, June 2000 (Denver).

Salari, K. & Knupp, P. (2000). Code verification by the method of manufactured solutions, SAND2000-1444, Sandia National Laboratories, Albuquerque, NM 87185, June 2000.

Shunn, L., Ham, F., & Moin, P. (2012a). Verification of variable-density flow solvers using manufactured solutions. *Journal of Computational Physics, 231*(9), 3801–3827.

Shunn, L., Ham, F., & Moin, P. (2012b). Verification of variable-density flow solvers using manufactured solutions. *Journal of Computational Physics, 231*(9), 3801–3827.

Sinclair, G. B., Beisheim, J. R., & Sezer, S. (2006). Practical convergence-divergence checks for stresses from FEA. In *Proceedings of the 2006 international ANSYS users conference and exposition*, 2–4 May 2006, Pittsburgh, PA.

Steinberg, S., & Roache, P. J. (1985). Symbolic manipulation and computational fluid dynamics. *Journal of Computational Physics, 57*(2), 251–284.

Steinberg, S., & Roache, P. J. (1986a). Variational grid generation. *Numerical Methods for Partial Differential Equations, 2,* 71–96.

Steinberg, S., & Roache, P. J. (1986b). Grid generation: A variational and symbolic-computation approach. In *Proceedings numerical grid generation in fluid dynamics conference*, July 1986, Landshut, W. Germany.

Steinberg, S., & Roache, P. J. (1992). Variational curve and surface grid generation. *Journal of Computational Physics, 100*(1), 163–178.

Wang, S. S. Y., Jia, Y., Roache, P. J., Smith, P. E., & Schmalz, R. A. Jr., (Eds.). (2009). *Verification and validation of 3D free-surface flow model*. ASCE/EWRI Task Committee.

Woods, C. N., & Starkey, R. P. (2015). Verification of fluid-dynamic codes in the presence of shocks and other discontinuities. *Journal of Computational Physics, 294,* 312–328.

Chapter 13
Validation Metrics: A Case for Pattern-Based Methods

Robert E. Marks

Abstract This chapter discusses the issue of choosing the best computer model for simulating a real-world phenomenon through the process of validating the model's output against the historical, real-world data. Four families of techniques are discussed that are used in the context of validation. One is based on the comparison of statistical summaries of the historical data and the model output. The second is used where the models and data are stochastic, and distributions of variables must be compared, and a metric is used to measure their closeness. After exploring the desirable properties of such a measure, the paper compares the third and fourth methods (from information theory) of measuring closeness of patterns, using an example from strategic market competition. The techniques can, however, be used for validating computer models in any domain.

Keywords Model validation · State Similarity Measure · Area Validation Metric · Generalized Hartley metric

13.1 Introduction

Validation of a computer model broadly means determining whether the model is behaving as expected, given the modeller's knowledge of the real-world phenomenon being modelled; validating can aid in the choice of the best model, as discussed below. This chapter uses the example of agent-based models. Agent-based computer simulations (or multi-agent systems) are a special case of computer simulations which model autonomous or semi-autonomous rule-based agents dynamically interacting out of

R. E. Marks (✉)
School of Economics, University of New South Wales, 6 Vincent Street, Balmain,
Sydney, NSW 2041, Australia
e-mail: robert.marks@gmail.com
URL: http://www.agsm.edu.au/bobm

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_13

equilibrium, for the purpose of observing the emergence of patterns of behaviour at the micro (agent) level or at a higher, macro (group) level which might not otherwise be predicted.[1] For agent-based models, validation poses special issues since the emergent behaviour of such models might be previously unobserved or unexpected. This chapter explains techniques of validation for such models, in particular, the choice of a validation metric.[2] But the metrics to be discussed below are applicable in principle to validation of computer models against observed data—time series or cross sectional—of applications in many fields in engineering, science, computer science or the social sciences. Indeed, any phenomenon in which one set of multivariate variables is compared against another, time series or cross section.[3]

Chapter 31 by Fagiolo et al. in this volume presents a clear overview of validation of agent-based simulation models.[4] They remark that there are many kinds of validation or validity: e.g. output validation, structural validation, theoretical validity, model validity and operational validity. The simulation model is an attempt to include the relevant variables in a mechanism to reflect the behaviour and hopefully to explain the phenomenon being examined. The phenomenon exhibits a certain (historical) behaviour; the simulation model can generate simulated behaviour. How closely the simulation model's behaviour reflects the observed behaviour is one measure of how well the simulation model reflects the phenomenon being modelled (output validation). Another is to identify the causal structures underlying the real-world phenomenon, as revealed in the historical data, and to compare them with the causal structures of the simulation model or models. This chapter focuses on output validation, asking how well do the model data track existing real-world data, possibly micro (at the agent level), possibly macro (at the aggregate level).[5]

In this chapter, four broad families of measures that can be used in this respect will be explained: what might be called *empirical likelihood measures*, so-called *stochastic area measures*, so-called *information-theoretic measures* and *pattern-based* or *strategic state measures*. There are trade-offs associated with these families of measures, and several metrics, so far, have been devised for each.[6]

[1] For an overview of types of computer simulation modelling, see Gilbert and Troitzsch (2005).

[2] We distinguish between broader *measure* and narrower *metric*—a metric is a measure, but a measure is not necessarily a metric—as discussed in Sect. 13.2 below.

[3] See Marks (2007), Midgley et al. (2007), Oberkampf and Roy (2010), and Liu et al. (2010) for further general discussions of validation.

[4] As Guerini and Moneta (2017) observe, the appearance of many measures to validate agent-based simulation models is an indication of "the vitality of the agent-based community."

[5] This chapter, in effect, focuses on techniques of output validation (see Chap. 30, Sects. 4.2, 5.1 and 5.2 by Fagiolo et al. in this volume), going into greater detail about three of the six measures they discuss.

[6] This chapter puts the work of Marks (2013) into a wider context.

In the fourth family, we describe in detail two metrics (the State Similarity Measure, SSM, and the Generalized Hartley Measure, GHM) which are applicable to validation of models the output of which is multivariate patterns, unlike other methods which assume univariate variables. The two measures can be thought of alternate methods of measuring the rowwise distance between any two matrices of equal dimension, **X** and **Y**.

13.2 Validation Metrics

As an example from the social sciences, consider the interactions over time among several brands, where each brand's market decisions (prices, promotions, etc.) in any period affect the other brands' volumes sold and profits, and the other brands respond with their own market decisions in the following period. (See Sect. 13.5.) This "rivalrous dance," as I have called it, generates a complex dynamical pattern of prices, profits, volumes sold, etc. The problem is not specific to simulation models and phenomena in the social sciences: researchers in the biological sciences face the same issue and have made some seminal advances in our understanding of the issues (Mankin et al. 1977).

13.2.1 Four Types of Measurement Scales

The variables compared in Ferson et al. (2008) and Roy and Oberkampf (2011), like almost all variables in scientific and engineering validation, share one property: they are *interval scales*. That is, they measure ordered magnitudes, defined so that the intervals of pairs of variables can be compared, or measured. (They could be *ratio scales*, such as Kelvin for temperature, with absolute zero and where ratios are meaningful,[7] but this is less common.)

Almost all validation methods in finance, science and engineering are applicable to interval-scaled variables, but not to *order-scaled variables*, in which the variables might be increasing (decreasing) in one (or the other) direction but where distances in these directions are meaningless because order is their highest characteristic. And such validation methods cannot be applied when the variables are *nominal scales* only: when their order is arbitrary, and their highest characteristic is unique identity, with arbitrary, separate names or numbers.

The main focus of this chapter is on methods of validation which can deal with nominal-scaled variables, or *patterns*, such as those that are seen in the historical phe-

[7] A temperature of 100K is twice as hot as 50 K, but 100 °C is not twice as hot as 50 °C: K is a ratio scale, but °C is only an interval scale ("by how much?"); "hotter" and "colder" is only an ordered scale.

nomena (and the computer programs written to simulate them) described in Sect. 13.5 below.

Two metrics in particular—the State Similarity Measure and the Generalized Hartley Measure—have been developed to deal with nominal-scaled data. These can be thought of as generalizations of the interval-scale-based metrics of Ferson et al. (2008). They also overcome an issue that does not arise with interval-scaled variables: the disappearance of any state in one but not the other of the two sets of matrices **X** and **Y**: interval-scaled measures do not exhibit gaps in which one state appears in **X** but not in **Y**, or vice versa.

13.2.2 The Desirable Properties of a Validation Metric

Ferson et al. (2008, p. 2415) state that "a validation metric is a formal measure of the mismatch between predictions [of the model] and data that have not previously been used to develop the model." And that the closer the match between the model output and the historical observations, the smaller the measure. Specifically, they argue that a desirable measure should exhibit six properties:

1. it should be objective (and quantitative) so that the same predictions and the same data will result in the same assessment, no matter who conducts it.
2. if there is a comparison between deterministic values without stochasticity, then the metric should generalize this in a reasonable way.
3. the metric should reflect all differences in the two distributions (of the predictions and of the history), not just the lower moments of these distributions (mean, standard deviation); it should not be too sensitive to outliers.
4. for ease of understanding, the unit of the metric should be the same as the unit of the variables, if possible.
5. the modulus of the measure should be unbounded above.
6. the measure should be a true *metric*: that is, it should be non-negative and symmetric, should satisfy the delta inequality:

$$d(x, y) + d(y, z) \geq d(x, z)$$

   and should satisfy the identity of indiscernables[8]:

$$d(x, y) = 0 \iff x = y.$$

Property 6 defines a metric. Properties 1, 3 and 6 are, I believe, crucial to any validation measure. Property 4 is desirable for interval-scaled variables and Property

[8] Lacking only symmetry, it is a quasi-metric; lacking only the identity of indiscernables, it is a semi-metric; lacking only the triangle inequality, it is a pseudo-metric.

2 is desirable for validation of stochastic models. Property 5 is not necessary and is inapplicable where the variables are order- or nominal-scaled.

The issue of measuring the distance between the dynamics of the output produced by a simulation model and the historical counterpart raises the question of how to define a metric to measure this distance. For simple phenomena (and simple models), the output might be simple too. Measuring the distance between two time series, say, is simple. When the phenomenon is dynamic and multivariate, with more than one interrelated time series output, however, the issue of defining and measuring the distance between the pair of sets of outputs is not simple.

If, moreover, the variables of the data and the model predictions are not interval-scaled, but only nominal-scaled, then the units of the measure will not in general be those of the data and predicted variables (Property 4). And it is not clear whether Properties 2 and 3 will be satisfied. First, in the application of oligopolistic pricing below, the historical data and the model predictions are deterministic. (The computer simulation is a deterministic model, mapping from market state, determined by curtailed historical data, to the next period's marketing actions—here, prices). It is not clear how to generalize this to a stochastic model, except perhaps by Monte Carlo simulations (Marks 2016). Second, the metrics we propose below are no more sensitive to (less frequent) data than they are to more frequent data: the tails are not too influential.

13.3 Four Families of Validation Measures

We can distinguish between four families of measures that are important in the context of validation. First, empirical likelihood measures; second, what might be called stochastic area measures; third, information-theoretic measures; and, fourth, strategic state measures that compare patterns of data.

13.3.1 Empirical Likelihood Measures

These measures include maximum likelihood, the generalized method of moments, the method of simulated moments and indirect inference (see Chen et al. 2012); to a greater or lesser extent, these demand knowledge of the true probabilistic dynamics of the models' output or require the use of assumptions about these dynamics.[9] But likelihood measures rely on summary statistics and do not explicitly compare the similarity of distributions or patterns between the data and the simulated data generated by the models.

[9] Guerini and Moneta (2017) present a new method of validation, based on comparing structures of vector autoregressive models estimated from both model and historical data.

In general, such measures satisfy Ferson et al.'s Properties 1, 4, 5 and 6, but, in only generating summary statistics, these measures ignore the information contained in the patterns, especially relevant in strategic, dynamic models; they do not satisfy Property 3. Moreover, such methods are usually seen not as validation methods, but as methods of calibration and estimation (see Chap. 31 by Fagiolo et al. in this volume, Sect. 31.4.1).

13.3.2 Stochastic Area Measures

These have been derived by Ferson et al. (2008) and Roy and Oberkampf (2011) and others. Specifically, these papers address models and observations with stochastic characteristics and univariate response quantities. That is, the model output Y and the observed data X are single random variables. Unfortunately, the generalization to multivariate responses is not straightforward.

Following Ferson et al. (2008), there are a variety of ways to compare univariate random variables, expressed as probability density functions (p.d.f.s) or cumulative distribution functions:

First, the random variables are "equal" or "surely equal" if their p.d.f.s are identical.

Second (more weakly), the random variables are "equal in mean" if the expectations of the absolute values of the differences between X and Y are zero.

Third (more weakly), if not quite equal in means, the *mean metric* provides a measure of their discrepancy

$$dE(X, Y) = E(|X - Y|) \neq |E(X) - E(Y)|,$$

where E is the expectation operator. This can be generalized to higher order moments of the distributions, where equality in the higher order moments implies equality in all lower order moments.

Fourth (more weakly), if the shapes of the distributions of the two variables are identical, then the random variables are "equal in distribution".

Fifth (more weakly), if the distributions are not quite equal in shape, there are many proposed measures, including the Kolmogorov–Smirnov distance:

$$dS(X, Y) = \sup_z |\Pr(X \leq z) - \Pr(Y \leq z)|,$$

which is the vertical distance between the cumulative distributions functions of the two random variables, where z takes on all values in the common range of historical observations X and model output Y. Other measures, such as the Kullback–Leibler divergence, are discussed below.

The variables compared in Ferson et al. (2008) and Roy and Oberkampf (2011), like almost all variables in finance, scientific, and engineering validation, share one property: they are interval scales. The Area Validation Metric (AVM) introduced by Ferson et al. (2008) can only be applied when the two variables are interval or ratio scales. The AVM measures the area between the cumulative distribution functions of the two random variables, that of the model predictions and of the historical data. The metric is not defined for ordered scales, in which the variables might be increasing (decreasing) in one (or the other) direction but where distances in these directions are meaningless because order is their highest characteristic.

And such interval-scale measures cannot be applied when the variables are nominal only: when their order is arbitrary, and their highest characteristic is unique identity, with arbitrary, separate names or numbers.

Indeed, the Smirnov distance is not applicable, even with an arbitrary ranking of the ordering of the states. But the the applications introduced below generate nominal-scale output, not interval-scaled.

Ferson et al.'s AVM, in measuring the divergence of the p.d.f. of the model output from the historical data, does take satisfy Property 3, but is limited in that it requires interval-scaled single variables of both output and observed data.

In what follows, we focus on methods that explicitly compare patterns in the data, both observed and simulated, and do not in general require interval-scaled variables. They are from the following two families.

13.3.3 Pattern-Based Measures I: Information-Theoretic Measures

Information-theoretic measures are derived from Shannon's measure of entropy (Shannon 1948), and include the Kullback–Leibler construct (Kullback and Leibler 1951), and more recent measures that attempt to overcome shortcomings of Kullback–Leibler, such as the *GSL-div* (Lamperti 2018a, b).

Such measures satisfy Properties 1, 3 and 5, but, as we discuss in Sect. 13.4 below, they do not in general satisfy Property 6, although that has not eliminated their use in model validation. I argue here that there are true metrics which should be considered instead.

13.3.4 Pattern-Based Measures II: Strategic State Measures

Strategic state measures include Marks' State Similarity Measure (Marks 2013) and Klir's 2006 Generalized Hartley Measure, from early set-theoretic work of Hartley's (Hartley 1928). These two measures satisfy Ferson et al.'s Properties 1, 2, 3 and 6, but not Property 4 (units of measurement), or Property 5 (the measures are bounded above); I argue that these two properties are not crucial for a validation metric.

13.4 Measures of Closeness or of Information Loss

Turn now to the third family of measures. The broad idea behind evaluating a distance between the model output and the real-world data in order to choose the model "closest" to the real-world data is as follows. If the real data are information full, then models of the underlying process capture only some of the information. Choosing the model that loses least information compared to historical data is the criterion for choosing the "best" model.

Information is often measured using Shannon entropy (1948) (SE).[10] It is based on probability and can be defined as

$$SE(p(x)|x \in X) = -\sum p(x)log_2(p(x))$$

where p is the probability distribution of random variable x. The function SE exhibits some useful properties such as additivity, branching, normalization and expansibility. Shannon entropy led to the Kullback and Leibler (1951) measure of information loss from historical to model; it has some attractions theoretically, but is not a true metric, as we shall see.

13.4.1 Kullback–Leibler Information Loss

The Kullback–Leibler (K-L) divergence or information loss (also known as relative entropy) provides a measure of the information lost when model g is used to approximate full reality f:

$$I(f, g) = \int f(x) \log\left(\frac{f(x)}{g(x|\theta)}\right) dx$$

in the continuous version, where the models g are indexed by θ, or

$$I(f, g) = \sum_{i=1}^{k} p_i \times \log\left(\frac{p_i}{\pi_i}\right)$$

in the discrete case, with full reality f distribution $0 < p_i < 1$, and model g distribution $0 < \pi_i < 1$, with $\sum p_i = \sum \pi_i = 1$. Here, there are k possible outcomes of the underlying process; the true probability of the ith outcome is given by p_i, while the $\pi_1, \ldots, \pi_k$ constitute the approximating model. Hence, f and g correspond to the p_i and π_i, respectively.

[10] Another measure used for information is Hartley information (see Sect. 13.7). Both are special cases of Rényi entropy (Rényi 1970). Both derive from work done at the Bell Labs.

But the K-L information loss is not a true metric: it is not symmetric and does not satisfy Property 6, since $I(f, g) \neq I(g, f)$.[11] Moreover, π_i must be positive for every i,[12] while in data, even for a coarse, dichotomous partition, this value is likely to be zero for some states, for either set of data (model predictions or real data).[13] As mentioned above, this is a stumbling block for the AVM technique of Ferson et al. (2008), although AVM is suitable for validation of models with univariate random variables for output and observations.

13.4.2 The Generalized Subtracted L divergence (GSL-div)

To overcome shortcomings of the Kullback–Leibler divergence, the symmetric L divergence (Lin 1991) was developed. From this, the *GSL-div* (Lamperti 2018b) has been derived to measure the degree of similarity between real and simulated dynamics by comparing the patterns of the time series. Lamperti discusses the procedure to obtain the *GSL-div*, and then presents results to discriminate among four different classes of stochastic processes. He also compares the *GSL-div* with alternative measures of fit (using several summary statistics) commonly used for calibrating ABMs, and concludes that *GSL-div* provides much more satisfactory performance at this. But neither K-L nor Lin's *L-div* (and hence *GSL-div*) satisfy Property 6, and, hence, are not proper metrics, despite the interesting properties of *GSL-div* (Lamperti 2018b).[14]

Let us now turn to the fourth family of measures, the strategic state measures, which include the author's State Similarity Measure (Sect. 13.7) (which uses rectilinear or Minkowski's L_1 or the cityblock distance), and Klir's Generalized Hartley Measure (Sect. 13.7). Both are true metrics. Before we present the measures, we describe the models for our example.

[11] It is a semi-quasi-metric.

[12] The K-L measure is defined only if $p_i = 0$ whenever $\pi_i = 0$.

[13] As Akaike (1973) first showed, the negative of K-L information is Boltzmann's entropy. Hence minimizing the K-L distance is equivalent to maximizing the entropy; hence the term "maximum entropy principle." But, as Burnham and Anderson (2002) point out, maximizing entropy is subject to a constraint—the model of the information in the data. A good model contains the information in the historical data, leaving only "noise." It is the noise (or entropy or uncertainty) that is maximized under the concept of the entropy maximizing principle. Minimizing K-L information loss then results in an approximating model g that loses a minimum amount of information in the data f. The K-L information loss is averaged negative entropy, hence the expectation with respect to f. Fagiolo et al. (2007, p. 211) note further that "K-L distance can be an arbitrarily bad choice from a decision-theoretic perspective ... if the set of models does not contain the true underlying model ... then we will not want to select a model based on K-L distance." This is because "K-L distance looks for where models make the most different predictions—even if these differences concern aspects of the data behaviour that are unimportant to us."

[14] Although, as (Lamperti 2018b) points out, so long as the simulated data are always compared with the historical data, and not with simulated data from other models, *GSL-div* might still allow model choice.

13.5 The Example: Models and Data

Return to our example of the interactions over time among several brands. We use three models from simulations described in Marks et al. (1995). Each model has three interacting brands, and each brand agent independently chooses its weekly price from its own set of four possible prices in order to maximize its weekly profit, in a process of co-evolution using the Genetic Algorithm (GA). With 1-week memory, each agent's action is determined by the state of the market in the previous week, which means $4^3 = 64$ possible market states for each agent to respond to. See results for 2- and 3-week memory below. The GA chooses the mapping from perceived state to action for each brand (with each brand's weekly profit as its "evolutionary fitness"). This means that the models are not derived from historical patterns of oligopolistic behaviour, and so can be used to predict these patterns.

Each model of the three brands' interactions corresponds to a separate run of the GA search for model parameters, using weekly profits of the brands as the GA "fitness". Given the complexity of the search space and the stochastic nature of the GA, each run "breeds" a distinct model, with distinct mappings from state to brand price, and hence different patterns of brand actions associated with each model.[15] Figures 1 and 3 of Midgley et al. (1997) and also of Marks (2013) show, respectively, the observed historical weekly prices and volumes sold of several brands of coffee competing in a U.S. supermarket chain, and a 50-week period of simulated interactions among three brand agents in Model A, where each brand chooses from one of four possible prices per week.

In order to reduce the number of degrees of freedom, we coarsen the partitioning of the data, using a dichotomous partition into High and Low prices for both the real data and the simulated data.

The distribution of the eight possible 1-week states in the historical chain store (H) with three brands or players and in three models (A, B and C) [16] of the models' outputs, using 50 weeks of data, are shown in Table 13.1, with "0" corresponding to a player's "High" price and "1" to a player's "Low" price.[17] Modelling deeper memory for the brands results in similar distributions, but the tables are 64 rows and 512 rows deep, with 2-week and 3-week memory, respectively, corresponding to 64 and 512 states.

[15] The three models differ in more than the frequencies of the eight states (Table 13.1): each model contains three distinct mappings from state to action, and, as deterministic finite automata (Marks 1992), they are ergodic, with emergent periodicities. Model A has a period of 13 weeks, Model B of 6 weeks, and Model C of 8 weeks. It is not clear that the historical data exhibit ergodicity, absence of which will make simulation initial conditions significant (Fagiolo et al. 2007). Initial conditions might determine the periodicity of the simulation model.

[16] In Midgley et al. (1997) and Marks (2013), Model A is called Model 26a, Model B is called Model 26b and Model C is called Model 11.

[17] Figures 2 and 3 of Marks (2013) plot these behaviours. State 000 corresponds to all three players choosing High prices; State 001 corresponds to Players 1 and 2 choosing High prices and Player 3 choosing a Low price, etc.

Table 13.1 State frequencies from history and three models

State	History	Model A	Model B	Model C
000	32	30	20	0
001	2	11	10	18
010	6	3	7	15
011	1	0	0	0
100	7	5	12	16
101	0	0	0	0
110	2	1	1	0
111	0	0	0	1
Total	50	50	50	50

The important thing to note here is that these are models of *strategic* interaction: it is not sufficient to examine a single brand's time series of actions, since these have affected—and in turn have been affected by—its rivals' actions over time. This is essentially a multivariate validation problem.

13.6 The State Similarity Measure (SSM)

Introduced in Marks (2010), the SSM counts the absolute difference in the frequency of each possible state in each of two sets of vectors (or time series), and sums these to obtain the SSM for the pair of sets of vectors. In effect, SSM treats each time series set as a vector $\mathbf{p}$ in an n-dimensional, non-negative, real vector space with a fixed Cartesian coordinate system, where there are n possible states in the sets of vectors. The SSM between two sets matrix $\mathbf{P}$ and matrix $\mathbf{Q}$ of vectors (or time series) is calculated as the rectilinear Minkowski's L_1 or cityblock distance (Krause 1986) d_1 between their two constructed vectors $\mathbf{p}$ and $\mathbf{q}$, given by

$$d_1^{\mathbf{PQ}} = d_1(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i|, \tag{13.1}$$

where p_i is the number of occurrences (or frequencies) of state i in vector set $\mathbf{P}$. That is, SSM is the sum of the absolute differences of the coordinates of the two sets of vectors as n-dimensional constructed vectors. (See Marks 2013, Appendix 1 for details of this procedure.)

As defined here, the SSM is an absolute measure, where its maximum distance D is a function of the equal length of the pair of sets of vectors. The lower the SSM, the closer the two sets of vectors.

The maximum D of an SSM measure occurs when the intersection between the states of the two sets of vectors is null, with $D = 2 \times S$, where S is the number of window states, which depends on the memory length, inter alia. In our example, maximum D would be 100 for 1-week memory, $2 \times 49 = 98$ for 2-week memory, and $2 \times 48 = 96$ for 3-week memory, (given that there are 50 observations per set of time series). It is possible to define a normalized measure.

13.6.1 Results for the Models

The six pairs of SSMs between the partitioned prices of the three models and the observed historical data, using 50-week data series, are presented in Table 13.2 for 1-, 2- and 3-week memory. Table 13.3 presents the distances between History, and the three simulations, Model C, Model A, and Model B from Marks et al. (1995), with 3-week memory. Model C is far from any of the other sets, and Model B is closest to Model A, but Model A is closer to the History historical data (at 54/96) than it is to the closest other simulation, Model B (at 60/96).

As the partitioning becomes finer (with deeper memory of past actions), the SSMs increase as the two sets of vectors (or time series) become less similar. This should not surprise us. We also note that with these four sets of time series, the rankings do not change with the depth of memory: (from closer to more distant) (History, Model A), (Model A, Model B), (History, Model B), (Model C, Model B), (Model C, Model A) and (History, Model C). Which of the three models is closest to the historical data of History? The SSM tells us that Model A is best, followed by Model B, with Model C bringing up the rear.

13.6.2 Monte Carlo Simulations of the SSM

We can, using Monte Carlo stochastic sampling (Marks 2016), derive some statistics to test whether any pair of sets is likely to include random series (see below).

As *Null Hypothesis* we choose: each of two sets of time series is random.

Table 13.2 SSMs calculated between the six pairs of sets

	Pair	1-week memory	2-week memory	3-week memory
b	History, Model A	18	36	54
f	Model A, Model B	22	42	60
c	History, Model B	28	48	68
e	Model C, Model B	42	60	80
d	Model C, Model A	62	76	88
a	History, Model C	70	88	92

Table 13.3 SSMs between observed history and three models

	History	Model A	Model B	Model C
History	0	54	68	92*
Model A	54	0	60	88*
Model B	68	60	0	80*
Model C	92*	88*	80*	0

With this null hypothesis, we can set 1% and 5% one-sided confidence intervals to the SSM numbers. (Note: * in Table 13.3 indicates we cannot reject the null at the 5% level.) With three brands and $S = 48$, the maximum D is 96. 95% of pairs of sets of three random time series are at least 80 apart, and 99% of pairs of sets of three random time series are at least 76 apart.[18] This means that, in Table 13.3, we reject the null hypothesis of random data for the pairs (History, Model A), (History, Model B), and (Model A, Model B), since all SSMs here are less than 76, so the data are significantly non-random, and the null hypothesis is rejected. The other three pairs (all comparisons with Model C), with SSMs above 80, are not significantly (5%) different from random, and the null hypothesis cannot be rejected. By construction, none of the simulated data sets is random, although they are not particularly similar (see Table 13.1). Figure 4 of Marks (2013) plots the Cumulative Mass Function (CMF) of the MC parameter bootstrap simulation against the six SSMs of the pairs.

13.7 Classical Possibility Theory

Possibility theory offers a non-additive method of assigning a numerical value to the likelihood of a system assuming a specific state, one of a given set of states. The likelihood expressed is that of *possibility*; for this reason, the possibility assigned to a collection of possible events is the maximum (rather than the sum) of the individual possibilities (Ramer 1989).

Hartley (1928) solved the problem of how to measure the amount of uncertainty associated with a finite set E of possible alternatives: he proved that the only meaningful way to measure this dichotomous amount (when any alternative is either in or out: no gradations of certainty) is to use a functional of the form:

$$c \log_b |E|,$$

[18] This number was determined by a Monte Carlo bootstrap simulation of 100,000 pairs of sets of four quasi-random time series, calculating the SSM between each pair, and examining the distribution. The lowest observed SSM of 64 appeared twice, that is, with a frequency of 2/100,000, or 0.002 percent.

where set E contains all possible alternatives from the larger (finite) set X, and where $|E|$ denotes the cardinality of set E: b and c are positive constants, and it is required that $b \neq 1$. If $b = 2$ and $c = 1$ (or more generally, if $c \log_2 = 1$), then we obtain a unique functional, H, defined for any basic possibility function, r_E, by the formula:

$$H(r_E) = \log_2 |E|,$$

where the measurement unit of H is bits. This can also be expressed in terms of the basic possibility function r_E as

$$H(r_E) = \log_2 \sum_{x \in X} r_E(x).$$

H is called a *Hartley measure* of uncertainty, resulting from lack of specificity: the larger the set of possible alternatives, the less specific the identification of any desired alternative of the set E. Clear identification is obtained when only one of the considered alternatives is possible. Hence, this type of uncertainty can be called *non-specific*.

This measure was first derived by Hartley (1928) for classical possibility theory, where any alternative element of set X is either possible (i.e. in set E) or not. The basic possibility function, r_E, is then

$$r_E(x) = \begin{cases} 0 & \text{when } x \in E, \\ 1 & \text{when } x \notin E. \end{cases}$$

and is derived explicitly in Klir (2006, pp. 28). To be meaningful, this functional must satisfy some essential axiomatic requirements.[19]

13.7.1 The Generalized Hartley Measure (GHM) for Graded Possibilities

Following Klir (2006), we relax the "either/or" characteristic of the earlier treatment and allow the basic possibility function[20] on the finite set X to take any value between zero and one: $r : X \to [0, 1]$. Note that

$$\max_{x \in X}\{r(x)\} = 1,$$

a property known as possibilistic normalization.

[19] See further discussion in Marks (2013), Appendix 2.

[20] It is not correct to call the function r a possibility *distribution* function, since it does not distribute any fixed value among the elements of the set X: $1 \leq \sum_{x \in X} r(x) \leq |X|$.

The Generalized Hartley Measure (GHM) for graded possibilities is usually denoted in the literature by U and is called *U-uncertainty*. U-uncertainty can be expressed in various forms. A simple form is based on notation for graded possibilities: $X = \{x_1, x_2, \ldots, x_n\}$ and r_i denotes for $i = 1, \ldots n$ the *possibility* of the singleton event $\{x_i\}$. Possibilities can (although need not) be estimated by frequencies. Elements of X are appropriately rearranged so that the possibility profile:

$$\mathbf{r} = \ < r_1, r_2, \ldots, r_n >$$

is ordered in such a way that

$$1 = r_1 \geq r_2 \geq \ldots \geq r_n > 0,$$

where $r_{n+1} = 0$ by convention. Moreover, the set $A_i = \{x_1, x_2, \ldots, x_i\}$ is defined for each $i \in \{1, \ldots, n\}$.

Using this simple notation, the U-uncertainty is expressed for each given possibility profile $\mathbf{r}$ by the formula

$$U(\mathbf{r}) = \sum_{i=2}^{n} (r_i - r_{i+1}) \log_2 i \tag{13.2}$$

Klir (2006, p. 160) notes something relevant to our purposes here: "Another important interpretation of possibility theory is based on the concept of *similarity*, in which the possibility $r(x)$ reflects the degree of similarity between x and an ideal prototype, x_P, for which the possibility degree is 1. That is, $r(x)$ is expressed by a suitable distance between x and x_P defined in terms of the relevant attributes of the elements involved. The closer x is to x_P according to *the chosen distance*, the more possible we consider x to be in this interpretation [our emphasis]."

13.7.2 Applying U-Uncertainty to Our Data

From the frequencies of Table 13.1 (one-week memory), we can reorder[21] the possibilities (observed frequencies) of the three runs and the historical data, to get the four reordered, non-normalized[22] possibility profiles (Table 13.4):

[21] It might be objected that this reordering loses information. But this overlooks the fact that the order of the states is arbitrary. It should not be forgotten that the definition of the states with more than 1 week's memory captures dynamic elements of interaction.

[22] Normalization here means $r_1 = 1$, not $\sum r_i = 1$.

Table 13.4 The four possibility profiles, one-week memory

History:	32	7	6	2	2	1	0	0
Model A:	30	11	5	3	1	0	0	0
Model B:	20	12	10	7	1	0	0	0
Model C:	18	16	15	1	0	0	0	0

Using Eq. (13.2), the four Hartley measures are calculated[23]:

1. History:

$$U(\mathbf{r}) = \frac{1}{32}(25\log_2 1 + 1\log_2 2 + 4\log_2 3 + 0\log_2 4 + 1\log_2 5 + 1\log_2 6)$$

$$= 0.383$$

2. Model A:

$$U(\mathbf{r}) = \frac{1}{30}(19\log_2 1 + 6\log_2 2 + 2\log_2 3 + 2\log_2 4 + 1\log_2 5)$$

$$= 0.516$$

3. Model B:

$$U(\mathbf{r}) = \frac{1}{20}(8\log_2 1 + 2\log_2 2 + 3\log_2 3 + 6\log_2 4 + 1\log_2 5)$$

$$= 1.054$$

4. Model C:

$$U(\mathbf{r}) = \frac{1}{18}(2\log_2 1 + 1\log_2 2 + 14\log_2 3 + 1\log_2 4)$$

$$= 1.399$$

The GHMs for the three models and History have been calculated for the three cases of 1-week, 2-week and 3-week memory, as seen in Table 13.5.

These GHMs are true metrics (they satisfy Property 6, unlike the K-L information loss), and so we can compare the differences of Table 13.6 between the four measures. We can readily see that Model A (0.516) is closest to the historical data

[23] For clarity, we have included the $(i = 1)$th element, $(r_1 - r_2)\log_2 1$, which is always zero, by construction, consistent with Eq. (13.2).

Table 13.5 GHMs calculated for three memory partitions

Process	1-week memory	2-week memory	3-week memory
History	0.383	0.495	0.782
Model A	0.516	0.679	1.085
Model B	1.054	1.657	2.542
Model C	1.399	2.179	2.787

Table 13.6 GHM differences calculated for the six pairs of sets

	Pair	1-week memory	2-week memory	3-week memory
b	History, Model A	0.133	0.184	0.303
e	Model C, Model B	0.345	0.522	0.245
f	Model A, Model B	0.538	0.978	1.457
c	History, Model B	0.671	1.162	1.760
d	Model C, Model A	0.883	1.500	1.702
a	History, Model C	1.016	1.684	2.005

of History (0.383); next is Model B (0.516), with Model C (1.399) furthest from the Historical data. Moreover, we can see that Model A is closer to the Historical data than it is to Model B.

Table 13.6 shows the six pairwise differences in GHM, derived from Table 13.5. It can be compared with the six pairwise SSMs of Table 13.2.

For 1-week memory the maximum GHM, corresponding to 50 equi-likely states, is $\log_2 50 = 5.644$; for 2-week memory $\log_2 49 = 5.615$, and for 3-week memory $\log_2 48 = 5.585$. These numbers are the maximum pairwise difference between GHMs; the minimum difference is zero in all three depths of memory.[24]

13.8 Comparing the Distances Measured by SSM and GHM

From Table 13.2, for 1-week memory, the SSMs are ranked (closest to farthest): {b, f, c, e, d, a}; but, from Table 13.6, the GHM differences are ranked (smallest to largest): {b, e, f, c, d, a}. Model A is closest to History using either measure, and Model C is farthest. Note, however, from Table 13.2, that although the SSM rankings are the same for 1-, 2- or 3-week memory, the GHM rankings are sensitive to the depth of memory (see Table 13.6). That is, the two methods do not always produce identical

[24] We could also define a normalized GHM.

rankings, although the degree to which these two measures result in similar rankings of distances is noteworthy, given their quite different foundations.[25]

13.9 Conclusions

Is a particular computer model the best model of a particular real-world phenomenon? “Best” can have several meanings, but here we mean whether the behaviour (“output”) of the simulation model is closest to the observed behaviour of the phenomenon. Measuring the closeness of the simulated behaviour and the observed (historical) behaviour might be simple (for example, for univariate, interval-scaled, deterministic variables) or not (for example, for multivariate output of nominal-scaled variables). Measuring this closeness is necessary to validate any model, and can be used to choose the best model of set of contenders.

We have examined the appropriateness of measures from four families of techniques, as characterized by the kinds of output observed and generated. Using Ferson et al.’s “desirable properties” of validation metrics, and focusing on the kind of phenomenon (oligopolistic, strategic interactions among sellers) which exhibits multivariate, nominal-scaled behaviour, we have argued that two contenders—SSM and GHM—are appropriate.

These two strategic measures, SSM and GHM, are true metrics that allow us to measure the degree of similarity between two sets of vectors (or matrices **X** and **Y**), here multivariate time series. The SSM between two sets of vectors is the absolute distance between two constructed vectors in non-negative, n-dimensional vector space, where n is the number of possible states that each set of vectors can exhibit. GHM is a measure of the possibility of any set **P** of vectors occurring as a vector **p** in n-dimensional space.

Since GHM is a metric, differences of sets of vectors’ GHMs are meaningful. SSM is also a metric (satisfying Property 6). As such, both measures can be used to score the distance between any two sets of vectors, such as sets of time series, which previously was unavailable.

The SSM and GHM strategic state measures have demonstrated closeness in measuring similarity of sets of time series, although the two measures’ rankings of distances are not identical, as seen above. The SSM is intuitive: it uses the cityblock metric to tally the differences in the states between two constructed vectors. It can be described in six simple steps, as outlined in Marks (2013), Appendix 1. The GHM is anything but intuitive, based on arcane possibility theory.

Using Occam’s Razor, the SSM, as a simpler, more transparent measure, is preferred.

The two strategic state measures, SSM and GHM, are not restricted to measuring the similarity of (or distance between) two sets of time series: they are more general, as we have reminded the reader, in that they can be applied to pairs of sets of (equal

[25] Exploration of these differences awaits further research.

length) vectors. The data used here are illustrative only: the two measures can be applied to any pairs of simulated data and historical data, so long as the number of observations of the model output and the historical data are equal, with equal numbers of vectors, or observations. Even more generally, the two measures can be thought of alternative methods of measuring the rowwise distance between any two matrices of equal dimension.

Acknowledgements I should like to thank Dan MacKinlay for his mention of the K-L information loss measure, Arthur Ramer for his mention of the Hartley or U-uncertainty metric and his suggestions, and Vessela Daskalova for her mention of the "cityblock" metric. The efforts of the editors of this volume and anonymous referees were very constructive, and have greatly improved this chapter's presentation.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Chen, S.-H., Chang, C.-L., & Du, Y.-R. (2012). Agent-based economic models and econometrics. *The Knowledge Engineering Review*, *27*(2), 187–219.

Fagiolo, G., Moneta, A., & Windrum, P. (2007). A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. *Computational Economics*, *30*(3), 195–226.

Fagiolo, G., Guerini, M., Lamperti, F., Moneta, A., & Roventini, A. (2019). *Validation of agent-based models in economics and finance*. pp. 763–787.

Ferson, S., Oberkampf, W. L., & Ginzburg, L. (2008). Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering*, *197*, 2408–2430.

Gilbert, N., & Troitzsch, K. G. (2005). *Simulation for the social scientist* (2nd ed.). Open University Press.

Guerini, M., & Moneta, A. (2017). A method for agent-based models validation. *Journal of Economic Dynamics & Control*, *82*, 125–141.

Hartley, R. V. L. (1928). Transmission of information. *The Bell System Technical Journal*, *7*(3), 535–563.

Klir, G. J. (2006). *Uncertainty and information: Foundations of generalized information theory*. New York: Wiley.

Krause, E. F. (1986). *Taxicab geometry: An adventure in non-euclidean geometry*, New York: Dover. (First published by Addison-Wesley in 1975.)

Kullback, J. L., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Lamperti, F. (2018a). An information theoretic criterion for empirical validation of simulation models. *Econometrics and Statistics*, *5*, 83–106.

Lamperti, F. (2018b). Empirical validation of simulated models through the GSL-div: An illustrative application. *Journal of Economic Interaction and Coordination*, *13*, 143–171.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151.

Liu Y., Chen W., Arendt P., & Huang H.-Z. (2010). Towards a better understanding of model validation metrics. In *13th AIAA/ISSMO Multidisciplinary Analysis Optimization Conference, Multidisciplinary Analysis Optimization Conferences*.

Mankin, J. B., O'Neill, R. V., Shugart, H. H., & Rust, B. W. (1977). The importance of validation in ecosystem analysis. In G. S. Innis (Ed.), *New Directions in the Analysis of Ecological Systems, Part 1, Simulation Council Proceedings Series, Simulation Councils, La Jolla, California* (Vol. 5, pp. 63–71). Reprinted. In H. H. Shugart & R. V. O'Neill (Eds.), *Systems ecology* (pp. 309–317). Hutchinson and Ross, Stroudsburg, Pennsylvania: Dowden.

Marks, R. E. (1992). Breeding hybrid strategies: Optimal behaviour for oligopolists. *Journal of Evolutionary Economics*, *2*, 17–38.

Marks, R. E. (2007). Validating simulation models: A general framework and four applied examples. *Computational Economics*, *30*(3), 265–290. http://www.agsm.edu.au/bobm/papers/s1.pdf.

Marks, R. E. (2010). Comparing two sets of time-series: The state similarity measure. In V. A. Alexandria (Ed.), *2010 Joint Statistical Meetings Proceedings-Statistics: A Key to Innovation in a Data-centric World, Statistical Computing Section* (pp. 539–551). American Statistical Association.

Marks, R. E. (2013). Validation and model selection: Three similarity measures compared. *Complexity Economics*, *2*(1), 41–61. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.401.6982&rep=rep1&type=pdf.

Marks, R. E. (2016). Monte Carlo. In D. Teece, & M. Augier (Eds.), *The palgrave encyclopedia of strategic management*. London: Palgrave.

Marks, R. E., Midgley, D. F., & Cooper, L. G. (1995). Adaptive behavior in an oligopoly. In J. Biethahn, & V. Nissen (Eds.), *Evolutionary algorithms in management applications* (pp. 225–239). Berlin: Springer.

Midgley, D. F., Marks, R. E., & Cooper, L. G. (1997). Breeding competitive strategies. *Management Science*, *43*(3), 257–275.

Midgley, D. F., Marks, R. E., & Kunchamwar, D. (2007). The building and assurance of agent-based models: An example and challenge to the field. *Journal of Business Research*, *60*(8), 884–893. (Special Issue: Complexities in Markets).

Oberkampf, W. L., & Roy, C. J. (2010). Chapter 12: Model accuracy assessment. *Verification and validation in scientific computing* (pp. 469–554). Cambridge: Cambridge University Press.

Ramer, A. (1989). Conditional possibility measures. *International Journal of Cybernetics and Systems*, *20*, 233–247. Reprinted in D. Dubois, H. Prade, & R. R. Yager, (Eds.). (1993). *Readings in fuzzy sets for intelligent systems* (pp. 233–240). San Mateo, California: Morgan Kaufmann Publishers.

Rényi, A. (1970). *Probability theory*. Amsterdam: North-Holland (Chapter 9, Introduction to information theory, pp. 540–616).

Roy, C. J., & Oberkampf, W. L. (2011). A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, *200*, 2131–2144.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Chapter 14
Analysing Output from Stochastic Computer Simulations: An Overview

Christine S. M. Currie

Abstract Proper statistical analysis of the output of a stochastic simulation model helps to ensure against drawing conclusions based on random fluctuations. In this chapter, we detail methods for analysing the output of stochastic simulation models. We consider terminating and non-terminating simulations and demonstrate how to set initial conditions for the former, and how to determine the length of the warm-up period in the latter using Welch's method and MSER-5. The chapter also describes methods for choosing the number of replications needed to draw statistically valid conclusions and how to compare between different options. We introduce some basic statistical concepts at the start of the chapter to help with understanding what follows and use two examples throughout the chapter to demonstrate the methods.

Keywords Simulation · Output analysis

14.1 Introduction

During the validation of a computer simulation, most often, the output of the simulation program is compared to data (see in particular Chap. 15 by Murray-Smith in this volume). This assumes that the output is well understood, which can be achieved using output analysis. Output analysis can be seen as a poor relation to the model-building process—it is less exciting and perhaps more mundane—but it can make or break a simulation project. If little consideration is paid to how the model outputs are interpreted, decisions can be made based on results that are not statistically significant and may simply be due to random fluctuations. Alternatively, much time can be wasted running many replications of long simulations unnecessarily if the statistics show that significant results can be obtained much earlier or more efficiently with a well-thought-out experimental structure.

C. S. M. Currie (✉)
Mathematical Sciences, University of Southampton, Southampton, UK
e-mail: christine.currie@soton.ac.uk

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_14

We consider here the analysis of computer simulation models with stochastic outputs. Such models use *pseudorandom numbers* to mimic the randomness inherent in real systems, where pseudo random numbers are generated by a deterministic computer algorithm to mimic true random numbers. Examples of the stochastic simulation model methodologies include discrete event simulation and agent-based modelling, which have a wide range of applications from modelling of processes such as hospitals, factories, call centres to modelling of individual behaviour in crowd situations, transmission of diseases, customer behaviour in markets, etc. Our focus is on how to obtain reliable results that take full account of the inherent uncertainty in the system. The output of non-stochastic or deterministic simulations, by contrast, is much more straightforward to analyse and is thus not considered in this chapter.

We use two examples to demonstrate the techniques covered in this chapter: Example 1 is a simple queueing system such as that observed in a bank or post office, with one server and unlimited waiting space; and Example 2 is a more complex individual-based model of tuberculosis (TB) and HIV, previously published in (Mellor et al. 2011). Both of these models have been programmed using discrete event simulation such that the system only changes state when an event occurs, e.g. an arrival or a departure from the queueing model. This means that the simulation can jump from one event to the next rather than using a fixed time step. Nonetheless, the methods we describe here will also work with fixed time step models.

The content of this chapter is fairly statistical and while we cover some preliminaries in the following section, readers may wish to revise some of their basic knowledge of random variables and statistics before moving on. There is an extensive literature in output analysis for simulation models that this work draws on and readers who wish to know more about the subject will find useful articles in the online archive of the proceedings of the Winter Simulation Conference (http://www.wintersim.org). There are also a number of excellent books that cover output analysis, which is referred to during the chapter.

We begin in the following section by defining some of the terminologies that we will use in the remainder of the chapter as well as going through some of the background knowledge that the reader will need. In the main part of the chapter, we focus separately on terminating and non-terminating simulations. These are defined below and have different challenges for output analysis. We then go on to discuss how to estimate the number of replications that are needed in Sect. 14.5 before describing methods for effective comparisons of different options in Sect. 14.6. Finally, we provide a discussion with details of where to find out more.

14.2 Preliminaries

14.2.1 Definitions

For the purposes of what follows, we split stochastic simulation models into two broad categories: **terminating** simulations and **non-terminating** simulations. Terminating simulations run until a particular event or set of events occur. This might be a set time or alternatively could be a certain type of random event such as the first time something happens. Of particular interest in the output analysis of terminating simulations is the setting of initial conditions. In contrast, non-terminating simulations do not have a well-defined endpoint and tend to be used to model *steady-state* behaviour, also described as stationary or equilibrium behaviour. A simulation model is said to be in a steady state if the initial conditions have no influence on the behaviour of the system, i.e. once the initial transient behaviour has settled down.

It is important to note that in the steady-state, output is not constant and does still vary, but it will vary according to a fixed, steady-state distribution. For these simulations, output analysis is needed to determine the length of the warm-up period. We consider terminating simulations in Sect. 14.3 and non-terminating simulations in Sect. 14.4.

We here define some of the common simulation terms that we will use in the rest of the chapter.

- **Initial conditions**: the initial conditions of a simulation model define the state that the system is in at the start of the first time step.
- **Warm Up**: the number of time steps for which the simulation is run before it is assumed to have reached its desired state. For terminating simulations, this could be the number of time steps until the model reaches its desired initial state. For non-terminating simulations, the warm-up is the number of time steps until the model can be viewed as being in the steady-state.
- **Replication**: a run of the simulation model.
- **Pseudorandom Numbers**: these are generated within the simulation model and should exhibit statistical randomness. Each replication is assumed to use a different set of pseudorandom numbers.

14.2.2 Background Statistical Knowledge

The **mean** of a random variable Y can be estimated as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{14.1}$$

where y_i is the ith observation out of a total of n observations.

The **variance** of a random variable Y can be estimated by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \tag{14.2}$$

where S is the standard deviation.

The **confidence interval** describes a range of likely values for a calculated statistic such as the mean of a random variable. Typically, 95% or 90% confidence intervals are calculated, meaning that if the same population is sampled several times and interval estimates are made on each occasion, the resulting intervals would include the true population mean in approximately 95% (or 90%) of the cases. Setting a confidence interval of 95% implies a **significance level** α of 5%.

Standard methods for calculating confidence intervals assume that the y_i follow a normal distribution. This is more likely to be true if the y_i are themselves sample means, which have been calculated by averaging over a large number of observations (e.g. if we use a long simulation run) or if n is large. If the data are not normal, the following calculation of the confidence interval is invalid and a technique such as bootstrapping (Efron and Tibshirani 1998) might be a suitable alternative.

There are a number of formal normality tests available that can be used to determine whether the y_i follow a normal distribution, with the EDF tests (Anderson–Darling, Kolmogorov–Smirnoff, Cramér-von-Mises) probably considered to be the most powerful. (See, for example, (D'Agostino and Stephens 1986) for how these can be implemented, although most statistics packages will have them built in). The above tests work best with a moderate to high number of data points. Shapiro–Wilk is another useful test, and tends to be more appropriate for small numbers of data points. Alternatively, a probability or q–q plot provides a visual check of normality. (A q–q plot or quantile-quantile plot to give it its full name plots the data on the x-axis against the statistical model—in this case, the normal distribution—on the y-axis. If the data follow the statistical model, the data points should be evenly distributed about the 45°, $y = x$ line).

Assuming the output is normal, the $100(1-\alpha)\%$ confidence interval around the estimate of the performance measure, $\bar{y}$, is given by

$$\left[\bar{y} - t_{1-\alpha/2;n-1}\frac{S}{\sqrt{n}},\ \bar{y} + t_{1-\alpha/2;n-1}\frac{S}{\sqrt{n}}\right], \tag{14.3}$$

where $t_{1-\alpha/2;n-1}$ describes the point on the t-distribution with $n-1$ degrees of freedom where the cumulative distribution function is equal to $1-\alpha/2$.

14.2.3 Setting Up the Problem

A simulation study will have a set of performance measures that are being calculated via simulation model outputs. For simplicity, we consider univariate outputs here, e.g. the number of people waiting in a queue or the number of new infections of TB disease in the population, and use y to describe the output that we are interested in. We run n replications of the simulation model and in each replication, we record m observations of y. Here, m can take on any value but will typically be either equal to 1 or T, the number of timesteps in each replication. Note however that in example 1, our queueing model, we have C observations, where C is the number of customers that enter the system.

The values

$$\mathbf{y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{pmatrix},$$

constitute the simulation output data, where y_{ij} is the jth observation in the ith replication of the simulation model. Simulation output data are random, as we are considering stochastic simulation models. In this sense, running the model is just like observing reality in that each replication of the model will generate a new set of results, assuming that we use a different set of pseudorandom numbers in each replication.

14.3 Working with Terminating Simulations

Working with terminating simulations, it is important that the simulation model is in the right state when we start collecting results. This means that when the data collection starts, it should obey the initial conditions. For some systems, the initial conditions are obvious from the setting, e.g. when simulating a shop from its opening time, the initial conditions would be that the shop starts empty. In other situations, starting the system empty is no longer valid. For example, if we are instead interested in how the shop behaves over the lunchtime period, it would be more realistic to have some customers in the system at the start of the results collection period. Before we consider how we can achieve this, let us think about how we can choose the correct initial conditions for a system.

For real systems, it may be feasible to observe the state of the system at a given time. For the shop example, this would involve observing the number of customers in the queue and being served at the start of the lunchtime period over a number of days. In more complex systems, there may be data available that suggest either the values of some of the states or alternatively values for the system outputs at the desired start point.

For complex models, where it is hard to determine each of the individual state values, the best way of obtaining the correct conditions is to run the model prior to the start of the results collection period. We demonstrate this via an example.

Example: Modelling TB and HIV We use an infectious disease model of tuberculosis (TB) and HIV to illustrate this point. The model is a stochastic, individual-based simulation model that was built to describe the impact of the HIV epidemic on the incidence rate of new cases of TB disease in Zimbabwe. To help with modelling TB transmission, the population were divided into households. Details of the model can be found in (Mellor et al. 2011) and we consider some experimentation with the model below.

When setting initial conditions, data were available on the incidence of TB disease (i.e. the number of reported cases of TB per year per 100,000 population) from 1980, shortly before the start of the HIV epidemic, until 2002 and the distribution of the population within households, based on a survey. The epidemic was not expected to reach a steady-state and so a terminating simulation model was built to describe the transmission of disease in the population and the progression of the disease within individuals. Our aim when setting up the model was to ensure that the simulation model reaches a state in 1980 that allows the impact of HIV on the population of Zimbabwe to be reproduced as accurately as possible. While we know the HIV prevalence and TB incidence in 1980 and beyond, we are unable to know the detailed information of each state, e.g. the number of people in each stage of TB disease, etc. As a result, we need to run the model from an earlier date so that it is in the correct state when results collection begins.

In this case, a long run length was needed due to the initial oscillation in the output, as shown in Fig. 14.1. To save computational time in running the scenario analysis later in the project, the state of the simulation model at the end of the warm-up was saved for a large number of warm-up runs, and the first step of each replication of the simulation model was to sample from this set of initial states, to choose the initial conditions for that particular replication.

14.4 Working with Non-terminating Simulations

When calculating statistics for the output of non-terminating simulations we concentrate only on the steady-state output, ignoring the initial transient. It is important to also take account of the variability in the output. There are two main approaches of doing this. In the first, we run the simulation for a number of replications and combine results from the replication, e.g. by taking averages. An alternative is to observe only one, long simulation run and split this up into smaller sections or batches. This is the technique of **batch means**. Some care needs to be taken in applying batch means to allow for **autocorrelation** in the output. If data are autocorrelated, there are a dependence between data points in the series of output data. An example of autocorrelated data might be daily ice cream sales over a year in the UK,

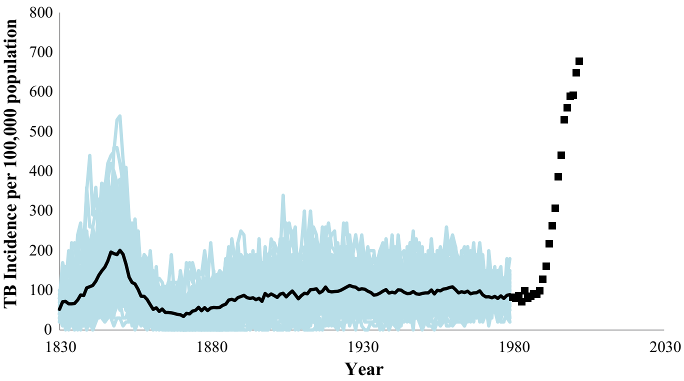


Fig. 14.1 Time series showing the simulation output of TB incidence and setting up the initial conditions. The TB data for Zimbabwe is shown as black squares; results for 100 runs are shown in pale blue; the average model output is shown as a black solid line. (© IEEE. Reprinted, with permission, from (Currie and Cheng 2016), Fig. 5)

where there is likely to be a connection between values on adjacent days due to the weather dependence of ice cream sales. We do not consider batch means any further here and instead refer the interested reader to (Robinson 2014) for a good review of batch means methods.

With non-terminating simulations, the first task of the output analysis concentrates on ensuring that the initial transient behaviour is not allowed to affect the results of the simulation study. Estimating the initial transient or warm-up duration can be done in a number of ways and we consider two options below. Having estimated the warm-up duration, any output data generated during this time is removed from the analysis.

Figure 14.2 shows the output from Example 1, what is known as a $M/G/1$ queue, in which customers arrive into the system following a Poisson distribution (M) and are then served with the service time following a general distribution (G). The one refers to the fact that there is just one server. Here, the output is the waiting time of each successive customer. Service is assumed to follow a truncated normal distribution (truncated to prevent negative service times). As discussed above, this could be used to describe the arrival and service process in a small shop or post office.

Although the output is random, it is possible to observe an initial period where there is an upwards trend, before the data series settles down to what can be regarded as its steady-state behaviour. The period of fluctuations at the start of the simulation run is often termed the **initial transient** and if output from this period is included in the calculation of statistics of interest (e.g. including the initial small values for the waiting time in this example), they are likely to bias the results such that we do not get a good approximation to the steady-state mean.

The most common way of dealing with the initial transient is to delete the output from this period, which we define to be the **warm-up period**. For example, if we

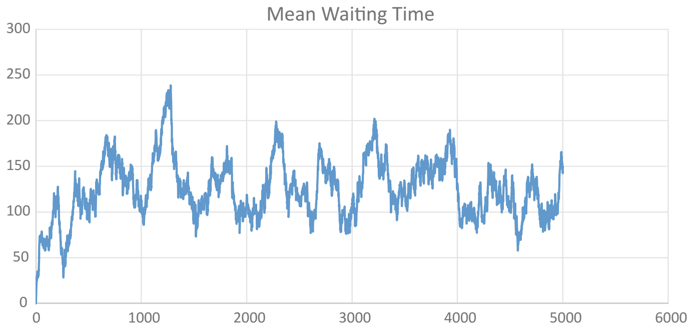


Fig. 14.2 Time series showing the mean waiting time of customers in the queue, averaged over 10 iterations

have a series of simulation output data $y_{i1}, y_{i2}, \ldots, y_{iT}$ coming from T observations of a particular model output y during the ith replication, we would calculate the mean for replication i, $\bar{y}_i$, as being equal to

$$\bar{y}_i = \frac{\sum_{t=t_0+1}^{T} y_{it}}{T - t_0},$$

where the warm-up period is set to be equal to t_0. Other statistics, such as the variance or range of variables would be calculated in a similar way, i.e. by ignoring the first t_0 observations.

Identification of the warm-up period has been a topic of research for many years and we recommend (Hoad et al. 2010) as a good review and test of different methods. In this chapter we consider two alternatives. The first, Welch's Method (introduced in (Welch 1981) and described clearly in (Law 2014)) is a graphical method, where the modeller decides on the warm-up duration based on observations of the time series. This is simple and easy to implement but subjective and hard to automate. The second method we introduce is MSER-5, which is a heuristic method introduced by (White 1997) and tested in (Hoad et al. 2010). The acronym stands for Marginal Standard Error Rule and the method makes a trade-off between precision and bias, as we discuss further below.

14.4.1 Welch's Method

The principle behind Welch's method is to set the warm-up period t_0 such that after t_0, a moving average of the output settles down to a relatively constant value. It is not possible to determine the warm-up period from just one run of the simulation model and the standard practice is to carry out at least five replications and find the average output. It is also important to run the simulation model for long enough to

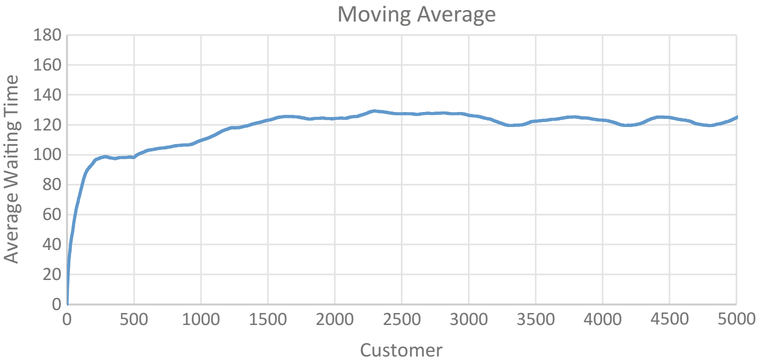


Fig. 14.3 Moving average of the average waiting time for the M/G/1 queue using 50 replications and a moving average window of 500

be certain that the system has reached a steady-state by the end of the run (see e.g. Chapter 9 of (Robinson 2014)). The procedure is as follows.

1. Run the simulation model n times, where $n \geq 5$ to obtain a time series of the output of interest, y.
2. Calculate the mean of the output over the n replications for each observation $t = 1, 2, ..., T$, writing these as $\bar{y}_t = \sum_{i=1}^{n} y_{it}/n$.
3. Calculate a moving average for the $\bar{y}_t$ using a window size of w, where $w \leq \lfloor T/4 \rfloor$ and T is the number of observations in each replication. The function $\lfloor T/4 \rfloor$ is known as the *floor function* and denotes the integer part of $T/4$.
4. Plot the moving average $\bar{y}(w)$.
5. If the data do not look smooth, increase the window size w and repeat steps 3 and 4.
6. The warm-up period t_0 is the point at which the moving average flattens out.

For reference, we include the equations for calculating the moving average below.

$$\bar{y}_t(w) = \begin{cases} \left(\sum_{\tau=1}^{2t-1} \bar{y}_\tau\right)/(2t-1) & t = 1, \ldots, w \\ \left(\sum_{\tau=0}^{2w} \bar{y}_{t-w+\tau}\right)/(2w+1) & t = w+1, \ldots, T-w \end{cases}$$

Considering the first example of the M/G/1 queue, we plot the moving average in Fig. 14.3. The output from this queue is highly variable and we find that we need 50 replications and a window length of 500 to obtain a relatively smooth plot. This suggests a warm-up duration of around 1500, i.e. we remove the waiting time of the first 1500 customers when calculating the mean and variance of the waiting times.

14.4.2 MSER-5

The MSER algorithm is based on the fact that, as the duration of the warm-up t_0 increases, the bias associated with the result will decrease, improving the estimate. However, a higher t_0 means that a smaller number of observations is used to calculate the output statistic, and the precision of the statistic will decrease. MSER-5 aims to balance the bias reduction with the increase in variance.

We define t_0^* to be the optimal warm-up duration. As with Welch's method, the simulation is run for n replications where $n \geq 5$. For the M/G/1 example, we consider below, we set n to be much higher than this (50) as the output is highly variable, and the computational time for each simulation run is very short, but this will vary from model to model.

The MSER Statistic gives a measure of the statistical error in the output. For a small t_0, the error is mainly due to bias and for a large t_0, it is mainly a result of increased variability. In order to calculate the MSER statistic, we first define the b batch averages

$$z_k = \frac{1}{5n} \sum_{i=1}^{n} \sum_{j=1}^{5} y_{i,5(k-1)+j}, k = 1, \ldots, b = \lfloor T/5 \rfloor$$

and then the estimated mean, ignoring the warm-up period

$$\bar{z}_{b,t_0} = \frac{1}{n(T - t_0)} \sum_{i=1}^{n} \sum_{t=t_0+1}^{T} y_{it}.$$

If we define $b_0 = t_0/5$, the number of batches in the warm-up period, then the optimal value for b_0 is given by the minimum of the MSER statistic

$$MSER = \frac{1}{(b - b_0)^2} \sum_{k=b_0+1}^{b} (z_k - \bar{z}_{b,t_0})^2.$$

As with Welch's method, we use a long run length when determining the warm-up duration such that $b >> b_0$.

If we apply MSER-5 to the M/G/1 queue example, the minimum occurs at $b = 229$, corresponding to $t = 1145$, with the plot shown in Fig. 14.4. This matches reasonably well with the estimate obtained using Welch's method but has the advantage that the methodology can be automated.

14.5 How Many Replications?

The number of replications should be decided based on the level of precision required for the results and the variability of the simulation output. In what follows, we show how to set the number of replications n to obtain the desired level of accuracy in the mean of our output of interest $\bar{y}$. The precision/accuracy is indicated by the sample variance S^2 and the confidence interval, which we wrote expressions for in Sect. 14.2.2.

Equation 14.3 shows that as the number of replications n increases, the width of the confidence interval decreases at a rate of $1/\sqrt{n}$. There is also a smaller effect coming from the t-distribution which has thinner tails as n increases. In order to determine the appropriate value of n to achieve a desired precision, we must first decide how we wish to measure it: *absolute error* or *relative error*. Absolute error is a measure of the difference between the estimated mean coming from the simulation model $\bar{y}$ and the unknown true mean of the output, μ. In contrast, relative error is the absolute error divided by the unknown true mean. In effect, the relative error allows for the scale of the observations, enabling two error rates to be compared. For example, an absolute error of 10 is large if the mean value is 20 but is unlikely to be important if the mean value is 2000. These two situations would have relative error rates of 0.5 and 0.005 respectively. The choice of which error rate to use is likely to be dictated by the nature of the results being output.

Considering absolute error initially, we define our precision as $|\bar{y} - \mu|$ and assume that we are setting n such that

$$|\bar{y} - \mu| < \epsilon \text{ is true with probability } 1 - \alpha.$$

In order to determine n, we must make a small number of initial replications, m, to give us an estimate of the underlying variance based on the first m observations, $\tilde{S}^2(m)$. Rearranging Eq. 14.3, we can then say that the number of replications required

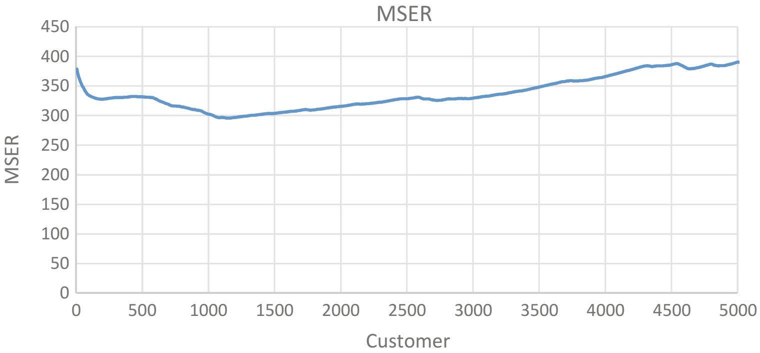


Fig. 14.4 Plot showing the variation in the MSER Statistic with the customer number

to obtain a precision of ϵ, n_ϵ^{abs} is the minimum value of n such that $n \geq m$ and

$$t_{1-\alpha/2;n-1}\frac{\tilde{S}(m)}{\sqrt{n}} \leq \epsilon.$$

When using the relative error, we set n such that

$$|(\bar{y}-\mu)/\mu| < \epsilon \text{ is true with probability } 1-\alpha,$$

i.e. the the estimator should be within $100\epsilon\%$ of the true value with probability $1-\alpha$. A similar principle is used to that for the absolute error, but we now need to take into account the mean of the relevant performance measure, $\bar{y}(m)$ calculated after the first m replications. In this case, the number of replications, n_ϵ^{rel} that are needed for the estimator to be within $100\epsilon\%$ of the true value with probability $1-\alpha$ is the minimum value of n such that $n \geq m$ and

$$t_{1-\alpha/2;n-1}\frac{\tilde{S}(m)}{|\bar{y}(m)|\sqrt{n}} \leq \frac{\epsilon}{1+\epsilon}.$$

For more details of the derivation of these formulae, see (Law 2014). These estimates are unlikely to be perfect as they rely on the estimates of $\bar{y}(m)$ and $\tilde{S}^2(m)$ calculated after only a small number of replications. Nonetheless, they can provide a good ballpark figure for the number of replications that should be made.

14.6 Making Comparisons

Many simulation projects set out to compare different system configurations or to find the best value for a given parameter. For a more in-depth study, the simulation optimisation methods described in (Fu 2015) provide an excellent way of optimising a system via simulation, taking account of the stochasticity of the output. There is insufficient space available to cover these methods here and we instead describe how confidence intervals can be calculated for a comparison between two systems. In what follows, we consider an example used previously in (Currie and Cheng 2016), which compares the impact of different interventions on the incidence of TB disease.

14.6.1 Comparing Two Systems

When comparing two systems, we wish to run sufficient replications to be confident that the system we observe to be better is the right one, usually up to a 95% or 90% confidence level. In what follows we assume that we record the mean value of our output in each replication, writing $\bar{y}_i(1)$, $\bar{y}_i(2)$, $i = 1, \ldots, n$ to denote the mean observed in replication i for systems 1 and 2, respectively.

The quantity of interest here is the difference between system 1 and system 2,

$$\Delta_i = \bar{y}_i(1) - \bar{y}_i(2).$$

Taking the mean over the n replications,

$$\bar{\Delta}(n) = \sum_{i=1}^{n} \Delta_i / n$$

and the variance of the Δ_i is given by

$$Var[\bar{\Delta}(n)] = \frac{\sum_{i=1}^{n}(\Delta_i - \bar{\Delta}(n))^2}{n(n-1)}.$$

If there is a significant difference between the two systems, the mean of the differences $\bar{\Delta}(n)$ will have a confidence interval that does not include zero. Assuming a normal distribution, the $(100 - \alpha)\%$ confidence interval is given by

$$\bar{\Delta}(n) \pm t_{n-1,1-\alpha/2}\sqrt{Var[\bar{\Delta}(n)]},$$

where $t_{n-1,1-\alpha/2}$ is the $1 - \alpha/2$ point of the t-distribution with $n - 1$ degrees of freedom. This can be found using a statistical package or even Excel.

Returning to the model of TB and HIV considered earlier, we demonstrate how to compare two interventions designed to counteract the impact of the two diseases. Intervention 1 mimics the impact of visiting households in which someone has previously been diagnosed with TB disease and intervention 2 visiting households in which one or more of the members is in the later stages of HIV. The output variable of interest here is the number of cases of TB disease found using the two interventions.

We run 50 replications with each intervention, keeping all other model parameters the same, and so $n = 50$, and observe a mean difference of 43 cases, with a variance of 170. Before calculating the confidence interval, we run the Anderson–Darling test to check for normality. The null hypothesis ($H0$) of the Anderson–Darling test is that the data are a random sample from the normal distribution. We obtain a p-value of 0.090. Therefore, using a significance level of 0.05, we cannot reject the null hypothesis that the data is normal, and we are safe to continue with our calculation of confidence intervals. The t-distribution has $n - 1 = 49°$ of freedom, resulting in a 95% confidence interval of $[-21, -65]$. As the interval does not straddle zero, we are safe to conclude that we can see a significant difference between the two interventions, with the second intervention finding more TB cases.

If the Anderson–Darling test suggests that the differences do not follow a normal distribution there are other ways of calculating a confidence interval, with the easiest method to follow being bootstrap resampling. There is insufficient space to discuss this here and we recommend (Efron and Tibshirani 1998) as a good introduction to this technique for the interested reader. When the variance is high, it may be necessary to run many replications to find a significant difference between the two systems. In this case, it can be helpful to use the technique of *common random numbers* as a way of reducing the variability (e.g. see (Law 2014)).

14.6.2 Comparing Many Systems

The statistical analysis when comparing many systems is more complicated than when comparing just two as it is necessary to take account of the fact that several hypothesis tests are being carried out simultaneously. For example, consider calculating 95% confidence intervals for 10 comparisons. For each of these comparisons, there is a 5% or 1 in 20 chance that the true difference lies outside the confidence interval, a so-called type 1 error. Therefore, the overall confidence interval for those ten comparisons, the chance that one of the true differences lie outside its confidence interval, is in fact as low as 60%.

There are methods available that can account for this, the best known being the Bonferroni correction. The Bonferroni correction is discussed at length in (Law 2014) and we refer the interested reader to this book for more details. In essence, if the required confidence level is $(1 - \alpha)\%$, and p comparisons are being made, then each of the individual comparison confidence levels need to be $(1 - \alpha/p)\%$. This increases the number of replications that need to be made in order to meet the required significance.

14.7 Conclusion

This chapter demonstrates some useful techniques in output analysis, which allows for the stochastic output of a simulation model. The key messages should be to think carefully about the meaning of the output, to plot output data where possible to gain a better understanding of how it is distributed, and to always run more than one replication of a stochastic simulation model.

Model verification and validation techniques determine whether a model is correct. Output analysis sits within that by ensuring that the results reported for stochastic simulation models are an accurate representation of the model output. The techniques described in this chapter will also help with providing an accurate description of the uncertainty in the simulation results, which is vital when using them to make decisions.

References

Currie, C., & Cheng, R. (2016). A practical introduction to analysis of simulation output data. In T. Roeder, P. Frazier, R. Szechtman, & Zhou, E. (Eds.), *Proceedings of the winter simulation conference* (pp 118–132). Winter Simulation Conference.

D'Agostino, R., & Stephens, M. (1986). *Goodness-of-Fit Techniques. Statistics: textbooks and monographs 68*. Marcel Dekker.

Efron, B., & Tibshirani, R. (1998). *An introduction to the bootstrap*. Chapman and Hall.

Fu, M. (2015). *Handbook of simulation optimization*. Springer.

Hoad, K., Robinson, S., & Davies, R. (2010). Automating warm-up length estimation. *Journal of the Operational Research Society*, *61*, 1389–1403.

Law, A. M. (2014). *Simulation modeling and analysis* (5th ed.) McGraw-Hill.

Mellor, G., Currie, C., & Corbett, E. (2011). Incorporating household structure into a discrete-event simulation model of tuberculosis and hiv. *ACM Transactions on Modeling and Computer Simulation*, *21*, 26.

Robinson, S. (2014). *Simulation: The practice of model development and use* (2th ed.) Palgrave.

Welch, P. D. (1981). *On the problem of the initial transient in steady-state simulation*. IBM Watson Research Center.

White, J. K. P. (1997). An effective truncation heuristic for bias reduction in simulation output. *Simulation*, *69*, 323–334.

Part IV
Methodology—Points of Reference and Related Techniques

Chapter 15
The Use of Experimental Data in Simulation Model Validation

David J. Murray-Smith

Abstract The use of experimental data for the validation of deterministic dynamic simulation models based on sets of ordinary differential equations and algebraic equations is discussed. Comparisons of model and target system data are considered using graphical methods and quantitative measures in the time and frequency domains. System identification and parameter estimation methods are emphasized, especially in terms of identifiability analysis which can provide valuable information for experiment design. In general, experiments that are suitable for system identification are also appropriate for model validation. However, there is a dilemma since models are needed for this design process. The experiment design, data collection and analysis of model validation results is, inevitably, an iterative process, and experiments designed for model validation can never be truly optimal. A model of the pulmonary gas exchange processes in humans is used to illustrate some issues of identifiability, experiment design and test input selection for model validation.

Keywords Graphical comparisons · Quantitative measures · Identifiability · Experiment design · Test inputs

15.1 Introduction

This chapter discusses issues associated with the use of experimental data for simulation model validation. This is central to many of the ideas of validation, as introduced in Chap. 4 by this author in this volume. As explained in that chapter, validation is concerned with the processes involved in establishing the extent to which a given model is consistent with the target system using information which is external to the model, as compared with the more internal processes of verification which are used to establish the correctness, or otherwise, of the code and algorithms used in the implementation of the simulation model.

D. J. Murray-Smith (✉)
School of Engineering, University of Glasgow, Rankine Building, Glasgow G12 8QQ, UK
e-mail: David.Murray-Smith@glasgow.ac.uk

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_15

The words ‘experimental modelling’ are used in this chapter to distinguish the development of a theoretical model using knowledge and understanding of the target system from methods of model development that also make use of data from measured or observed responses from that system. An appropriate model structure may often be established on a theoretical basis, but some details, such as precise values of specific constants (parameters) within the equations, may not be immediately available. For example, in deterministic dynamic models used in physics, engineering, physiology and many other fields, the structure of the model (usually nonlinear in form) may often be found using scientific laws and principles. However, experimental data from the target system may often be needed to provide estimates of parameters within that chosen model structure if they cannot be determined from a priori information. This type of approach to the development of simulation models has interesting and important implications in terms of simulation model validation methods.

Although we are concerned here with models that may be developed through a combination of theoretical modelling and experimental modelling methods, it is important to note that it is also possible to develop simulation models entirely from measured responses from the target system. This ‘system identification’ process involves determining both the model structure and the parameter values from experimental data and this may be thought of as the most extreme form of experimental modelling. System identification is closely related to methods of time series analysis (see, for example, Chatfield 1996) and is much used in the analysis and design of feedback control systems. Many system identification techniques lead to linear models and they are applied most often in situations where little prior information is available about the structure of the system under investigation.

In many cases, the target systems described by models derived using a combination of theoretical and experimental modelling methods involve one or more variables that can be regarded as ‘inputs’ in the sense that they can be changed in an independent fashion, while other variables of the system respond in some way to those applied changes. For example, the level of water in a bathtub depends on the flow rates at the hot and cold taps and these could be regarded as inputs for a model of the system, while the water level could be regarded as an ‘output’ variable. Models of practical systems may involve many variables, some of which may be regarded as inputs while others may be output variables (usually those that are important for the intended application of the model and for which measurements are often available). Model variables that are not described as ‘inputs’ or ‘outputs’ are known as intermediate variables, and measurements are often not available for these.

It must be noted that experimental modelling approaches for model development and validation are only appropriate in situations where the system being modelled is available for testing and this is clearly impossible in some cases, such as in the initial stages of engineering design. However, even when no target system exists, experimental data from other systems that resemble the target system may provide insight. This is especially true if the new project involves subsystems that are identical or very similar to some existing subsystems and, thus, to sub-models developed previously.

In situations where a model is derived from theory, test data from a target system may be essential for model validation in the context of the planned model application. However, experimental data sets are likely to be of limited value for validation of a dynamic system simulation model unless the tests carried out excite the system in such a way that the whole of the relevant range of operation is covered. In this respect, the requirements for experimental design for validation are essentially the same as those for system identification. Experience gained in system identification and parameter estimation is therefore directly relevant to the success of experimental modelling techniques in model validation. The most important point in both situations is that key variables of the target system must be perturbed during validation tests, usually through the application of externally applied test inputs and the choice of initial conditions. The measured time histories of those key variables from the target system must cover the complete range of importance. Both the amplitudes and rates of change of variables must be considered, and this has an influence both on the magnitude and the frequency content of the test inputs applied. Careful experiment design is therefore very important and best practice requires that experiments should be tailored precisely to the application, and it is always important that they are fully documented so that the resulting data sets can be used with confidence in the future.

Experimental data may also allow some aspects of a system to be eliminated from a theoretically based simulation model. For example, part of a complex theoretical model could be replaced by measured response data obtained from tests on the target system. This may simplify both the development of the model and the validation process. An example of this is given in a later section of this chapter within a case study involving a model of pulmonary gas exchange processes in human subjects where one key variable of the model (the gas flow into and out of the lungs) is based on measured data from the target system. The resulting model is, of course, specific to the individual human subject being considered.

The aim in validation of a model is usually to establish the range of conditions over which model predictions agree with the behaviour of the target system to some specified level of accuracy, but experimental data can also provide insight into possible sources of model deficiencies in the context of the intended application. Experimental conditions should always be chosen to maximize information about the target system, with special emphasis being given to measured quantities that relate directly to aspects of the model where most uncertainties exist.

Ideally, we would like to have confidence intervals for model predictions, but these are not easily found. Usually, the best we can do is to assemble all the available quantitative information about the model and the target system, together with results from face validation involving people with expert practical knowledge. Thus, an aircraft model for a ground-based piloted simulation facility might be tested initially through quantitative comparisons of simulation model histories and corresponding flight test data, but could also be assessed by experienced test pilots who would make suggestions for model improvement using a more subjective approach involving handling qualities and manoeuverability criteria applied to the simulator and to the real aircraft during flight tests.

This chapter starts with a section discussing general issues concerning data used during the development and testing of deterministic models. An important distinction is made between applications in which data are based only on observations of a target system and applications in which it is possible to carry out experiments involving imposed changes of operating conditions. This leads on to a section in which an introduction to experimental modelling methods is presented and, in the two sections that follow, some graphical and quantitative ways comparing data sets from the target system and the simulation model are discussed. A brief outline is then given of some system identification and parameter estimation approaches, which are central to the experimental methods in model validation and the concept of 'identifiability' is emphasized. This can provide insight about the structure and choice of parameters in theoretically based models as well as being important for the design of experiments that are efficient in terms of extracting information from the target system. Some methods that have been found to be particularly effective for experiment design in the context of model validation are then presented, together with discussion of model complexity issues and the problems of 'over-fitting' and 'under-fitting' of parameters. This leads on to a section in which experimental modelling methods are applied to a theoretically based simulation model of human pulmonary gas exchange processes where identifiability and experiment design methods in simulation model validation are stressed. The final two sections provide some general discussion and conclusions relating to issues concerning the use of experimental data in the validation of nonlinear simulation models that are dynamic and essentially deterministic in form.

15.2 Data Sets for Model Development and Testing

Data from the target system, whether from historical sources or from specially designed experiments, may provide important information for use in the development of a simulation model. The experimental data form the reference against which competing models are compared during the validation process. It is, therefore, essential to know the limits of accuracy of the data since, without that information, model predictions cannot be properly assessed.

Errors in data from the target system may be random or systematic in form. Random errors are often associated with measurement noise. The scatter of results obtained when experiments are repeated can give a rough indication of the significance of such errors. Systematic errors, on the other hand, produce bias in results and this cannot be reduced through additional testing. Such errors arise from sensor offsets, sensor calibration errors and problems within the hardware and software used for data collection, as discussed by many authors (for example, Hughes and Hase 2010). Although random errors are always present in measurements, much can be done to eliminate or reduce systematic errors through, for example, introducing redundancy within the hardware used and employing sensor fusion or state estimation methods such as Kalman filtering (see, for example, Tischler and Remple 2006).

Some modelling applications, such as arise in astrophysics, climatology or economics, do not offer possibilities for experimentation and data for validation must come from observations. Sometimes, only historical data sets are available. In such cases, data must be split into subsets for model development (e.g. using system identification and parameter estimation methods) and subsets for subsequent model testing. A data set should never be used in the estimation of unknown model parameters and, again, at the validation stage.

Although experimental modelling techniques, such as system identification and parameter estimation, are usually seen as being important for the development of a model, these tools can also be very valuable for validation. For example, trends in estimated parameter values over a range of operating conditions may be compared with corresponding parameter values found from the theoretical model.

15.3 Comparison Methods for Model and Target System Data Sets

In practical applications, simulation models often involve many different output variables and can generate large quantities of data. In terms of measures of model quality, outputs must be chosen which are meaningful for the planned application.

Analysis of model quality may be based on many different model and system comparison measures, such as particular features of time histories for specific variables, including steady-state values, or frequencies of oscillation, or rates of decay of output variables during a transient. It could also involve comparisons of complete time histories for the target system and the model, recorded over a specified period for several variables. A combination of quantitative measures and graphical representations is often emphasized in discussions of model validation, since graphical representations can often provide insight not available from quantitative measures on their own.

15.3.1 Graphical Methods for System and Model Data Comparisons

Simple graphical methods allow comparisons of experimental data and the corresponding model-generated outputs and any major differences between predicted output variables from the simulation model and the corresponding time histories of measured variables of the target system may suggest that the model structure is wrong, especially if records display distinctive features that are similar across data sets for different variables.

Graphical presentations are usually based on plots of simulated variables and the corresponding observed or measured values against an independent variable which, for dynamic models, is usually time. Additional plots showing differences between the model and measured variables versus time may emphasize mismatches and autocorrelation functions of these residuals can often highlight modelling errors (see, e.g. Gustavsson 1972). Ideally, the residuals should resemble white noise and the autocorrelation function should show a distinct peak around the time origin with other values close to zero. Tools such as parameter sensitivity analysis can also be used to gain more understanding of residual differences.

Graphs involving plots of simulated values against equivalent measured values can be useful and should show a straight line at an angle of 45° to the axes in the ideal case. Points above or below the 45-degree line give a measure of model discrepancies, provided experimental errors are negligible.

Wherever possible, estimates of measurement errors should be shown as error bars superimposed on the nominal values found experimentally (most often represented by discrete points). In experimental situations where different sensors are available for the measured variables, it may also be possible to check for consistency. Simple visual checks may highlight inconsistencies and computational tools for data evaluation and reconstruction, such as the Extended Kalman Filter, can then provide further checks. Situations in which data evaluation and reconstruction methods have been especially successful for validation of experimentally derived models include the testing of aircraft and helicopter models where flight trials can involve different types of inertial and air data sensors (e.g. Tischler and Remple 2006).

While a graphical output in terms of time history plots is easy to interpret for two or three variables, such comparisons become harder to use as the number of variables is increased. The simple time history plot is not the only graphical approach used to display and compare measured and simulated time histories. Other useful representations include a form of scaled polar diagram in which key quantities can be represented as points on radial lines. These points can be linked to create polygon figures and separate polygons of measured and simulated results on the same diagram then provide a measure of model and system agreement for the chosen quantities. This form of representation closely resembles the Kiviat diagram used for visualization of computer performance metrics and has proved useful for several simulation model applications (see, for example, Kammel et al. 2005; Smith et al. 2007; Murray-Smith 2015).

15.3.2 Some Quantitative Measures for System and Model Comparisons in the Time Domain

When data values generated from a simulation are compared with data from the corresponding target system, some specific quantitative measures of differences may be used instead of graphs and may be very useful in the validation of simulation models. For example, with n sets of values, differences between observed values y_i and simulated values $\hat{y}_i$ may be used to give the mean-squared difference:

$$J_{mse} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15.1}$$

There are many other possible measures and further discussion may be found in Chap. 18 by Saam in this volume which deals with broader issues associated with validation benchmarks and benchmark metrics.

One measure that has been used widely for some model validation applications is Theil's Inequality Coefficient (TIC) which is defined, for a single variable, as

$$J_{TIC} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2}} \tag{15.2}$$

This measure lies between zero and unity. A model capable of perfect predictions for the data used in the comparison gives a value of J_{TIC} of zero. On the other hand, a value of J_{TIC} of unity indicates complete failure in terms of model predictions. Although the useful range of J_{TIC} values for a credible model depends on the application and is inevitably subjective, values of J_{TIC} about 0.2 or 0.3 are often suggested as indicating acceptable agreement. This type of measure can also be used with more sets of variables, but a single number is then being used to describe the model quality for several different variables, leading to possible problems of interpretation.

All model quality measures of this kind, involving the use of one number, can mask differences that may be more obvious from graphical output. This has been discussed elsewhere by Jachner et al. (2007) and by the present author (Murray-Smith 2015) using specific examples.

15.3.3 Frequency-Domain Measures and Comparisons

An alternative to the methods of system and model comparisons based on time histories, which are outlined in Sect. 15.3.2, involves moving from the 'time domain' to the 'frequency domain'. Responses from the target system and model are compared in terms of their magnitude and phase (relative to some reference value) over a range of frequencies that is significant for the intended application of the model. The transformation of the system and model time histories to the frequency domain can be carried out using Fast Fourier Transform (FFT) algorithms and this is a routine procedure within the time series analysis and digital signal processing fields (see, e.g. Chatfield 1996).

Graphs of magnitude and phase against frequency for each relevant variable can often provide insight not so readily gained from time-domain analysis alone. Several quantitative measures for frequency-domain data are in common use and one such measure is the Frequency Response Assurance Criterion (FRAC) (Heylen and Lammens 1996). As with the Theil's Inequality Coefficient discussed in the previous section, this measure gives values lying between zero and unity. For identical frequency response functions, the FRAC value is unity, while unrelated responses give a value of zero. For structural dynamics and other engineering applications, such as pipeline modelling, the FRAC may have advantages over other approaches in that it makes direct use of measured frequency responses and thus captures the damping characteristics of the system, including nonlinear effects. Direct examination of measured and theoretical frequency response functions in graphical form may provide useful information concerning model validity. All frequency values that are likely to be important for the model must be included in such comparisons and this requires careful consideration when experiments are being planned. Any test input applied to the system and the model must include frequencies that cover the entire range of interest for the intended application.

Another measure which is widely used in experimental modelling is the 'ordinary' or 'magnitude squared' coherence function between two data sets (see, e.g. Priestley 1981). This provides a measure of the dependence between time histories as a function of frequency. The ordinary coherence function has a value zero when the two signals are unrelated and approaches a value of one as the two signals become more similar in form. In the case of input–output analysis of a physical system, measurement noise, un-modelled disturbances and nonlinearity are always present so that a value of ordinary coherence between input and output data sets of less than one may suggest that the input–output relationship is nonlinear, or that the output includes components from un-modelled inputs, including measurement noise. Evaluation of coherence from measurements on the target system can be very useful in checking that assumptions underlying theoretical models are justifiable. Interesting examples of the use of spectral methods and coherence functions in the context of model development and validation may be found in work on helicopter flight mechanics models (see, for example, Tischler and Remple 2006) and in neural system modelling (see, for example, Rosenberg et al. 1982).

15.4 System Identification and Parameter Estimation in Model Validation

15.4.1 A Brief Overview of System Identification and Parameter Estimation

Use of system identification and parameter estimation methods can lead to a functional or 'black box' type of model, obtained entirely from measured input and output data. Although black box models are commonly used within control algorithms that require regularly updated information about the system being controlled, they are less useful in design and in hypothesis testing where all available knowledge about the target system is usually included at the model formulation stage. In most cases where experimental modelling methods are used, the model is based on a theoretical structure derived from the application of generally accepted laws and principles, with only some unknown parameters remaining to be estimated using experimental data. The latter type of model usually involves a set of ordinary differential equations, partial differential equations or difference equations, together (possibly) with algebraic equations.

As well as providing estimates of model parameters, system identification techniques provide analytical and computational tools that offer important insight during the development of theoretically based models, including the validation stage. Classical methods of system identification may involve linear discrete-time models or continuous-time models. For nonlinear models having a structure derived from theory, parameter estimation is usually based on 'local' or 'global' optimization techniques. Local methods may not reach a true optimum, becoming stuck at a local extremum with the true maximum or minimum remaining undetected. Global methods, which can help to overcome such difficulties, involve use of random components to reduce the risk of the algorithm becoming trapped. These global methods include the use of evolutionary algorithms such as genetic algorithms (GA), genetic programming (GP) and simulated annealing (SA) (see, for example, Murray-Smith 2012).

Although widely used as a basis for experimental modelling of dynamic systems based on ordinary differential equations and algebraic equations (often termed 'lumped parameter' models), system identification methods have received less attention in the case of partial differential equation (PDE) models which are often termed 'distributed parameter' models. The use of system identification and parameter estimation techniques for distributed parameter models usually involves repeated solution of the PDEs, which is a computationally intensive procedure.

Theoretical finite-element models (which are much used in engineering in, for example, structural dynamics applications) can present difficulties since modelling errors and uncertainties are not easily estimated. However, experiments on a real structure may, in principle, be possible and measurements may be obtained with appropriate actuators and sensors mounted within the structure. Dynamic properties, such as eigenvalues and mode shapes, can be compared with equivalent measured quantities in the target system, and conclusions may be reached about the credibility of the theoretical model. Frequency-domain methods are often used to compare finite-element model responses with measurements. This approach can provide evidence about the parts of the frequency range over which model deficiencies are greatest. Such an approach is outlined in (Bryce et al. 1976; Murray-Smith 2015) in the context of a lumped parameter representation of large hydraulic pipelines which formed a sub-model within a larger model of a hydroelectric generator system.

One very important issue in system identification and parameter estimation is the accuracy of parameter estimates. Inevitably, this depends on the form of model since a large scatter in estimated parameter values often indicates errors in the model structure. In one approach, direct use is made of the statistical scatter for repeated parameter estimations. This is relevant when many sets of repeated test measurements are available for the same conditions but is seldom useful in practical applications involving continuous system simulation models since the number of repeated estimates is usually insufficient to provide statistically significant results.

A second way of assessing the accuracy of estimated parameters involves a more theoretical approach based on the Cramer–Rao Inequality which is used to define a Cramer–Rao bound (see, e.g. Söderström and Stoica 1989). This bound is always smaller than or equal to the standard deviation of the corresponding estimates that would be found from scatter analysis from many repeated tests. Although the bounds provide theoretical values for the standard deviations of parameter estimates, it must be noted that, with most system identification methods (e.g. the maximum-likelihood approach), the theoretical values may have to be multiplied by factors of five or ten to give realistic estimates of scatter. These factors allow for modelling errors and the effects of non-Gaussian noise. Noise effects can, of course, be reduced through filtering but the introduction of a factor of two is still recommended, even for cases where the noise is properly modelled or completely filtered out. Although variance values associated with parameter estimates are important indicators of model quality, it is should be noted that no direct comparisons can be made of variance values obtained for different model structures.

For the purposes of validation, techniques of system identification and parameter estimation provide an alternative to direct comparisons of system and model behaviour. These methods are appropriate for testing a model of an existing system and can be used when a prototype system or test rig is available. However, we must distinguish between system identification and parameter estimation processes used during model development to obtain information about the target system to be incorporated in the model and procedures involving system identification and parameter estimation to investigate the quality of an optimized model which may be based partly on experimentation and partly on theory.

Detailed information about methods of system identification and parameter estimation may be found elsewhere (see, for example, Ljung 1999; Nelles 2001; Raol et al. 2004). Different approaches have strengths and weaknesses which must be properly understood in the context of the intended application, not only for the estimation of poorly defined parameters or investigation of structural issues within a model under development but also for model validation. Software for system identification and parameter estimation is widely available on a commercial and open-source basis.

15.4.2 Issues of Identifiability

When using system identification and parameter estimation methods, the aim is to derive reliable estimates for all of the model parameters. This may not be feasible in practice and, in the testing of models, it is important to know the extent to which estimation of model parameters is theoretically possible. The concept of ‘identifiability’ provides a way of handling this and allows potential problems to be found before an identification method and test input signal are chosen. Although identifiability issues are of greatest importance in the context of identification and parameter estimation, they are also highly relevant for experimental design within the model validation process. For example, tests of identifiability can help to overcome the problem of ill-defined parameters which is an important issue when investigating the quality of large and complex simulation models.

Identifiability involves at least two different issues. Questions of 'global' or 'structural' identifiability are encountered when the number of parameters in a model is too large to allow them all to be found, whatever input is applied. This form of identifiability depends on the model structure and is a necessary condition for obtaining unique parameter estimates. Bellman and Åström (1970) highlighted the practical importance of global identifiability. They showed that expressing coefficients within a specific linearized form of a model in terms of the parameters appearing in the original equations may allow a set of nonlinear algebraic equations to be derived. It was shown that the linearized form of model is then identifiable, in the global sense, if these algebraic equations have a unique solution. Although it deals with a linearized version of the model, investigation of this form of identifiability can provide valuable insight about problems of parameter estimation for the original nonlinear model structure.

The second form is known as 'pathological' or 'numerical' identifiability. Unidentifiability of this kind arises if a model which is structurally identifiable is used with experimental data sets that are, in some way, inappropriate. This may happen if the test record is short compared with the dynamic properties of the system, such as dominant time constants or periods of any oscillations. Problems of pathological unidentifiability are also found if measured data are inaccurate due, perhaps, to unmodelled measurement noise. This form of unidentifiability has also been discussed by Brown and Godfrey (1978) who introduced the word 'determinancy' in describing it.

In simple cases, pathological unidentifiability may be established from time histories of parameter sensitivity functions. These functions provide an indication of how much each variable of the model is affected by changes of each parameter. Model parameters may be estimated only if parameter sensitivity functions for output variables with respect to each parameter are linearly independent (see, for example, Beck and Arnold 1977). This idea can be examined in more detail through properties of the 'sensitivity' matrix:

$$\boldsymbol{X} = \begin{bmatrix} \frac{\partial y_1}{\partial q_1} & \cdots & \frac{\partial y_1}{\partial q_p} \\ \vdots & \vdots & \vdots \\ \frac{\partial y_n}{\partial q_1} & \cdots & \frac{\partial y_n}{\partial q_p} \end{bmatrix} \tag{15.3}$$

and also using the closely related 'parameter information' matrix $\boldsymbol{M} = \boldsymbol{X}^T\boldsymbol{X}$, where the variables y_i are the outputs of interest and the parameter sensitivities $\frac{\partial y_i}{\partial q_j}$, as a function of time, give an indication of the extent to which the time history of variable y_i is influenced by a parameter q_j. The sensitivity matrix $\boldsymbol{X}$ allows parameter interdependencies to be found that are more complex than those that can be seen from direct inspection of time histories of model variables or their sensitivity functions. Specifically, pathological unidentifiability arises if columns of the matrix $\boldsymbol{X}$ are linearly dependent. This may also be seen from the condition number of the matrix $\boldsymbol{M}$ (the ratio of the largest eigenvalue of the matrix to the smallest eigenvalue) or from the value of the determinant of $\boldsymbol{M}$. If the condition number is large or if the determinant is small, the confidence region for the estimates is large and parameter estimates are poorly defined.

The inverse of the parameter information matrix ($\boldsymbol{M}^{-1}$) is the 'dispersion' matrix, denoted by $\boldsymbol{D}$. It provides a useful and highly practical indicator of pathological unidentifiability through the value of its determinant. This is because, for an efficient estimator (e.g. a maximum-likelihood estimator), elements of $\boldsymbol{D}$ are related to the variance of the estimated parameters, through the Cramer–Rao bound.

Inspection for situations involving significant correlations between pairs of parameters may be carried out using the 'parameter correlation' matrix which is denoted by $\boldsymbol{P}$ and defined in terms of its elements as

$$p_{ij} = \frac{m_{ij}^{-1}}{\sqrt{m_{ii}^{-1} m_{jj}^{-1}}} \tag{15.4}$$

where p_{ij} is the element of $\boldsymbol{P}$ in row i and column j and m_{ij}^{-1} is the element of $\boldsymbol{M}^{-1}$ in row i and column j. The diagonal elements of $\boldsymbol{P}$ are all unity and off-diagonal elements lie between -1 and 1. Small values of the off-diagonal elements of $\boldsymbol{P}$ show that the parameters are essentially decoupled, but a model is close to being unidentifiable if the modulus of one or more of the off-diagonal terms is close to unity, with 0.95 being widely regarded as a practical upper limit.

It should be noted that, in addition to providing an indication of the accuracy of estimates, large values of Cramer–Rao bound for specific parameters may also indicate problems of identifiability. Such a situation suggests that those parameters might be better viewed as fixed quantities within the model or could be removed completely.

15.4.3 Applications of System Identification and Parameter Estimation to the Processes of Validation

Although many methods of system identification are applicable only for linear models, these techniques can still provide useful insight for validation of nonlinear models. Using parameter estimation methods, linear models can be found for several test signal amplitudes and several points across the operating envelope of the system. Trends in values of estimated parameters can then be compared with trends in parameters of linearized theoretical descriptions derived from the nonlinear model for the same operating conditions. Differences in trends may then provide insight about limitations of the nonlinear model. This approach has been applied in testing nonlinear physically based helicopter flight mechanics models (Bradley et al. 1990).

Non-parametric methods of system identification, such as those based on frequency-domain identification techniques, may also provide empirical models to which physically based descriptions may be fitted. Thus, for example, transfer function estimates could be found experimentally, and parameter estimates could also be derived for an equivalent physically based model. Comparisons could be made between the experimentally estimated and theoretical values for these quantities within the model and sensitivity issues could be explored. In the case of single-input single-output linear models, many methods are available for fitting transfer function descriptions to time-domain or frequency-domain data obtained experimentally. For example, measurements using sinusoidal test signals allow amplitude ratios and phase shift of the measured response relative to the test input to be estimated directly. Spectral analysis methods may also be used, and broadband test signals are appropriate. However, any linear transfer function model found using experimental modelling methods in this way needs to be validated using a form of test input that is different from the one used for the identification process.

Large standard deviations in parameter estimates found using different test data sets suggest that there may be model structure problems or errors in the measured data. Parameter correlation issues can also limit the extent to which parameter estimates are useful for validation. In such situations, identifiability analysis may offer valuable insight. Examining special cases where there is good prior information available, such as a model parameter which depends on the acceleration due to gravity, can also be useful and any significant deviation from expected values should be examined carefully and explanations sought. It may be wise to redefine the experimental modelling problem, in some cases, so that well-known quantities are given fixed values and thus excluded from the parameter estimation process. Time series analysis methods (see, e.g. Priestley 1981; Chatfield 1996), can provide additional information about the significance of the experimentally derived evidence.

If adequate agreement between identified and theoretical models is found, the second stage of validation can be attempted, involving comparison of time histories for larger input perturbations. If the agreement is satisfactory over a range of operating conditions, the model may then be considered for acceptance.

Accepted models can be applied until a new situation is found where the model performance is considered inadequate. When such situations are encountered additional information or data must be sought and further model refinement undertaken, leading to further verification and validation tests.

When testing nonlinear models using large inputs, validation methods based on direct comparisons of model and system outputs are of limited value, whether based on graphical or quantitative measures. Holistic face validation methods involving the opinions of experts involved with the target system (e.g. pilots in the case of aircraft or operators in the case of industrial systems) may provide more useful insight since subjective testing allows attention to be focused on aspects of the model performance that are judged important. Model structure refinement and parameter adjustment processes must be used repeatedly until selected model time histories match those of the system to some predefined level. If the search space is small, we can examine possible solutions and find an adequate one by trial and error. However, such simple search procedures are unlikely to be useful in most practical modelling situations and computer-based optimization methods are usually needed.

For nonlinear simulation models, the selection of test conditions is more difficult than for linear descriptions because the system must be excited over the entire frequency range of interest and all the significant nonlinearities must also lie within the test signal range. Confidence intervals for the predictions made using a complete model are not generally available but, for nonlinear simulation models, some types of non-parametric description, such as Gaussian process models, do provide such information (see, for example, Kocijan et al. 2005; Thompson 2009).

The application of system identification and parameter estimation techniques may lead to the use of 'frequentist' methods (McFarland and Mahadevan 2008). Such statistical techniques involve hypothesis testing and may allow acceptance or rejection of a specific set of model parameters. Examples include the use of the multivariate Gaussian Hotelling T^2 test (Goodwin and Payne 1977) and more recent research on significance tests (see, for example, Huynh et al. 2012).

The spectral properties and amplitude distributions of data sets applied in validation must be similar to those applied during identification and parameter estimation. Although similar in terms of these properties, experimental test records for validation should differ in terms of their time histories from the records used for parameter estimation. As mentioned earlier, simulation output data being used for comparison with experiments should always be subjected to the same signal processing procedures (e.g. filtering) as the corresponding experimental data.

15.5 Design of Experiments and Selection of Inputs for Model Testing

Well-designed validation test specifications should include initial conditions and any boundary conditions, the form of inputs and the measuring equipment used. All relevant operating conditions and the full range of parameter values must be considered. Specification of the accuracy of measurements is also important since data provide the standard against which competing models are assessed.

Validation measures should be established at an early stage in the modelling process. These should be linked to model requirements and involve the main variables of interest for the eventual application. Model quality measures also need to take account of the potential difficulties and costs of data collection, both for the complete system model and for sub-models.

The design of experiments and test signals for system identification and validation must have a quantitative basis. Simple inputs such as steps or pulses are commonly used, but these are not 'persistently exciting' as they can involve significant time periods when inputs are constant or have zero value. There are also difficulties with test inputs based on sinusoidal signals where inputs with different frequencies are applied sequentially and the system must reach a steady state following each frequency change. Tests may also have to be repeated for a range of different input amplitudes. This can be time-consuming and, if there is insufficient information about the characteristics of the system, the applied inputs may not cover a wide enough range of frequencies and amplitudes initially and tests may have to be repeated.

The use of swept sine waves (also known as frequency-sweep inputs) eliminates some problems encountered with sine wave testing. With a swept sinusoidal test input, the frequency is varied from some initial low value to the highest frequency of interest, continuously over a period of time. This is a form of persistently exciting signal, and its use can significantly reduce the time required for testing compared with conventional sine wave signals.

The design of persistently exciting signals can also lead to square wave and binary multi-step inputs which involve irregular step-like changes between two predefined levels and, ideally, have no component at zero frequency (i.e. no steady offset). Pseudo-random binary (PRB) signals belong to this class and have an auto-spectrum like that of white noise over a limited frequency band.

Measures of model quality, such as those discussed in Sects. 15.3.2, 15.3.3 and 15.4.2, can be useful for experimental design and the selection of input test signals since, even when the form of test input has been decided, other properties (e.g. amplitude and frequency content, etc.) must still be chosen. For example, for models linearized about a specific operating point, the sensitivity matrix $\boldsymbol{X}$, the parameter information matrix $\boldsymbol{M}$ and the dispersion matrix $\boldsymbol{D}$ can all be used in establishing measures of the quality of experiments. The cost functions all involve a general expression:

$$J = f(M) \tag{15.5}$$

where f is a scalar function and J is the cost function. However, such measures depend on model parameter values and this means that experiment designs for model validation using such a cost function are never optimal. Several different performance measures of this kind are based on the dispersion matrix, $\boldsymbol{D}$. Examples include the A-optimal criterion based on minimization of the trace of $\boldsymbol{D}$, the D-optimal criterion in which the determinant of $\boldsymbol{D}$ is minimized and the E-optimal criterion which depends on the value of the maximum eigenvalue of the matrix $\boldsymbol{D}$. Some advantages have been claimed for the D-optimal criterion compared with the others (Hunter et al. 1969; Federov 1972). It has the form

$$J_D = \det(D) = \det\left(M^{-1}\right) \tag{15.6}$$

and gives the same level of emphasis to all parameters within a model. However, if only a small subset of the parameters are of interest use of a 'truncated' D-optimal criterion may be preferred (Hunter et al. 1969). This criterion has the form

$$J_{Dt} = \det\left(M_{ii}^{-1}\right) \tag{15.7}$$

where $\boldsymbol{M}_{ii}$ is a sub-matrix of the full information matrix involving only a subset of i parameters.

These two D-optimal criteria involve sensitivity matrix elements and thus depend on the model parameters. These measures can therefore only be used to assess and compare experimental designs in a general way. Optimal design is impossible due to the need for a perfect model, which is never available in practice.

15.6 Model Structure Optimisation

In simple terms, model complexity depends upon the number of equations and the number of adjustable parameters. The words 'under-fitting' and 'over-fitting' are widely used in discussions of model testing and validation. Under-fitting arises if a model has a structure that does not allow it to match observed test data adequately. If, on the other hand, a relatively complex model is used but predictions are poor, over-fitting may be the reason. An incorrect model structure is one possible underlying cause in both situations, but errors in the estimation of model parameters from experimental data due, for example, to bias caused by measurement noise is another possibility. Criteria commonly used for guidance in model selection include Akaike's information criterion (AIC) and the Bayesian modification of this (BIC). These criteria and other approaches which penalize models with many estimated parameters are discussed in many texts on time series analysis methods (see, e.g. Priestley 1981; Chatfield 1996). However, care must be taken in applying such approaches to deterministic models because many of these methods depend on an assumption that data are normally distributed.

In general, if any part of a model is derived using experimental modelling techniques, such as system identification and parameter estimation, the model should be assessed using a test data set that is different from any of the data sets used in model development, as already mentioned. This allows assessment of the predictive capability of the model for experimental situations that are different from those used in the development of the model.

15.7 Experimental Data for Validation: A Physiological Modelling Example

The example chosen to illustrate the importance of data in simulation model development involves a simple physiological model of human respiratory processes. It not only illustrates some aspects of validation, such as the insight provided by identifiability concepts and the importance of experimental design methods, but also provides an example of a data-driven simulation model. The research upon which this is based is not recent, but the example considered is ideal for the purpose since it involves a compartmental model structure and it is hoped that the main findings and conclusions are understandable for readers from all fields.

Dynamic compartmental models of pulmonary gas exchange processes that take place in the lungs and circulatory systems in humans are of interest, not only in terms of physiological knowledge but also for the development of diagnostic techniques. Several studies have used gas exchange models that incorporate periodic breathing and thus reflect, properly, the inspiratory and expiratory phases of the breath cycle (see, for example, Murphy 1969; Tomlinson et al. 1994; Hahn and Farmery 2003). The gas exchange processes may form a sub-model within a much larger model of the complete respiratory control system involving multiple feedback pathways from which the breath cycle pattern involved in lung ventilation is generated. However, in the model being considered here the requirement was for a representation of the pulmonary gas exchange processes in a form that could be applied to individual human subjects. The feedback control systems for the control of breathing are therefore not included in the model and the representation of the breathing cycle is based on measured ventilation data from the subject.

The background to this model has been described by Pack (1976), Bache (1981) and Bache et al. (1981). The model structure involves a rigid dead space compartment representing the upper airways and a single compliant alveolar compartment describing regions of the lungs in which gas exchange takes place between the air spaces and the blood. The model also includes a single compartment to represent the tissues, connected to the lungs through the bloodstream. The version of the model considered here involves carbon dioxide exchange and the structure is shown in Fig. 15.1. For the levels of partial pressure (concentrations) within the alveolar and tissue compartments arising in this application, the dissociation curves for carbon dioxide for mixed-venous and arterial blood are represented by a set of parallel straight lines.

Mass transfer principles are commonly used in establishing mathematical models within compartmental systems of this kind. Basically, in each compartment, the rate of change of mass of a specific substance must equal the difference between the input and output flow rates for that substance. This is a form of 'lumped parameter' model since, within each compartment, it is assumed that properties are uniform.

Considering the transfer of carbon dioxide between the compartments the following model may be derived (see, e.g. Bache et al. 1981):

$$\frac{dP_A}{dt} = \frac{S\dot{V}}{V_A}\left(P_I^* - P_A\right) + \frac{\dot{Q}}{V_A}[a_d + b(P_{TC} - P_A)]\gamma \tag{15.8}$$

$$\frac{dP_{TC}}{dt} = \frac{\dot{M}}{bV_{TC}} - \frac{\dot{Q}}{bV_{TC}}[a_d + b(P_{TC} - P_A)] \tag{15.9}$$

The variables $P_A(t)$ and $P_{TC}(t)$ represent the partial pressures of carbon dioxide in the alveolar and tissue compartments, respectively. The quantity $\dot{Q}$ is a fixed parameter which represents the mean blood flow through the lungs, V_A is the volume

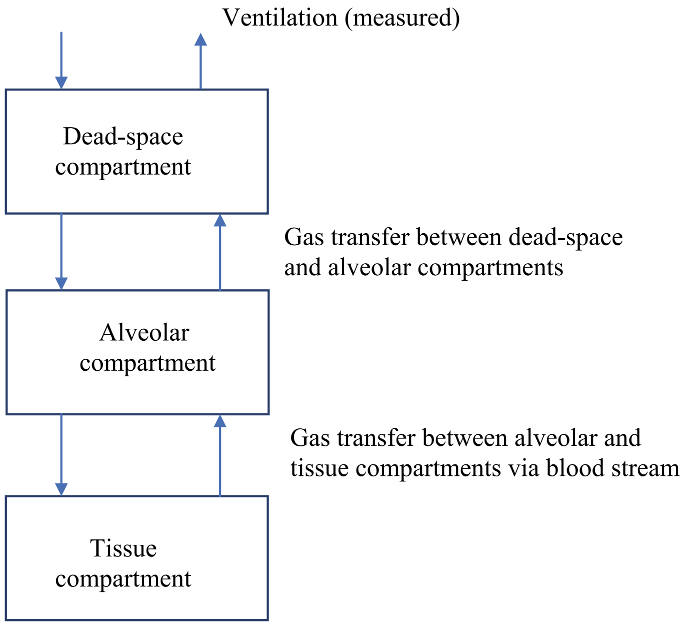

Ventilation (measured)
Dead-space compartment
Gas transfer between dead-space and alveolar compartments
Alveolar compartment
Gas transfer between alveolar and tissue compartments via blood stream
Tissue compartment

Fig. 15.1 Compartmental model representing pulmonary gas exchange processes

of the alveolar compartment about which volume changes occur during the different phases of the breath cycle and V_{TC} is the tissue compartment volume. The quantity $\dot{V}$ in (15.8) is the ventilation, which is a function of time and is a measured quantity, thus providing the data-driven element of the model, mentioned earlier. The quantities a_d and b which describe a linearized approximation to the dissociation curve are known, as is the constant γ.

The breathing cycle is divided into three stages. Stage 1 involves transfer to the alveolar compartment of gas which was in the dead space at the end of the previous breath cycle. The condition defining this stage is

$$\dot{V}(t) \geq 0 \text{ and } \int_{t_I}^{t} \dot{V}(t)dt \leq V_D \tag{15.10}$$

where t_1 defines the start time of the inspiratory phase. The quantity S in (15.8) is a switching variable which has value 1 when $\dot{V} \geq 0$ and value 0 when $\dot{V} \leq 0$.

The variable $P_I^*(t)$ has a form that varies over the breath cycle. During the first part of the inspiratory phase, gas entering the alveolar compartment is gas left in the dead space at the end of the previous expiration. Thus, throughout Stage 1, the variable $S =$ 1 and $P_I^*(t) = P_D(t)$ where $P_D\ (t)$ is the partial pressure of gas entering the alveolar $P_A\ (t)$ compartment and is taken to be the flow-weighted mean of the alveolar partial pressure during the final (end-tidal) portion of the previous expiration. The second stage of the breath cycle involves transfer of gas inspired during the current breath cycle to the alveolar compartment from the dead space compartment. The relevant condition is

$$\dot{V}(t) \geq 0 \,\text{and} \int_{t_I}^{t} \dot{V}(t)dt > V_D \tag{15.11}$$

The switching variable S remains unchanged at a value of 1 but the variable $P_I^*(t) = P_I(t)$ where $P_I(t)$ is now the partial pressure of carbon dioxide in the inspired gas mixture. In Stage 3, expiration takes place, for which the condition is

$$\dot{V}(t) < 0 \tag{15.12}$$

and throughout this phase of the breath cycle the variable $S = 0$.

This model structure describes gas exchange processes for human subjects with normal lungs and circulation for experiments of 10 min or less, not only for carbon dioxide but also for oxygen and other gases, provided modifications are made to the representation of the dissociation curve. Statements about the assumptions and approximations made in developing this lumped parameter compartmental model may be found elsewhere (see, e.g. Pack 1976; Bache 1981).

The ventilation $\dot{V}(t)$ can be related to an 'effective' ventilation $\dot{V}_E$ for the alveolar compartment by taking account of the dead space volume. For each inspiration, only part of the air breathed in at the mouth reaches the alveolar compartment and the effective ventilation under steady-state breathing conditions would be given by

$$\dot{V}_E = \dot{V} - fV_D \tag{15.13}$$

where V_D is the volume of the dead space compartment and f is the breathing frequency.

A second important variable is the partial pressure of carbon dioxide in the inspired mixture $P_I(t)$ and this, together with the ventilation $\dot{V}(t)$, provides a basis for generating test inputs to the system.

The partial pressure of carbon dioxide in the alveolar compartment $P_A(t)$ forms an output variable. Direct measurement of $P_A(t)$ in human subjects presents difficulties but an estimate is available from measurements, at the mouth, of carbon dioxide partial pressure during the final ('end-tidal') phase of each expiration. The variable $P_{TC}(t)$ in (15.9) may be viewed as being equivalent to the partial pressure of carbon dioxide in mixed-venous blood and can also be regarded as an output. It cannot, however, be measured readily on a continuous basis.

Investigation of global identifiability issues for this model may be carried out using a linearized version of the model of (15.8)–(15.13) written as

$$\frac{d^2P_A}{dt^2} + a_1\frac{dP_A}{dt} + a_2P_A = b_1\frac{du}{dt} + b_2u(t) + terms\ independent\ of\ u(t) \tag{15.14}$$

where the variable $u(t)$ is an input variable defined in terms of the effective ventilation variable $\dot{V}_E(t)$ and the partial pressure of carbon dioxide $P_I(t)$ in the inspired mixture. Details of this linearization process may be found elsewhere (see, e.g. Bache and Murray-Smith 1983; Murray-Smith 2015).

Equation (15.14) can be used to establish the global identifiability of the model and which parameters may be estimated independently and thus determines the scope for validation. The coefficients within the numerator and denominator of each term on the right-hand side of (15.14) can be expressed in terms of the parameters of (15.8)–(15.13). For example,

$$a_1 = \frac{\dot{Q}}{V_{TC}} + \frac{\dot{V}_E}{V_A(0)} + \frac{k\dot{Q}b}{V_A(0)} \tag{15.15}$$

$$a_2 = \frac{\dot{Q}\dot{V}_E}{V_{TC}V_A(0)} \tag{15.16}$$

$$b_1 = \frac{1}{V_A(0)} \tag{15.17}$$

$$b_2 = \frac{\dot{Q}}{V_{TC}V_A(0)} \tag{15.18}$$

Since the effective ventilation $\dot{V}_E$ is a measured quantity, the set of algebraic equations can be manipulated to give unique expressions for each of the parameters of (15.8) and (15.9), provided the parameters a_d and b describing the dissociation curves for carbon dioxide in mixed-venous and arterial blood are known, together with the constant γ. Since these three parameters can be determined independently, the model is thus globally identifiable.

Although global identifiability is established, there could also be issues of pathological unidentifiability for this model. Pathological unidentifiability may be detected retrospectively by looking for linear dependence of columns of the sensitivity matrix $\boldsymbol{X}$, as defined in (15.3). Unidentifiability or near-unidentifiability is also reflected in the condition number of the parameter information matrix $\boldsymbol{M}$ and this matrix should therefore also be examined. If pathological unidentifiability is encountered, careful consideration should be given to noise levels and to the design of the experiment from which the measured response data are obtained.

15.7.1 Experimental Constraints

Concentrations of inspired carbon dioxide must be constrained since levels above normal can only be tolerated for short periods of time. In addition, the cardiac output, which corresponds to the parameter $\dot{Q}$ (assumed constant in the model), is affected by the inspired carbon dioxide concentration. However, carbon dioxide concentrations of 5% or less have little influence on the cardiac output if breathed for only a few minutes. Therefore, the chosen gas mixture involved 5% carbon dioxide, 21% oxygen and 74% nitrogen, with a maximum test duration of 10 min (Bache et al. 1981).

Test inputs intended for routine clinical investigations need to be applied in a straightforward fashion. Input perturbations, in this case, involved manual switching of the inspired gas between dry air and input from a cylinder containing the chosen gas mixture, with the switching frequency under the control of the experimenter.

The tests started with one minute of air breathing which was intended to allow the subject to become comfortable. Following this, the chosen pattern of input was applied, with gas concentration and flow rate data being sampled ten times per second. All channels of data were passed through suitable low-pass filters. Details of the signal processing and parameter estimation procedures are given elsewhere (Bache 1981; Bache et al. 1981).

15.7.2 Experimental Design and Test Signal

A maximum-likelihood approach (see, e.g. Ljung 1999, Söderström and Stoica 1989) was chosen for parameter estimation but success in applying this method was found to be highly dependent on experimental design. Tests using a step function form of input involving the subject breathing air for 40 s and then being switched to a gas mixture containing 7% carbon dioxide for a further 80 s (i.e. a form of step function input) produced large off-diagonal values in the parameter correlation matrix for parameter pairs $\dot{M}$ and V_{TC} (0.999), $\dot{M}$ and V_A (0.840), V_A and V_{TC} (0.841) and $\dot{Q}$ and P_{TC} (0) (−0.952). This suggests that a step function input leads to problems of pathological unidentifiability and that a more persistently exciting input could be a better choice (Bache et al. 1981). The input must have suitable frequency content and the concentration of carbon dioxide and other gases in the inspired gas mixture and the test duration must take account of the experimental constraints discussed above.

Persistent excitation could involve periodic switching between air and the chosen mixture of gases with an appropriate switching pattern. Suitable forms for this binary signal were investigated using measures involving the parameter information matrix, the dispersion matrix and the parameter correlation matrix, as discussed in Sect. 15.4.2. Elements of these matrices depend on model parameters and, therefore, analysis aimed at finding suitable switching periods for parameter estimation, or for model validation, must be based on some initial model. Since parameter values are unknown, initially, a nominal set of parameters for a ‘normal’ human subject of average build was chosen for the experiment design process.

One case involved finding the best form of input for simultaneous estimation of all the parameters of the model using data from a single system identification test. The design procedure in this case involved the D-optimal criterion discussed in Sect. 15.5, with the cost function of (15.6). For resting conditions with a typical subject, the optimum switching period was found to be approximately 55 breaths.

A second case involved finding a persistently exciting input for estimation of individual parameters. The truncated D-optimal design approach and the cost function of (15.7) suggested that estimation of the tissue compartment parameters $\dot{M}$ and V_{TC} requires a switching period that was longer (say 60 breaths) than the period found for estimation of all the parameters simultaneously. On the other hand, for the parameter $V_A(0)$ (the initial steady-state volume of the alveolar compartment), a higher switching frequency, corresponding to a switching period of about 15 breaths, was shown to be desirable. This is consistent with physiology since the alveolar compartment dynamics are fast compared with those of the tissue compartment and the parameter $V_A(0)$ is a factor within a time constant for the alveolar compartment in the linearized model. Estimation of the parameter $\dot{Q}$, which represents the total blood flow through the lungs (the cardiac output for normal subjects), involved a clear optimum with a switching period of about 24 breaths.

Since the design is dependent on model parameter values, which are initially unknown, the results above can be used only as a guide in selecting test signals. The form of input chosen involved alternating periods of air and gas mixture involving 5% or 7% carbon dioxide, with an appropriate switching period depending on the parameters of interest. Measured quantities were the ventilation and carbon dioxide partial pressure at the mouth. All subjects had been previously assessed using other clinical tests and were judged to be normal in terms of lung function. Estimated values for each parameter of the nonlinear model were incorporated into the simulations model for each patient.

Validation of the models developed for each subject involved a separate set of experimental results in which a different form of persistently exciting test input was applied. This persistently exciting input involved a switching period of about 30 breaths for adults (approximately two minutes). This input was again chosen using results of the D-optimal and truncated D-optimal test input design optimization procedures outlined above. The validation test results also showed good agreement between measured and simulated outputs, with uncorrelated residuals (Bache et al. 1981). Analysis of sets of results obtained using the model validation test input also gave small off-diagonal elements in the parameter correlation matrix, indicating small interactions between parameters. For example, the results for one typical subject the element showing interactions between parameters $\dot{M}$ and V_A was small (0.067) compared to the result obtained using the step input (0.840). For parameters $\dot{M}$ and V_{TC}, the correlation matrix element was -0.017, which was also small compared to the value of 0.999 for the case of the step input (Bache et al. 1981; Bache and Murray-Smith 1983). Similar results were found for other validation data sets. This information, combined with the satisfactory values found for residuals, suggested that the chosen form of test signal could ensure that pathological unidentifiability issues were avoided and could provide reliable parameter estimates. The resulting simulation models were therefore judged to be suitable for the intended application for each of the subjects tested.

It is important to be able to distinguish between normal and abnormal patients in the clinical testing of lung function. Abnormalities may sometimes be associated with gas exchange inhomogeneity and lumped parameter gas exchange models of the type described here may be used to describe any maldistribution of ventilation or blood flow. One approach has involved the introduction of additional alveolar compartments, and a model similar to that of Fig. 15.1 was developed having two alveolar compartments. Global identifiability analysis applied to this slightly more complex model structure showed that it was now impossible to distinguishing between different models of this kind from measured ventilation and gas concentration data at the mouth. Estimation of parameters within this model structure was impossible without additional measured variables that could only be accessed in more invasive fashion (Bache 1981; Bache and Murray-Smith 1983). Also, it was shown that meaningful validation tests would not be possible for this modified model structure using the available measured variables.

15.8 Discussion

In the validation of deterministic dynamic models, emphasis is always placed on the assessment of the overall credibility and quality of a given model, developed for a stated set of requirements and for a specified application. Thus, the successful development of a simulation model requires close liaison between those responsible for drawing up the model requirements and those responsible for testing the resulting model through the processes of verification and validation.

Consideration of issues of availability and accuracy of measured data is an important part of the planning process for the validation procedures. There are many currently available methods for the validation of simulation models that could be used to better effect. Too many simulation model developers and users are content with superficial tests involving simple graphical comparisons. More attention should be given to issues of experiment design. Model properties, such as identifiability, could be used to provide significant additional insight.

More generally, it is fair to state that within many organizations the introduction of improved procedures for the management of simulation models and measured data from target systems, together with improved processes for the planning and execution of model tests and validation procedures could provide immediate benefits. Careful documentation of models and the associated data sets is also very important (Murray-Smith 2015; see also Chap. 25 by Reinhardt et al. in this volume).

The example involving the pulmonary gas exchange model was included to provide an illustration of simulation model development processes involving theoretical model development based on physical laws and principles, together with experimental modelling procedures for estimation of parameters and for validation. The example also illustrates how identifiability concepts may be used for investigation of parametric coupling and for experiment design. In that project, the primary requirement in terms of validation was that the test input applied should be capable of fully exciting the system over the relevant frequency range and covering the full operating range for each of the variables of the model while taking account of experimental constraints. The resulting test input applied to system and model provided results showed good agreement between the model and the system for a number of human subjects. That case study also provided an illustration of the use of identifiability analysis to investigate how available measurements could limit the possibilities for validation of more complex model structures involving additional variables. In such cases, identifiability analysis could establish the variables of the target system that would have to be measured to allow validation to be applied successfully.

15.9 Conclusions

In cases where experimental testing of the target system is possible, considerable additional insight can be gained if the experiments on the target system are designed appropriately and the test inputs applied cover the full operating range of the system in terms of both amplitudes and frequencies. It is always important to ensure that as much useful information as possible is included within any data sets from the target system that may be used for simulation model validation. It is suggested that the requirements for experimental design in simulation model validation are essentially the same as the requirements for experiment design for system identification and parameter estimation. Thus, a well-designed experiment for system identification and parameter estimation should also be a good experiment for the purposes of model validation and established methods for experiment design from that field can therefore be applied to simulation model validation.

It is clear that the same data set should never be used in both model development and model testing. In general, measured data sets should be split into two separate records, one being used for model development tasks, such as parameter estimation and the other for validation. When using data obtained by others, it is important to be aware of the source of the data, any preprocessing that may have been carried out (such as filtering or averaging) and possible errors in measured values.

Identifiability analysis is a particularly important tool that can be used to avoid fruitless attempts to separate effects of parameters that are inherently coupled. Identifiability also has a role in experimental design and especially in the selection of test inputs which ensure that the important parts of the frequency and amplitude ranges are covered. Historical data and data obtained from tests carried out for purposes other than system identification or validation may be of limited value.

System identification and parameter estimation, which are central to experimental modelling, are relatively mature and well-understood methods that are routinely used in specialist application areas such as control engineering. However, there are interesting research developments under way at present which relate to experimental modelling methods and model validation for some other types of application. For example, as mentioned previously, one area of current research concerns the use of Gaussian process descriptions which can provide useful information about confidence intervals within identified models.

In conclusion, the process of model validation gives rise to an obvious dilemma since models must be available for experimental design and this is unlikely at the early stages of the model development cycle. Thus, in applications where tests on the target system are possible, initial experiment design must be based only on approximate models. These approximate models may be refined later in the validation process and, in some cases, the experiments themselves may have to be repeated. This transition from approximate models to more refined models and improved experiments means that data collection and analysis for model testing and validation is an iterative process.

References

Bache, R. A. (1981). *Time-domain system identification applied to non-invasive estimation of cardio-pulmonary quantities*. Ph.D. thesis, University of Glasgow, UK.

Bache, R. A., Gray, W. M., & Murray-Smith, D. J. (1981). Time-domain system identification applied to non-invasive estimation of cardio-pulmonary quantities. *IEE Proceedings*, *128*, Part D, 56–64.

Bache, R. A., & Murray-Smith, D. J. (1983). Structural and parameter identification of two lung gas-exchange models. In G. C. Vansteenkiste & P. C. Young (Eds.), *Modelling and data analysis in biotechnology and medical engineering* (pp. 175–188). Amsterdam, The Netherlands: North-Holland.

Beck, J. V., & Arnold, K. J. (1977). *Parameter estimation in science and engineering*. New York, NY: Wiley.

Bellman, R., & Åström, K. J. (1970). On structural identifiability. *Mathematical Biosciences, 7,* 329–339.

Bradley, R., Padfield, G. D., Murray-Smith, D. J., & Thomson, D. G. (1990). Validation of helicopter mathematical models. *Transactions of the Institute of Measurement and Control, 12,* 186–196.

Brown, F., & Godfrey, K. R. (1978). Problems of determinacy in compartmental modeling with application to bilirubin kinetics. *Mathematical Biosciences, 40,* 205–224.

Bryce, G. W., Foord, T. R., Murray-Smith, D. J., & Agnew, P. (1976). Hybrid simulation of water-turbine governors. In R. E. Crosbie & J. L. Hay (Eds.), *Simulation councils proceedings series* (Vol. 6(1), pp. 35–44). La Jolla CA: Simulation Councils Inc.

Chatfield, C. (1996). *The analysis of time series. An introduction* (5th ed.). London, UK: Chapman and Hall.

Federov, V. V. (1972). *Theory of optimal experiments*. New York, NY: Academic Press.

Goodwin, G. C., & Payne, R. L. (1977). *Dynamic system identification: Experiment design and data analysis*. New York, NY: Academic Press.

Gustavsson, I. (1972). Comparison of different methods for identification of industrial processes. *Automatica, 8,* 127–142.

Hahn C. E. W. & Farmery, A. D. (2003). Gas exchange modelling: no more gills please. *British Journal Anaesthesia, 91*(1): 2–15.

Heylen, W., & Lammens, S. (1996). FRAC: A consistent way of comparing frequency response functions. In M. I. Friswell & J. E. Mottershead (Eds.), *Proceedings of Conference on Identification in Engineering Systems, Swansea, 1996* (pp. 48–57). Swansea, UK: University of Wales.

Hughes, I. G., & Hase, T. P. A. (2010). *Measurements and their uncertainties: A practical guide to modern error analysis*. Oxford, UK: Oxford University Press.

Hunter, W. G., Hill, W. J., & Henson, T. L. (1969). Designing experiments for precise estimation of $^{\beta}$some of the constants in a mechanistic model. *Canadian Journal of Chemical Engineering, 47,* 76–80.

Huynh, D. P. B., Knezevic, D. J., & Patera, A. T. (2012). Certified reduced basis model characterization: a frequentistic uncertainty framework. *Computer Methods in Applied Mechanics and Engineering, 201,* 13–24.

Jachner, S., van den Boogaart, K. G., & Petzoldt, T. (2007) Statistical methods for the qualitative assessment of dynamic models with time delay (R Package qualV). *Journal of Statistical Software*, *22*(8).

Kammel, G., Voigt, H. M., & Neβ, K. (2005). Development of a tool to improve the forecast accuracy of dynamic simulation models for the paper process. In: J. Kappen, J. Manninen, & R. Ritala (Eds.), *Proceedings of Model Validation Workshop, 6th October 2005, Espoo, Finland*. Espoo, Finland: VTT Technical Research Centre.

Kocijan, J., Girard, A., Banko, B., & Murray-Smith, R. (2005). Dynamic system identification with dynamic processes. *Mathematical and Computer Modelling of Dynamic Systems, 11,* 411–424.

Ljung, L. (1999). *System identification: Theory for the user* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

McFarland, J., & Mahadevan, S. (2008). Multivariate significance tests and model calibration under uncertainty. *Computer Methods in Applied Mechanics and Engineering, 197*(29–32), 2407–2479.

Murphy, T. W. (1969). Modelling of lung gas exchange - mathematical models of the lung. *Mathematical Biosciences, 5,* 427–447.

Murray-Smith, D. J. (2012). *Modelling and simulation of integrated systems in engineering: Issues of methodology, quality, testing and application*. Cambridge, UK: Woodhead Publishing.

Murray-Smith, D. J. (2015). *Testing and validation of computer simulation models: Principles, methods and applications*. Cham, Switzerland: Springer.

Nelles, O. (2001). *Nonlinear system identification*. Berlin, Germany: Springer.

Pack, A. I. (1976). *Mathematical models of lung function*. Ph. D. thesis, University of Glasgow, UK.

Priestley, M. B. (1981). *Spectral analysis and time series*. San Diego, CA: Academic Press.

Raol, J. R., Girija, G., & Singh, J. (2004). *Modelling and parameter estimation of dynamic systems*. IET Control Engineering Series No. 65. London, UK: IET.

Rosenberg, J. R., Murray-Smith, D. J., & Rigas, A. (1982). An introduction to the application of system identification techniques to elements of the neuromuscular system. *Transactions of the Institute of Measurement & Control, 4,* 187–201.

Smith, M. I., Murray-Smith, D. J., & Hickman, D. (2007). Verification and validation issues in a generic model of electro-optic sensor systems. *Journal of Defense Modeling and Simulation, 4,* 17–27.

Söderström, T., & Stoica, P. (1989). *System identification*. New York, NY: Prentice-Hall.

Tischler, M. B., & Remple, R. K. (2006). *Aircraft and rotorcraft system identification*. USA: AIAA.

Thompson, K. R. (2009). Implementation of gaussian process models for non-linear system identification. Ph.D. thesis, University of Glasgow.

Tomlinson, S. P., Tilley, D. G., & Burrows, C. R. (1994). Computer simulation of the human breathing process. *IEEE Engineering in Medicine and Biology Magazine, 13,* 115–124.

Chapter 16
How to Use and Derive Stylized Facts for Validating Simulation Models

Matthias Meyer

Abstract Stylized facts are often evoked in a very casual manner in the context of model validation, a practice that stands in contrast with their popularity and their potential. Given this situation, it is the aim of this chapter to characterize the concept of stylized facts from an epistemological perspective, which includes clarification of interesting ideas behind the concept, development of a definition and discussion of its possible uses for validating simulation models. The latter includes not only output validation but input validation and theory validation as well. Second, the need for a more systematic derivation of stylized facts is addressed. For this purpose, several approaches to establish stylized facts are presented and assessed. In this context, an additional approach is presented, which tries to overcome the shortcomings of current practice. Together, these interrelated aims represent an attempt to strengthen the basis for the validation of simulation models using stylized facts.

Keywords Assumptions · Methodology · Simulation models · Stylized facts · Validation

16.1 Introduction

Extant literature in economics and other social sciences includes widespread references to stylized facts as support for the validity of models; a typical example of which is found in a paper by Kahn (1987) in the American Economic Review. Kahn introduces his paper with a stylized fact about the inventory behavior of firms, according to which it is a well-established empirical observation that "the variance of production exceeds the variance of sales". The rest of his paper is organized around this stylized fact. The paper begins with a discussion of how far the existing approaches can explain this stylized fact. Then, it develops a model which, according to Kahn, allows accounting for this stylized fact. As he can reproduce this stylized fact with

M. Meyer (✉)
Hamburg University of Technology, Hamburg, Germany
e-mail: matthias.meyer@tu-hamburg.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_16

his model, Kahn considers this to be support for its empirical validity. This reference to stylized facts to support the validity of models is quite frequent in economics and other social sciences (Troitzsch 2008, p. 667, see also Chap. 31 by Fagiolo et al. in this volume).[1]

The popularity of the concept of stylized facts can be observed in the area of simulation as well. For example, a review on calibration and validation approaches to agent-based models (Windrum et al. 2007) shows that stylized facts play a prominent role in the primary alternatives currently discussed for model validation.[2] Stylized facts are used as an empirically based reference point for simulation model calibration and validation in (a) the indirect calibration approach; (b) the Werker–Brenner calibration approach; and (c) the history-friendly approach (Windrum et al. 2007, paragraphs 4.4–4.25). With respect to the level of simulation models addressed, stylized facts are often used to specify regularly observed patterns at the societal or macro level (Gilbert 2008; Grimm et al. 2005), but can also be used to specify relevant characteristics at the level of agents (Grimm et al. 2005).

This widespread use of stylized facts and its prominent role in the process of model validation is, however, in contrast with the amount of attention typically devoted to the concept. Most authors still refer—if at all—to Kaldor (1968), who uses the concept primarily for rhetorical purposes and who, moreover, is not explicit as to how he derives his set of stylized facts. Such an approach leads to at least two problems.[3] First, there seems to be no clear and shared understanding concerning the epistemological characteristics of stylized facts. This is reflected in the fact that, thus far, few authors discuss the interesting ideas behind the concept of stylized facts from a general, epistemological perspective (e.g., Boland 1994; Lawson 1989; Schwerin 2001). Second, in most applications, stylized facts are simply "stated" (in the best case with reference to some supporting sources) and thus far, only a few studies attempt to systematically derive stylized facts (e.g., Schwerin and Werker 2001; Cont 2001; Heine et al. 2005). Such approaches can easily lead to credibility problems concerning the respective sets of stylized facts. Given these two problems, any attempts to validate simulation models based on stylized facts are in danger of criticism as to lack of solid basis. In other words, one may question whether stylized facts can be considered to provide a solid epistemological basis for validating simulation models.

Given this situation, it is the aim of this chapter to elaborate on Kaldor's original contribution and to provide an epistemological discussion of the concept of stylized facts for use in model validation. In particular, the present chapter addresses the two

[1] A search in Google Scholar for the English and American version of the term "stylized facts" results in a total of 98,800 hits. [Query 10.07.2018]

[2] For the purposes of this paper, validation is defined as the process of determining whether a simulation model is an accurate representation of the system for the objectives of the study. For a similar understanding, see Law (2006) and Burton and Obel (1995). This paper mainly uses agent-based models as examples, but its ideas can be transferred to other simulation modeling techniques as well.

[3] For a more general critique concerning the replication of stylized facts see Lux and Zwinkels (2017).

problems described above. First, it aims to characterize the concept of stylized facts from an epistemological perspective, which includes clarification of interesting ideas behind the concept, offering a precise definition and discussing the implications for the validation of simulation models. Second, the need for a systematic derivation of stylized facts is addressed. For this purpose, several approaches to derive stylized facts are presented and assessed. Together, these interrelated contributions aim to strengthen the epistemological basis for validating simulation models via stylized facts.

The chapter is structured as follows. The second section addresses the origin and epistemological foundations of the concept of stylized facts. In this section, several uses of stylized facts for simulation model validations are also specified. In the third section, the process of establishing stylized facts and existing approaches in this respect are presented, which are, at present, scattered throughout different literature streams. Based on the problems associated with these different approaches, a fourth approach is developed in the fourth section of the present chapter. The chapter concludes with a summary of the main results.

16.2 Epistemological Foundations of Stylized Facts

16.2.1 Development and Definition of the Stylized Facts Concept

The concept of stylized facts is introduced by the economist Kaldor in the context of debate on growth theory in 1958. In particular, he uses the concept to convince his peers that his model of economic growth and capital accumulation is a better representation of reality than the existing neoclassical models of his time (Kaldor 1968, p. 179).[4] A closer look at the strategy behind his argument helps to obtain a first-hand impression of the basic ideas behind the concept.

Kaldor starts by arguing that theories and corresponding models are based on abstractions, which must be appropriate to the characteristic features of a phenomenon “as recorded by experience” (Kaldor 1968, p. 178). For Kaldor, this implies that theory and model development should start with a summary of the facts that can be regarded as relevant for the problem under investigation. The problem is, however, that “facts as recorded by statisticians, are always subject to numerous snags and qualifications, and for that reason are incapable of being summarized” (Kaldor 1968, p. 178).[5] To handle this problem, Kaldor suggests that theorists “should be free

[4] Like many economists, Kaldor uses the terms “model” and “theory” interchangeably. For the purposes of this paper, a model is understood as an isomorphic or homomorphous formal mapping of a real system. Similarly, a theory is understood as a system of statements (theorems, axioms, hypotheses, assumptions) on empirical relationships. For a detailed discussion of these two terms and their usage, see Morgan (1998).

[5] For this problem, see also Leamer (1983, pp. 42–43), with additional references.

to start off with a stylised view of the facts—i.e., concentrate on broad tendencies, ignoring individual detail" (Kaldor 1968, p. 178). With respect to the resulting, broad tendencies that emerge from different sources of empirical data, Kaldor coins the term "stylized facts." These "stylized facts" form the reference point for theory construction (Kaldor 1968, p. 178). Based on these general methodological considerations, Kaldor names, in the second step, six stylized facts of macroeconomic growth[6] and argues that none of the neoclassical models of his time explain these stylized facts (Kaldor 1968, p. 179). The remainder of his paper is devoted to constructing a model that at least addresses some of these stylized facts.

Kaldor's exposition had a considerable impact on the discussion of growth theory and the term "stylized fact" can often be found as an example of an early research agenda in this field.[7] However, one can argue that the approach outlined by Kaldor can be fruitful also in more general contexts. In particular, his argument entails two interesting ideas from a methodological perspective.[8] First, he suggests an interesting way to link empirical research with research based on modeling. Stylized facts can be interpreted as stepping stones linking models with existing empirical observations concerning the phenomenon investigated. The aim is to identify characteristics of a phenomenon that are broadly supported by the existing empirical evidence in such a way that researchers can regard it as important enough to require an explanation (see also Lawson 1989, p. 65). This "summary" of existing evidence provides a means to formulate an empirically grounded research agenda for a research community. Second, he introduces the idea of a "stylized view" to deal with the diversity of empirical findings and the associated problems in summarizing the existing empirical evidence.[9] He suggests to ignore minor and even contradictory details and to focus on broad tendencies over the diverse empirical findings. The aim is to formulate broad generalizations over the empirical data that need not be universal.[10] Based on these two basic ideas, the following definition is suggested: stylized facts are broad, but not necessarily universal generalizations of empirical observations and describe the supposed essential characteristics of a phenomenon that requires an

[6] "As regards the process of economic change and development in capitalist societies, I suggest the following "stylized facts" as starting point for the construction of theoretical models: (1) The continued growth in the aggregate volume of production and in the productivity of labor at a steady trend rate; no recorded tendency for a falling rate of growth of productivity. (2) A continued increase in the amount of capital per worker, whatever statistical measure of "capital" is chosen in this connection" (Kaldor 1968, p. 178).

[7] See Schwerin (2001). However, in other areas, this term also can be found regularly. Boland (1997, p. 243) critically remarks that, in addition to widespread use of the term, little of Kaldor's idea is understood today.

[8] For a discussion of the concept, see also Boland (1994, 1997), Lawson (1989) and Schwerin (2001). For detailed discussion from a methodological perspective, see also Heine et al. (2007).

[9] This becomes even more intricate when different empirical research methods are involved, including qualitative designs. For a discussion of these issues, see Sects. 16.3 and 16.4.

[10] As these broad generalizations do not claim universality, this approach is not directly confronted with the problem of induction. Nevertheless, in line with Popper (1959), the importance of critical discussion in deriving and/or revising stylized facts should be emphasized. Similarly, stylized facts can be refined and/or revised over time like hypotheses.

explanation (Heine et al. 2007).[11] Given this function as a reference point for the research community, they should be based on a high level of consensus.[12]

A discussion of the concept of stylized facts also has to address briefly how the concept and related ideas are applied nowadays in economics and other sciences. While some researchers explicitly use this term (Aoki 1988; Marcet and Nicolini 2003), others use different terms but often mean the same or a related concept. Examples are "patterns" (Grimm et al. 2005; Hayek 1964), empirical regularities (Constance 2007) or statistical properties of a phenomenon (Cont 2001). A characteristic common to many, but not all is that many of the researchers stress the complexity of the phenomenon under investigation. A very early example in this respect is Hayek. More recently, this claim is central to the econophysics position. They identify a number of stylized facts of financial markets and state that some of them such as fat tails are an indicator of the complexity of these phenomena (Cont 2001).

An important area where the concept is used quite regularly is macroeconomics. In this area stylized facts are currently understood as broad, but robust enough, statistical properties pertaining to a certain economic phenomenon (Marcet and Nicolini 2003, pp. 245–290). But also other fields such as management (Constance 2007; Heine et al. 2005) or information systems research (Houy et al. 2015) start to use this concept. Moreover, it is used to map scientific progress in a research area using modeling as research method and this map can be used as a guide for future model development (Heine et al. 2007; Meyer 2011).

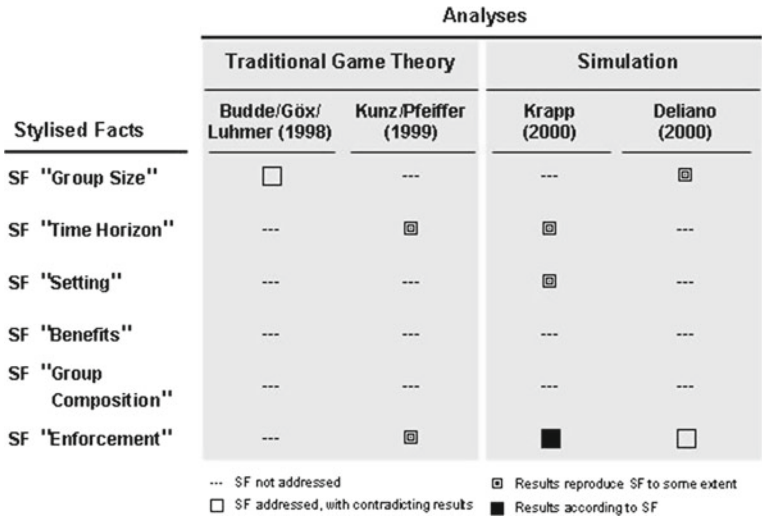
16.2.2 Using Stylized Facts for Simulation Model Validation

Having a set of stylized facts can not only provide a reference point for model construction ex ante (as Kaldor suggests) but also for the purpose of model validation. With reference to a given set of stylized facts of a phenomenon, models and their basic assumptions can be analyzed comparatively, with focus on their productive implications without distraction by other, minor issues that are also covered by the models: "[A]s long as we can come to an agreement regarding the "stylized" facts, the comparative appropriateness of competing explanatory abstractions can be brought into clear and decisive focus" (Boland 1994, pp. 535–536).

Two steps can be distinguished in such a model assessment. In the first step, the contributions of the models are measured. From the stylized facts perspective, the contribution of a respective model is in its ability to explain the stylized facts of a phe-

[11] This is already an interpretation and elaboration of Kaldor's original ideas. For example, it might be of interest to explore in which respect his concept of stylized facts can be regarded as the construction of phenomena in the sense of Bogen and Woodward (1988).

[12] For an earlier definition, see Lawson (1989, p. 65) and Schwerin (2001). The aspect of consensus is not sufficiently emphasized in earlier definitions, although very different philosophers, such as Popper, Peirce or Habermas emphasize the importance of social and discursive processes that enable criticism and lead to a convergence in the beliefs of scientists (Hands 2001, pp. 218–221). The issue of consensus is addressed further in Sects. 16.3 and 16.4.



	Analyses			
	Traditional Game Theory		Simulation	
Stylised Facts	Budde/Göx/ Luhmer (1998)	Kunz/Pfeiffer (1999)	Krapp (2000)	Deliano (2000)
SF "Group Size"	□	---	---	▣
SF "Time Horizon"	---	▣	▣	---
SF "Setting"	---	---	▣	---
SF "Benefits"	---	---	---	---
SF "Group Composition"	---	---	---	---
SF "Enforcement"	---	▣	■	□

--- SF not addressed
□ SF addressed, with contradicting results
▣ Results reproduce SF to some extent
■ Results according to SF

Fig. 16.1 Results of comparative model assessment by Heine et al. (2005)

nomenon. A model that contributes to explaining a stylized fact is therefore regarded as more valuable than another model that is geared toward side issues (Boland 1994, p. 536). If there are several stylized facts, the contribution of a model increases with its ability to explain more stylized facts alongside the absence of contradictions relating to the other stylized facts. Marcet and Nicolini (2003, pp. 245–290) provide a good example of this first step by demonstrating the value addition of their model with respect to a list of well-established stylized facts concerning hyperinflation. The authors show that existing models are not consistent with all of the stylized facts of hyperinflations on the one hand. On the other hand, their model adds value by being the first to explain all facts under study consistently.

In the area of social simulation, a study by Heine et al. (2005) provides another instructive example. In their study, the authors use a set of stylized facts as empirically grounded reference points to compare simulation models with game theoretic models (for a discussion of how stylized facts can be used as validation benchmarks see Chap. 18 by Saam in this volume). In particular, they use these stylized facts to assess the comparative advantages of these two modeling approaches. By referring to a set of stylized facts, they are not only able to provide an overview of the state of research in that field in an economic way, but also to specify the comparative advantages of simulation methods with respect to a specific phenomenon (see Fig. 16.1). The authors show that simulation models can be related to more stylized facts and achieve an equivalent or better reproduction of these facts. Moreover, the authors use these stylized facts to be very concrete about the potential of simulation models and argue that they exhibit far greater potential for incorporating yet unaddressed stylized facts, such for capturing the dynamic nature of the stylized fact changing group composition (Heine et al. 2005).

Thus far, it has only been discussed how simulation models can be validated with respect to their output using stylized facts as an empirically grounded reference point. If one does not wish to hold a simple instrumentalist position, a second step must be included.[13] Since there are different modes of reproducing stylized facts within a model, it should be analyzed in a second step how exactly the respective stylized facts are reproduced. In other words, one should "look under the hood" of the model results.

This expression stems from Hausman, who uses an insightful analogy of buying a used car to illustrate his argument as to why models should not be judged merely on the basis of their predictions concerning the phenomenon they attempt to explain. Hausman argues that when someone buys a used car, it is not sufficient to judge its future performance simply as a result of a road test. It makes sense to ask a mechanic to look at the engine and to judge how well the components serve their different purposes as well. Accordingly, much can be learned by carefully assessing model construction decisions and underlying assumptions, particularly when a model is extended to new circumstances or when it must be revised in the face of predictive failure (Hausman 1995, p. 219). This means that after successful reproduction of stylized facts, the necessity for testing shifts to the next level. In this level, the model mechanisms used to reproduce these facts must be assessed as well.[14]

In this respect, the stylized facts concept offers a clear guide as to how to "look under the hood". It helps to reduce complexity in pinpointing relevant components in the model and what they should do. With one or more stylized facts that a model can relate to, stylized facts can be used as a starting point to identify the basic assumptions and parameters in the model that are responsible for the reproduction of stylized facts. Hence, stylized facts can be used to isolate the corresponding model elements and thereby to enable subsequent model validation that is sufficiently focused to yield significant results. This is of particular interest when stylized facts are not reproduced adequately. Moreover, one can check whether any direct links from basic assumptions and parameters to stylized fact properties can be identified. This introduces a check for interpretability and plausibility of the basic model elements.[15] Such an interpretability check is particularly important when the modeling approach allows for certain degrees of freedom, e.g., by setting parameters and their values and/or when leaving a comparatively well-codified area, such as traditional game theory, to computer simulation models. Finally, this perspective can provide indications for future modeling options.

[13] A standard instrumentalist would contend that this step may be omitted. This position is often attributed to Friedman, who argues that "the only relevant test for the *validity* of a hypothesis is comparison of its predictions with experience" (Friedman 1989, pp. 8–9).

[14] See also Lawson (1989, pp. 62, 66). An explanation is not limited to the deduction of correct predictions, but also means isolating the "mechanism" that leads to the results. Models explain by generating insights, they depict how results are produced.

[15] It should be noted that the "realism" of assumptions is not assessed directly; rather, it is discussed whether appropriate and productive abstractions concerning the problem are being analyzed.

An illustrative example of this approach with respect to agent-based modeling is provided in Grimm et al. (2005).[16] They argue that combining several stylized facts can lead to more structurally realistic models. In particular, checking whether a set of stylized facts can be reproduced in a meaningful way allows discrimination between competing assumptions (e.g., theories about agent behavior) or even allows one to decide about parameter values. The authors illustrate this with respect to an ecological model of trout behavior.[17] Validating this model, one must decide between several competing theories about how individual fish select their habitats. Only one of the theories is able to reproduce all three observed patterns concerning feeding hierarchy, response to predatory fish and competing species, and response to reduced food availability (Grimm et al. 2005). This ability to select between competing assumptions at the level of agents is remarkable, as each of the patterns by itself is quite weak in terms of its discriminatory power and only their combination allows for elimination of the other theories.

In this second step, stylized facts can be interpreted as a spotlight for input validation. In other words, with reference to a set of stylized facts, it is possible to scrutinize the model mechanics systematically and without unnecessary complexity. Given the internal complexity of many simulation models, this use can be particularly beneficial in the area of simulation modeling. This analysis of the workings of a model can be fruitfully complemented by systematic sensitivity analyses (Lorscheid et al. 2012; Saltelli et al. 2004) and robustness tests used to identify the core mechanisms of a model, which are responsible for producing the observable model behavior (Grimm and Berger 2016).

Summing up this section, stylized facts can be used for different types of validation exercises. First, output validity of simulation models can be tested via using stylized facts as an empirically grounded reference point. Second, using stylized facts as a spotlight offers a clear guide as to how to "look under the hood". They pinpoint relevant components in the model, e.g., the modeling of agents and their interactions. This introduces a systematic approach to assess the input validity and the theory validity of simulations as well.[18]

[16] Grimm et al. (2005) call this model validation strategy "pattern-oriented modeling", but it follows a very similar basic logic.

[17] The same approach has been helpful in agent-based modeling applied to identify characteristics of stock market investors or to model the decisions of nomadic herdsman (Grimm et al. 2005).

[18] The use of stylized facts does not exclude using additional validation strategies. For example, Klingert and Meyer (2018) use stylized facts to validate the macro-output of their model and use data from laboratory experiments for an additional micro-validation. In this spirit and more generally, newer developments in validation techniques (Lamperti 2017; Richiardi et al. 2006) can be seen as an important complement to the use of stylized facts. Given the many possible degrees of freedom of simulation models, a broader set of validation techniques is a clear benefit. In comparison to other validation techniques, I see the main advantage of stylized facts, in their characteristic, that they are ideally based on a broad set of observations. This complements data sets typically used for validation which provide more details about a specific case or system.

16.3 Existing Approaches to Establish Stylized Facts

In order to realize the benefits that stylized facts can provide, it is of crucial importance to ensure the quality of the stylized facts in use. This problem is already recognized by Solow in his direct response to Kaldor, quoting his stylized facts with the comment, "[t]here is no doubt that they are stylized, though it is possible to question whether they are facts" (Solow 1988, p. 2). In addition to the critical nature of this statement, it also pinpoints a central aspect of a productive use of the concept: the explicit statement of a specific set of stylized facts by an individual researcher should not be the end, but rather the beginning of a critical discussion among experts in a particular field. This discussion makes explicit and open to criticism what is otherwise only used implicitly by individual researchers ("their view of the phenomenon").[19] Ideally, at the end of such a process, a consensus emerges, at least with regard to some stylized facts. Before such a consensus is reached, the transparency of derivation,[20] the amount and consistency of empirical results and the independence from specific theories or streams of literature may serve as supporting indicators of stylized fact quality.

Such a discussion must begin with existing empirical data. In order to reduce biases that might result from focusing on specific empirical methods, such as the use of only surveys, triangulation is an important concept. By combining different empirical research methods, the respective shortcomings of one individual method are compensated. More generally, the integration of different perspectives and research backgrounds also provides a means of coping with the problem of theory-ladeness (Popper 1959, p. 107).

Consequently, the derivation of stylized facts is ideally based transparently on a broad set of empirical investigations and ought to be subject to discussion among experts in the field. Ultimately, good stylized facts are the result of critical discussions within the expert community, with a high level of consensus that they represent a robust tendency in a substantial number of different empirical observations.

In the following section, three practices as to how stylized facts are currently generated are discussed from this perspective.

(1) "Ad Hoc" Approach

This "ad hoc" practice follows the early example of Kaldor, who generates stylized facts immediately for his specific rhetorical purposes and does not explain how he derives his stylized facts with regard to economic growth. Basically, two variants of such an ad hoc approach are distinguishable. With the first variant, the respective stylized facts are simply stated. Examples of such an approach are found in the papers already mentioned by Kahn (1987) or Marcet and Nicolini (2003). In such case, only the scientific expertise of the author(s) can provide an indication with respect

[19] Stylized facts are fallible. In other words, they can be revised in subsequent discussions and in light of new empirical findings. Critical discussion is central to Popper's understanding of science (Popper 1959).

[20] In the case of Kaldor's stylized facts, a lack of transparency is criticized particularly by Schwerin (2001). For a similar criticism concerning the current practice, see Boland (1997, pp. 243–245).

to the quality of the stylized facts, assuming that well-experienced researchers are familiar with relevant empirical literature. In the second variant, presentation of a set of stylized facts is supported with reference to certain supporting empirical studies (e.g., Geroski 1995). In the best case, these stylized facts are based on descriptive overviews of empirical literature that are typically provided by review papers. Taken together, these two variants are considered the most common current practice.

In assessment of the ad hoc approach, its main advantage is the comparatively low level of effort it requires to generate a set of stylized facts. As a result, the stylized facts can be used immediately for the purpose at hand.

However, as the response to Kaldor by Solow shows, this comes at the price of usually low credibility of the stylized facts stated. At least three reasons for this low credibility of ad hoc-stated stylized facts can be identified concerning the first variant. First, it is not transparent as to which empirical observations are used to support the stylized facts. Second, the logic of the process in which they are generated is not transparent as well. Finally, the process is not designed to include other experts, which limits not only credibility, but also awareness with respect to the generated stylized facts in a scientific community. As a result, it is very difficult for a set of stylized facts to be established as a focal point for a community of researchers. These three lines of criticism also apply to the second variant, with the qualification that the data used as input is made explicit. However, as long as there is no explanation concerning the selection of these studies, the data basis is possibly susceptible to strategic behaviors and biases, again limiting the acceptance of these stylized facts.

(2) **Survey Approach**

A second approach to generating a set of stylized facts is to ask experts in a certain field what they consider to be the empirically well-established characteristics of a phenomenon. Methodologically, this is an application of the survey method of collecting expert knowledge concerning relevant characteristics of a phenomenon. Illustration of such an approach is provided in the study of Whaples (1995), among economic historians. In this study, 178 experts in the field of economic history are asked whether they agree about 40 hypotheses that refer to North American economic history. With respect to eight hypotheses, there was a high level of agreement, up to 90% (e.g., in that slavery was inefficient from an economic point of view). Another example is a study by Frey, Pommerehne, Schneider, and Gilbert (1984) who surveyed 2072 economists from five countries regarding basic hypotheses in areas such as monetary or economic policy. This study shows considerably lower levels of agreement.

In assessing the survey approach, one of its clear advantages is that it offers an explicit procedure as to how the stylized facts are generated. Collected statements are easy to aggregate and the level of agreement offers a clear measure of consensus that can be reported. Moreover, the judgment of experts is generally considered a quite effective means to an assessment, particularly with respect to more complex questions. Finally, the survey approach creates, as a by-product, a higher level of awareness concerning the stylized facts in a field. Not only is expert attention drawn to the issue, but discussions are stimulated and results are reported.

One of the drawbacks of this approach, however—especially given the intended use of the generated stylized facts for model validation and assessment—is the possibility of strategic behaviors of the respondents. This susceptibility to opportunistic behaviors can undermine the credibility of a set of stylized facts. A second problem is that it can be quite difficult to derive agreement in some research areas. For example, while the study by Whaples (1995) shows quite high levels of agreement, in the other study by Frey et al. (1984) the highest level of agreement on a single hypothesis was only 67.8%. Finally, the derivation of stylized facts via survey is not based immediately on empirical data. Although one can expect that in the most cases the judgment of experts is based on data, it is not transparent as to which data forms the basis of such judgment. As a result, it is difficult to assess whether experts' judgments are based on a comprehensive set of empirical studies and to what extent there is a possibility of bias with respect to the studies used.

(3) **Statistical Approach**

A third practice of deriving stylized facts is based on statistical analysis of large data sets. This practice is quite prominent in extant finance literature in the form of sophisticated statistical analyses of large primary data sets from financial markets. This approach comes in two flavors. In the first variant, the derivation of stylized facts is based on analysis of one large data set. An example in this respect is the study by Lampenius (2008), which analyzes the Dow Jones Industrial Average. Lampenius shows that, for a time period of almost 80 years, certain general patterns in the data are observed, such as heavy-tailed distributions in the returns of assets. The second variant is also based on statistical analysis of large data sets but aims at generalizing the results of several such analyses of different asset markets. This generalization is done in a qualitative form, "by taking a common denominator among the properties observed in studies of different markets and instruments" (Cont 2001, p. 224). Examples of the derived stylized facts are properties of assets returns, such as gain/loss asymmetries in returns or volatility clustering (Cont 2001).

Evaluating this third practice, one of the main advantages of the statistical approach is its transparent and systematic way of data analysis and aggregation. Moreover, the derivation of stylized facts is immediately based on empirical observations. As a result, the respective data input can be scrutinized with respect to comprehensiveness and possible biases. Finally, the ability of this approach to process large amounts of data can be considered a further advantage. All these characteristics are expected to foster the credibility of the resulting stylized facts.

Still, these advantages come at the price of quite restrictive requirements with respect to data input. First, in many cases the availability of primary data is not as

Table 16.1 Comparison of the three different approaches to establish stylized facts

	Ad hoc approach	Survey approach	Statistical approach
Transparency concerning data input	Low (variant 1)/High (variant 2)	Medium	High
Transparency concerning aggregation process	Low	High	High
Consensus-fostering elements in the process	Low	Medium	Medium
Amount of empirical data	Cannot be assessed (variant 1)/Varies, but usually not high (variant 2)	Cannot be assessed	High
Comprehensiveness and restrictions concerning data sources	Cannot be assessed (variant 1) or Difficult to assess (variant 2)	Cannot be assessed	Suitable data necessary

straightforward as with respect to financial markets—researchers in this area are in a comparatively fortunate position. Second, generating a sufficient level of consensus and awareness with respect to derived stylized facts can be a problem, as in the process typically a community of experts is not integrated. This is of particular relevance with respect to intended generalization in the second variant and it seems desirable to amend the process in such a way as to integrate the judgment of several experts and to foster consensus. To summarize, given its restrictive requirements with respect to data input, the statistical approach should, at best, be considered a special case for a more general approach. In other words, the statistical approach can offer improvements in very specific settings, but cannot be considered the only method of deriving stylized facts.

To conclude, none of the approaches discussed thus far can be considered satisfactory with respect to all desirable dimensions. Nevertheless, some lessons can be learned from existing approaches to developing stylized facts. The survey approach has strengths in terms of its transparency concerning the aggregation process and its possibility of involving a large number of experts. Both should foster the acceptance and awareness of stylized facts. The statistical approach immediately links relevant empirical observations and offers the ability to process large amounts of data. Table 16.1 provides a summary of the main results of the discussion.

16.4 An Alternative Process to Derive Stylized Facts

In this section, a fourth possible way to derive stylized facts is presented, building on the respective strengths of both survey and statistical approaches. With respect to the design of such a process, one can draw on existing standards. As it involves the synthesis of existing empirical studies, this alternative process has a general nature that is similar to the compilation of reviews (e. g., Cooper 1987; Fink 2005; Petticrew and Roberts 2006). In this respect, Fink (2005), differentiates the descriptive survey (qualitative synthesis) from meta-analysis (statistic synthesis)[21] as different forms of the synthesis of scientific literature. The derivation of stylized facts and meta-analysis have in common that they seek to filter a broad tendency from a high number of individual empirical observations, in a systematic and reproducible way. For this reason, the basic approaches behind meta-analyses provide a good orientation.

However, meta-analyses are limited mainly to statistical evidence from quantitative studies.[22] As stylized facts should be based ideally on a broad range of methodological and theoretical approaches, empirical evidence for stylized facts should not be restricted to quantitative studies. To integrate qualitative results, the approach presented draws also on the methodology of systematic reviews (Petticrew and Roberts 2006) and the concept of structural regularities (Schwerin 2001). As a result, the high level of formalization of a meta-analysis, especially in regard to statistic synthesis, can mostly not be retained in the suggested derivation of stylized facts. Consideration of all available empirical findings requires modification of the approach, so that qualitative as well as quantitative results and a variety of different empirical methods may be considered.[23] Therefore, a processing of verbal statements at the center of the derivation of stylized facts must exist analogically to a synthesis in the form of a descriptive survey. Furthermore, a specification of the derivation of stylized facts exists in the fact that an adequate agreement beyond the sources must be given, whereas literature synthesis can also display a spectrum of different results. As a result, a traceable demonstration of the level of agreement is taken care of during the derivation of stylized facts, from which the identification of a stylized fact is justified.[24]

[21] Meta-analyses as a special form of synthesis is introduced by Glass et al. (1984) and are since methodically refined and applied.

[22] The emphasis of meta-analyses is on the integration of bi-variate results (e.g., Cooper and Lindsay 1998, p. 332). Utilization of statistic methods in a meta-analysis does not exclude qualitative papers, because with the method of "content analysis" and/or the computer-based evaluation of encoded, qualitative papers, the possibilities for the statistic evaluation exist, however, within tight limits (Babbie 2004, pp. 314–324; 375–392).

[23] As a special case, stylized facts can possibly be derived quantitatively in order to be approached analogical to meta-analyses. However, the general case—without such a limitation—is supposed here.

[24] The precise determination of a threshold from which the level of agreement is assumed as sufficiently large, presents a challenge. It depends significantly on the informative value of the underlying studies, from when a generalization of empirical results can be vindicated. General criterions can insofar not be given (Schwerin 2001, pp. 106–111).

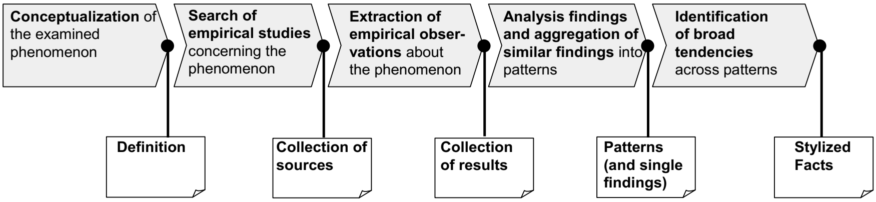

Conceptualization of the examined phenomenon
Search of empirical studies concerning the phenomenon
Extraction of empirical observations about the phenomenon
Analysis findings and aggregation of similar findings into patterns
Identification of broad tendencies across patterns
Definition
Collection of sources
Collection of results
Patterns (and single findings)
Stylized Facts

Fig. 16.2 Process of deriving stylized facts

Based on these thoughts, in Fig. 16.2 an outline of the process to derive stylized facts can be presented. In order to provide high transparency with regard to the process, an explicit intermediary result is presented as an output for each of the five steps. These elements in the process aim at fostering consensus, as this enables the retracing of the process in small steps and creates the basis for a constructive discussion. The following discussion of the different steps refers to excerpts of a previous study on deriving the stylized facts of the stability of collusion,[25] to illustrate the process steps, and in particular, the possible intermediary results.[26]

The starting point for the derivation of stylized facts consists analogical to the approach of meta-analyses, in a precise *conceptualization of the examined phenomenon*. This results in a precise definition of the phenomenon and the problem investigated. In this context, it is important to decide whether a phenomenon should be defined very generally or whether a certain context is already determined. If there is a high level of similarity of the phenomenon beyond different contexts, then this opens the possibility of widening the source material in order to achieve greater reliability of the stylized facts. The derivation of the stylized facts of collusion can provide a good illustration in this respect. Collusion can either be defined very generally as in the cooperation of agents for the increase of their benefit on the expense of a superior third *or* already in a specific context, such as the collusion of managers who are compensated after the Groves-mechanism. In this context, collusion can then be defined more specifically as the cooperation of managers for the increase of their compensation at the expense of the company. In this case, it seems reasonable to exclude the limitation to the specific context of the Groves-mechanism, since it can be initially assumed that in different contexts, the abstract mechanisms, which lead to the development of collusion, are sufficiently similar.

The second step, the *search of empirical studies* concerning the phenomenon is also in line with the approach of a meta-analysis. Here, the chosen definition serves as a criterion for the inclusion and exclusion of individual studies. The retrieval strategy should be transparent and should ideally comprise the relevant works as comprehensively as possible (Schwerin 2001, p. 106). The identified sources are then inspected for their respective representativeness for the empirical phenomenon and the compli-

[25] With this, the aforementioned criticism of Schwerin (2001, pp. 96–97, 99–100), on the missing transparency with respect to derivation of the stylized facts by Kaldor, are addressed.

[26] See Heine et al. (2005, 2007).

ance of other methodical standards (Fink 2005, p. 207). For the derivation of stylized facts of collusion, a limitation to scientific journals as a form of quality control is carried out.[27] The starting point is the literature database of the Social Science Citation Index (SSCI), a source often used for bibliometric analyses (similarly, Schwerin 2001, pp. 138–139).[28] Due to the focus on scientific journals, the compliance of methodical standards does not typically demonstrate a shortage and is capable of being evaluated on the basis of individual cases.[29]

In the third step, based on the source collection, the *empirical observations about the phenomenon are to be extracted*. As the derivation of stylized facts aims at the integration of qualitative results as well, an important difference with meta-analyses becomes apparent in this step. Basically, all empirical characteristics of the examined phenomenon, which are supported by a paper, are listed in verbal form. In the ideal case, these characteristics are observed in a large sample of papers. Since stylized facts are only formed where a broad agreement of the results exists, also individual empirical observations, which on their own are not representative (in the statistical sense), can be integrated, such as in case studies. A formalization of the synthesis, as with meta-analyses, is not reasonable due to the scope of qualitative findings, so that the traceable presentation of the intermediary result is given special importance. Figure 16.3 shows a possible example in tabular form in which the sources, identified in the previous step, are listed with the results to the collusion phenomenon extracted there. The table also includes a note about the research method employed in the respective studies in order to inform the reader in this respect as well.

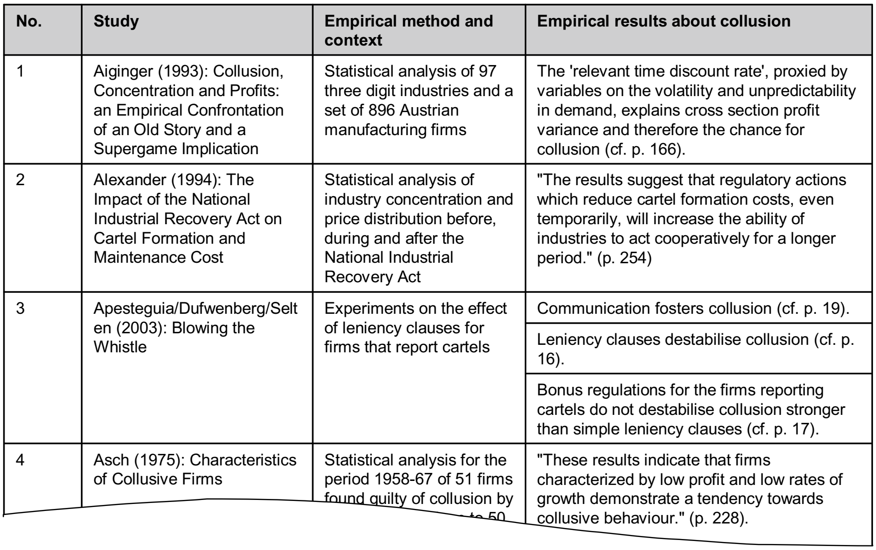
In the next step, one must process the resulting collection of empirical results in order that the general tendencies in the data extracted from different studies can be identified. In a meta-analysis, a statistical instrument completes this process. In the derivation of stylized facts, an *aggregation of similar findings to patterns* takes this place, whereby transparency and traceability are crucial. To foster this, first, the results are rephrased in order to abstract from the concrete study context in favor of implications for the observed phenomenon. The resulting "findings" are then equivalent to a hypothesis (in the sense of an "if-then-formulation") and nominate a variable ("if-part"), which shows an association with a significant criterion of the phenomenon ("then-part").[30] Based on this, the subsequent aggregation occurs in the form of formulating comprehensive patterns. These patterns of findings are based on

[27] The criteria that lead to publication in scientific journals are partly discussed in a controversial way (Schwerin 2001, pp. 122–125). The problem of a possible "publication bias" is eased in the context of the derivation of stylized facts since a positive (and thus, easier to publish) study result need not necessarily be consistent with the stylized fact, but can also confirm the opposite hypothesis (Fink 2005, p. 205).

[28] However, literature databases show distortions in their journal selections due to economic and practical influences. Therefore, the addition of further databases, as well as relevant but not covered journals, is carried out and transparently displayed. See the addition of the database EconLit by Stanley (2001, p. 135).

[29] The intermediary result is not displayed separately here, but is integrated in the next illustration.

[30] Using the structure of "if-then-formulations" does not necessarily imply a causal relation. Stylized facts are in a first step robust associations.



No.	Study	Empirical method and context	Empirical results about collusion
1	Aiginger (1993): Collusion, Concentration and Profits: an Empirical Confrontation of an Old Story and a Supergame Implication	Statistical analysis of 97 three digit industries and a set of 896 Austrian manufacturing firms	The 'relevant time discount rate', proxied by variables on the volatility and unpredictability in demand, explains cross section profit variance and therefore the chance for collusion (cf. p. 166).
2	Alexander (1994): The Impact of the National Industrial Recovery Act on Cartel Formation and Maintenance Cost	Statistical analysis of industry concentration and price distribution before, during and after the National Industrial Recovery Act	"The results suggest that regulatory actions which reduce cartel formation costs, even temporarily, will increase the ability of industries to act cooperatively for a longer period." (p. 254)
3	Apesteguia/Dufwenberg/Selten (2003): Blowing the Whistle	Experiments on the effect of leniency clauses for firms that report cartels	Communication fosters collusion (cf. p. 19).
			Leniency clauses destabilise collusion (cf. p. 16).
			Bonus regulations for the firms reporting cartels do not destabilise collusion stronger than simple leniency clauses (cf. p. 17).
4	Asch (1975): Characteristics of Collusive Firms	Statistical analysis for the period 1958-67 of 51 firms found guilty of collusion by ... to 50	"These results indicate that firms characterized by low profit and low rates of growth demonstrate a tendency towards collusive behaviour." (p. 228).

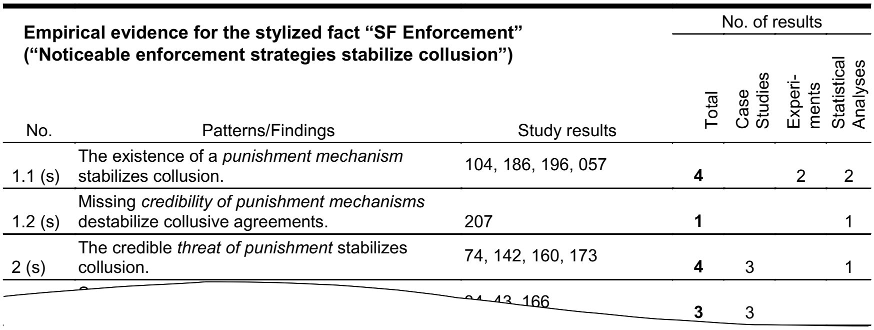
Fig. 16.3 Example for the extraction of empirical results

generalization of the "if-components" and/or the "then-components," under which larger numbers of the extracted results can be subsumed. In this process, insignificant details in comparison to the similarities can be neglected.[31] For the collusion example, rephrasing in regard to the mutual "then-component," "stability of collusion" details of the "if-component," such as context characteristics, which seem irrelevant, are neglected. Subsequently, patterns are formulated by the described generalization, such as stabilization of collusion with the existence of punishment mechanisms.[32]

In the final step, *broad tendencies over the different patterns must be identified*—the stylized facts. This encompasses a further generalization as to whether two or more patterns that are (only) different in detail can be further subsumed under a more general statement. In the tables, the findings are aggregated in groups with similar patterns and single findings. Then, it must be decided whether a sufficient level of agreement and representativeness of the underlying empirical results is given in order to assume a stable tendency, which justifies the identification of a stylized fact. To illustrate this final step, Fig. 16.4 provides an excerpt of the table that underlies the derivation of the stylized facts for enforcement strategies with respect to collusion. There, two of the patterns (and the individual empirical finding 1.2) are displayed, which can be subsumed under this general aspect.

[31] The question as to when an aspect is to be classified as an insignificant detail can only be decided in the respective context of this problem (Schwerin 2001).

[32] An illustration of the intermediary result is combined with the presentation of the result of the next step.



Empirical evidence for the stylized fact "SF Enforcement"
("Noticeable enforcement strategies stabilize collusion")

No.	Patterns/Findings	Study results	No. of results Total	Case Studies	Experi-ments	Statistical Analyses
1.1 (s)	The existence of a *punishment mechanism* stabilizes collusion.	104, 186, 196, 057	**4**		2	2
1.2 (s)	Missing *credibility of punishment mechanisms* destabilize collusive agreements.	207	**1**			1
2 (s)	The credible *threat of punishment* stabilizes collusion.	74, 142, 160, 173	**4**	3		1
		24, 43, 166	**3**	3		

Fig. 16.4 Example of aggregation to patterns and the identification of a stylized fact

These steps encompasses some challenges. The underlying aggregation and generalization comprise an inductive element.[33] Also, the determination of a threshold for the identification of a stylized fact is hard to determine precisely. However, indications, among other observations, can be obtained from the amount and conformity of the statements, the utilized empirical research methods and the sample sizes of the individual examinations. Here, particular situations stand out, in which a very evident, stable tendency exists, or in which the exact opposite is the case. Border cases, on the other hand, can hardly be decided in one or the other direction. However, this problem is to a degree inherent in science and thus it must be faced by the critical discourse in a community. Again, at this point, discourse is to be supported by a maximum of transparency and traceability, whereby also explicit notes to limits and/or restrictions (analogical to the meta-analysis) are made.[34]

Table 16.2 gives an illustration of the expected result of the described process. It depicts the derived stylized facts of collusion based on 109 empirical studies, which contributed 147 results concerning the stability of collusion.

Finally, assessing the approach presented in this section along the dimensions used in Table 16.1, there are some clear advantages from an epistemological perspective. The process of derivation is transparent with respect to data input and the aggregation process. Moreover, the approach can encompass a large amount of observations and a broad set of research methodologies and theoretical approaches. Finally, it comprises several consensus-fostering elements, which can be further complemented via application of the Delphi method or using the increasing possibilities of social software.[35] Still, this approach also requires certain accumulated empirical evidence

[33] Inductive conclusions are mostly problematic when they aim at deriving a universal statement (Popper 1959). However, this is not the case here, since stylized facts are generalizations without strictly universal claims.

[34] See also the discussion of the criterions in Schwerin (2001, pp. 140–142).

[35] To foster transparency, the derivation of stylized facts could be posted to a Website. There, discussion forums and voting mechanisms can be implemented to foster critical discussions among experts.

Table 16.2 Stylized facts of collusion (elaborated version of results in Heine et al. (2005, 2007)

No.	Short name SF	Stylized fact	$\sum$
1	Group size	A small effective group size stabilizes collusion	23
2	Setting	Uncomplicated settings stabilize collusion	44
3	Benefits	Actual or expected high benefits stabilize collusion	33
4	Time horizon	A long time horizon stabilizes collusion	10
5	Group composition	Little change in group composition stabilizes collusion	19
6	Enforcement	Noticeable enforcement strategies stabilize collusion	18
			147

as input. Moreover, one should not underestimate the effort necessary to derive a set of stylized facts.

16.5 Conclusion and Outlook

This chapter argues that in comparison to the widespread use of stylized facts, the amount of attention devoted to the concept and its adequate use is less. This applies both to the epistemological foundations of the concept and to the derivation of stylized facts. This, however, inhibits realization of the full potential of stylized facts, which can be particularly useful for validating simulation models. Against this backdrop, the paper clarified in the first step the epistemological basis of the concept and encouraged a more reflected use of stylized facts. To this end, Kaldor's original contribution is elaborated. This includes developing a precise definition: stylized facts are broad, but not necessarily universal generalizations of empirical observations and describe the supposed essential characteristics of a phenomenon that requires an explanation.

In addition, the chapter contributed to clarifying the interesting ideas behind the concept, especially with respect to model validation. Stylized facts can be used in the first place for a comparative assessment of simulation models with traditional modeling techniques. In our short case study, it was shown that in a debate on the Groves-mechanism simulation models have the potential to explain more stylized facts than game theoretic models. Moreover, they allow to scrutinize the structure of simulation models and their basic assumptions, e.g., at the level of agents. The use of stylized facts as a spotlight for "looking under the hood" goes beyond a simple instrumentalist approach of validating models and has, given the internal complexity of many simulation models, particular potential in this area. Hence, stylized facts have the potential to be used not only for output validation, but for input validation and theory validation as well.

In the second step, the need for a more systematic derivation of stylized facts is addressed. First, the ad hoc approach, the survey-based approach and the statistical approach to deriving stylized facts are presented and assessed. Based on the problems associated with these different approaches, a fourth approach is developed, which aims at transparent analysis and generalization of quantitative and qualitative empirical studies. Both steps—ideally in combination—are expected to strengthen the foundations for validating simulation models via stylized facts.

Given the potential of stylized facts for model validation, future research on this concept seems warranted. Among other things, future research could spell out in more detail how the use of stylized facts can be combined fruitfully with other validation approaches. Moreover, from an epistemological perspective a comparison with the approach of "saving the phenomena" in the sense of Bogen and Woodward (1988) seems to be worthwhile.

Acknowledgements This paper benefits from many discussions with Bernd-Oliver Heine, with whom I developed many ideas presented in this paper. In this respect see also Heine et al. (2007). I also thank Volker Grimm and Uta Berger for very helpful discussions about their related concept of pattern-oriented modeling. Finally, I would like to thank the two reviewers and the editors of this handbook for their constructive and helpful comments.

References

Aoki, M. (1988). *Information, incentives, and bargaining in the Japanese economy*. Cambridge: Cambridge University Press.

Babbie, E. (2004). *The practice of social research*. Belmont: Thomson Wadsworth.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review, 97*(3), 303–352.

Boland, L. A. (1994). Stylized facts. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *The new palgrave dictionary of economics* (Vol. 4, Repr. with corrections ed., pp. 535–536). London: Macmillan Press.

Boland, L. A. (1997). *Critical economic methodology*. London: Routledge.

Burton, R., & Obel, B. (1995). The validity of computational models in organization science: From model realism to purpose of the model. *Computational and Mathematical Organization Theory, 1*(1), 57–71.

Constance, E. H. (2007). Stylized facts, empirical research and theory development in management. *Strategic Organization, 5*(2), 185–192.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance, 1,* 223–236.

Cooper, H. M. (1987). Literature-searching strategies of integrative research reviewers. *Knowledge: Creation, Diffusion, Utilization, 8*(2), 373–383.

Cooper, H. M., & Lindsay, J. J. (1998). Research synthesis and meta-analysis. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 315–337). Thousand Oaks: Sage Publications.

Fink, A. (2005). *Conducting research literature reviews: From the internet to paper* (2nd ed.). Thousand Oaks: Sage Publications.

Frey, B. S., Pommerehne, W. W., Schneider, F., & Gilbert, G. (1984). Consensus and dissension among economists: An empirical inquiry. *American Economic Review, 74*(5), 986–994.

Friedman, M. (1989). The methodology of positive economics. In M. Friedman (Ed.), *Essays in positive economics* (15th ed., pp. 3–43). Chicago: University of Chicago Press.

Geroski, P. A. (1995). What do we know about entry? *International Journal of Industrial Organization, 13*(4), 421–440.

Gilbert, N. (2008). *Agent-based models*. Los Angeles, Calif. [u.a.]: Sage Publications.

Glass, G. V., MacGaw, B., & Smith, M. L. (1984). *Meta-analysis in social research*. Beverly Hills: Sage.

Grimm, V., & Berger, U. (2016). Robustness analysis: Deconstructing computational models for ecological theory and applications. *Ecological Modelling, 326,* 162–167.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., et al. (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science, 310,* 987–991.

Hands, D. W. (2001). *Reflection without rules: Economic methodology and contemporary science theory*. Cambridge: Cambridge University Press.

Hausman, D. M. (1995). Why look under the hood? In D. M. Hausman (Ed.), *The philosophy of economics: An anthology* (2nd ed., Reprint ed., pp. 217–222). Cambridge: Cambridge University Press.

Hayek, F. A. (1964). The theory of complex phenomena. In M. Bunge (Ed.), *The critical approach to science and philosophy* (pp. 332–349). Free Press of Glencoe.

Heine, B. -O., Meyer, M., & Strangfeld, O. (2005). Stylised facts and the contribution of simulation to the economic analysis of budgeting. *Journal of Artificial Societies and Social Simulation, 8*(4).

Heine, B.-O., Meyer, M., & Strangfeld, O. (2007). Das Konzept der stilisierten Fakten zur Messung und Bewertung wissenschaftlichen Fortschritts. *Die Betriebswirtschaft, 67*(5), 583–601.

Houy, C., Fettke, P., & Loos, P. (2015). Stylized facts as an instrument for literature review and cumulative information systems research. *CAIS, 37*, Article 10.

Kahn, J. A. (1987). Inventories and the volatility of production. *American Economic Review, 77*(4), 667–679.

Kaldor, N. (1968). Capital accumulation and economic growth. In F. A. Lutz & D. C. Hague (Eds.), *The theory of capital* (Reprint ed., pp. 177–222). London: Macmillan.

Klingert, F., & Meyer, M. (2018). Comparing prediction market mechanisms: An experiment-based and micro validated multi-agent simulation. *Journal of Artificial Societies and Social Simulation, 21*(1).

Lampenius, N. (2008). *Stylized facts of financial markets*. Manuscript.

Lamperti, F. (2017). An information theoretic criterion for empirical validation of simulation models. *Econometrics and Statistics, 5,* 83–106.

Law, A. M. (2006). *Simulation modeling and analysis* (4th ed.). Boston: McGraw-Hill.

Lawson, T. (1989). Abstraction, tendencies and stylised facts: A realist approach to economic analysis. *Cambridge Journal of Economics, 13*(1), 59–78.

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review, 73*(1), 31–43.

Lorscheid, I., Heine, B.-O., & Meyer, M. (2012). Opening the 'black box' of simulations: Increased transparency and effective communication through the systematic design of experiments. *Computational and Mathematical Organization Theory, 18*(1), 22–62. https://doi.org/10.1007/s10588-011-9097-3.

Lux, T., & Zwinkels, R. C. (2017). Empirical validation of agent-based models. *Handbook of Computational Economics* (Vol. 4, pp. 437–488). Elsevier.

Marcet, A., & Nicolini, J. P. (2003). Recurrent hyperinflations and learning. *American Economic Review, 93*(5), 1476–1498.

Meyer, M. (2011). Bibliometrics, stylized facts and the way ahead: How to build good social simulation models of science? *Journal of Artificial Societies and Social Simulation, 14*(4), 4.

Morgan, S. M. (1998). Models. In J. B. Davis, D. W. Hands, & U. Mäki (Eds.), *The handbook of economic methodology* (pp. 316–321). Cheltenham: Elgar.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford: Blackwell.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Richiardi, M., Leombruni, R., Saam, N., & Sonnessa, M. (2006). A common protocol for agent-based social simulation.

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity analysis in practice: A guide to assessing scientific models*. Wiley.

Schwerin, J. (2001). *Wachstumsdynamik in Transformationsökonomien: Strukturähnlichkeiten seit der Industriellen Revolution und ihre Bedeutung für Theorie und Politik* (Vol. 12). Köln: Böhlau.

Schwerin, J., & Werker, C. (2001). Learning innovation policy based on historical experience. *Structural Change and Economic Dynamics, 14*(4), 385–404.

Solow, R. M. (1988). *Growth theory: An exposition* (Paperback ed.). New York, NY: Oxford University Press.

Stanley, T. D. (2001). Wheat from chaff: Meta-analysis as quantitative literature review. *Journal of Economic Perspectives, 15*(3), 131–150.

Troitzsch, K. G. (2008). Stylized fact. In W. A. Darity Jr. (Ed.), *International encyclopedia of the social sciences* (2nd ed., Vol. 8, pp. 189–190). Detroit: Macmillan Reference.

Whaples, R. (1995). Where is there consensus among American economic historians? The results of a survey on forty propositions. *Journal of Economic History, 55*(1), 139–154.

Windrum, P., Fagiolo, G., & Moneta, A. (2007). Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation, 10*(2).

Chapter 17
The Users' Judgements—The Stakeholder Approach to Simulation Validation

Nicole J. Saam

Abstract This article presents a sociological perspective on the stakeholder approach to simulation validation using the validation of socio-ecological simulation models as an example. I develop an argument-based approach to simulation validation which can be applied in the natural and social sciences and argue that it is the constructionist camp of action researchers which has to consider the stakeholders' judgements as an indispensable point of reference for simulation validation. Only the stakeholders can validate that the model makes explicit their tacit knowledge. Only the stakeholders' willingness to accept and act upon the scenarios can decide issues of credibility. Obtaining the stakeholders' judgements in such a framework is an iterative communicative procedure that requires a strong background in qualitative methods of empirical social research as well as gaming simulation.

Keywords Simulation validation · Argument-based approach · Action research · Participatory modelling · Stakeholder

17.1 Introduction

Besides data (see Chap. 15 by Murray-Smith in this volume) and stylized facts (see Chap. 16 by Meyer) the user's judgements can serve as points of reference in simulation validation—meaning that these points of reference are compared to simulation results. However, while at first sight, every simulation model has a user, the user's judgement is not always relevant for simulation validation. This chapter introduces a particular context—action research using simulation models—in which the users' judgements are indispensable additional points of reference for simulation validation. The term 'user' takes on a special meaning in this context, a meaning that is better addressed with the concept of stakeholder. A stakeholder of a simulation model is anyone who has an interest in the model and its results. In the case of

N. J. Saam (✉)
Institute for Sociology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
e-mail: nicole.j.saam@fau.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_17

action research, such stakeholders are typically citizens and the simulation models are models of socio-ecological systems. Here, laypersons are included in the process of model development in some way or other. For instance, farmers participate in the process of model development to explore rules for rural credit (Barnaud et al. 2008), fishermen participate in developing rules for fishery management (Worrapimphonga et al. 2010), or herders and foresters participate in developing rules for landscape management (Dumrongrojwatthana and Trébuil 2011). These laypersons are referred to as participants and the modelling approach is called participatory modelling (Barreteau et al. 2013), companion modelling (Étienne 2014) or modelling with stakeholders (Voinov and Bousquet 2010). A model of a socio-ecological system typically includes agents which represent the stakeholders and their behaviour and tacit knowledge as well as an environment which represents the environment in which the stakeholders are living. The stakeholders also participate in the validation process. As Barreteau et al. (2013, p. 213) note, 'validation is the compulsory stage where stakeholders will have the opportunity to check the effectiveness of the computer model in representing correctly their behaviours and ways of acting'. The validation of these models is considered to be 'a difficult task because they mix different epistemological references' (Voinov and Bousquet 2010, p. 1277). Here, Voinov and Bousquet refer to the natural and the social components in the models. Although the stakeholder approach to simulation validation has been applied in many companion modelling studies, neither the methodology nor the epistemological and ontological foundations have been clarified. Depending on the nature of the research being conducted, action researchers can hold various ontological positions (Nicholas and Hathcoat 2014, p. 572). As Voinov and Bousquet (2010, p. 1272) point out, there is an ongoing debate between the positivist and the constructivist paradigm among action researchers using simulation models.

This chapter will present a sociological perspective on the stakeholder approach to simulation validation using the validation of socio-ecological simulation models as an example. I develop an argument-based approach to simulation validation which can be applied in the natural and social sciences. I argue that it is the constructionist camp of action researchers which has to consider the stakeholders' judgements as an indispensable point of reference for simulation validation. Only the stakeholders can validate that the model makes explicit their tacit knowledge. Only the stakeholders' willingness to accept and act on the scenarios can decide issues of credibility. Obtaining the stakeholders' judgements in such a framework is an iterative communicative procedure that requires a strong background in qualitative methods of empirical social research as well as gaming simulation.

The argument in this chapter is developed in seven steps. Initially, action research is described as a type of research in the social sciences based on philosophical pragmatism which supports social change and involves the participation of citizens. Here, simulation models may be used as intermediate objects.[1] The simulation model is used as a shared representation of a social system and its future. It helps us communicate about possible futures and find a solution to a problem (Sect. 17.2). In

[1] Meaning an object that mediates the communication between participants.

Sect. 17.3, I argue that a logical empiricist concept of simulation validation is not adequate for validating such simulation models. In Sect. 17.4, I propose a general concept of simulation validity that is suited for applications in the natural and the social sciences and an argument-based approach to validation. Both are applied to a hypothetical socio-ecological simulation model. In Sect. 17.5, techniques from qualitative social research and gaming simulation are presented which serve to obtain the stakeholders' judgements. In the discussion (Sect. 17.6), I reconsider the concept of validation and address the concept of judgement and the question whether the stakeholders' judgements could be suitable for validation purposes in other research fields too. Finally, I ask whether the stakeholders' judgements constitute some sort of face validity. The conclusion provides a very short summary and an outlook on important questions for further research.

In the following, the SCS definition of simulation validation will serve as a logical empiricist point of reference and starting point for this chapter's discussion: simulation validation is the 'substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model' (Schlesinger et al. 1979, p. 104). This definition directs the attention to the domain of applicability and to the model's intended application. It will be shown that the intended application of a simulation model in action research makes the stakeholders' judgements an indispensable additional point of reference for simulation validation.

17.2 Action Research and the Use of Simulation Models

Action research is a particular approach of social science research with a focus on topics and issues that will have implications for people's everyday lives (Reason and Bradbury 2008, p. 2). As opposed to other social science research where the researcher takes the position of an observer of social reality, in action research s/he is an active participant in social change, which gave rise to the very term 'action' research. The action researcher collaborates with members of a social setting on the diagnosis of a problem and the development of a solution based on that diagnosis. As Greenwood and Levin (1998, p. 62) state, action research 'aims to solve pertinent problems in a given context through a democratic inquiry where professional researchers collaborate with participants in the effect to seek and enact solutions to problems of major importance to the local people'. In action research processes, qualitative and quantitative methods of empirical social research are applied when the problem requires it. Popular areas of application of action research include social and community work, business and organization studies, nursing, healthcare, education, and development studies.

17.2.1 Meta-Theoretical Foundations of Action Research

Action research can be assigned to the participatory inquiry paradigm (Heron and Reason 1997; Lincoln et al. 2011). United by a common interest in participative and change-oriented initiatives, action research is used in different contexts and with different approaches (Reason and Bradbury 2008). In particular, there is a distinction between a pragmatic (e.g. Greenwood and Levin 1998, who in particular refer to Dewey 1900, 1902, 1991a[1927], b) and a critical orientation (e.g. Kemmis 2008) in action research. Johansson and Lindhult (2008) argue that both orientations suit different research contexts. The pragmatic orientation is well suited to contexts in which concerted action or practical knowledge is required (e.g. in the form of techne, or local knowledge) and in which power needs to be built in a situation of fragmentation and compartmentalization. In contrast, the critical orientation is preferable when reflection is crucial, when the development of a more abstract kind of knowledge is required (e.g. in the form of theory, episteme or reflective knowledge), and when the situation is characterized by unequal power relations or invisible structures that hamper thinking and action (Johansson and Lindhult 2008, p. 110). The focus of the latter is on emancipation of the participants, rather than the workability of the solution, which is favoured by the former. The goal of emancipation is connected to the core activity of reflection. Johansson and Lindhult (2008, p. 111) contrast this with experimentation as the core activity in pragmatically oriented action research. This distinction fits the empirical observation that action researchers with pragmatic orientation use computer simulation models as an intermediate object while those with a critical orientation do *not* use them. This article supports Johansson and Lindhult's (2008, p. 110) argument that the distinction of both orientations is rather an ideal-type construction. In the following, the critical orientation will not be neglected, but the pragmatic orientation will always serve as a starting point for the analysis.

Both orientations assume a constructionist subjectivist ontology and an interpretive epistemology pertaining to the social world. For the purpose of this chapter, I rely on Greenwood and Levin (1998) to explain these positions—which they summarize as hermeneutic and which they contrast with a logical positivist view. They claim that 'the world' is only available subjectively (and not objectively given) and that our epistemic task is to negotiate interpretations of this subjective world (and not 'to acquire the truth'; Greenwood and Levin 1998, p. 56). They stress that meanings are constructed and re-interpreted in collaborative communicative processes between researchers and participants in the action research inquiry process (Greenwood and Levin 1998, p. 63) and explicitly take a (social) constructivist position by referring to Berger and Luckmann (1966) as a point of departure. A recent reference (Coghlan and Brydon-Miller 2014), however, shows that action researchers follow not only a Berger and Luckmannian (social) constructivist ontology, but also other (social) constructivist traditions (see Shotter 2014, who prefers to call them social constructionist; in the following, I adopt his terminology), radical constructivism as put forward by von Glasersfeld (1982, see Hershberg 2014), critical constructivism grounded in the Frankfurt School's formulation of critical theory (see Steinberg 2014) or critical realism as developed by Bhaskar (1978, see Houston 2014).

In the following, I assume that action researchers using computer simulations take a social constructionist ontology as their basis. Thus, action researchers consider social reality to be continuously negotiated in interactions by social actors rather than as something external to them and totally constraining them. In particular, social realities are perceived as being local, specific and socially constructed. The local community whose problem is being addressed by the action research is perceived as experts on their own experience. Their local knowledge is explored through communication with the action researcher. Action researchers appreciate and respect diverse (social) realities and commitments of the participants. A reconstruction of social reality by the researcher is just another socially constructed reality with no epistemic priority or superiority. There is no such thing as a single (social) reality (Guba and Lincoln 2005).

17.2.2 The Validity of Action Research Knowledge

Referring to pragmatist philosophy, Greenwood and Levin (1998, p. 63) argue that the *'credibility-validity' of action research knowledge* depends on whether actions that arise from it solve problems (workability) and increase participants' control over their own situations. The framework of action research requires that the participants must be able to use the knowledge that emerges. The knowledge must support the enhancement of the participants' goals. Action research supports the creation of new knowledge that builds on a critical understanding of history and political contexts within which the participants act and that potentially can be liberating. Greenwood and Levin (1998, p. 64) describe the inquiry process as a communication process characterized by mutual relationships in which participants contribute their local knowledge, historical consciousness, and everyday experience, and researchers their skills in facilitating learning processes, technical skills in research procedures, and comparative and historical knowledge of the subject under investigation. The ideas of the superiority of scientific knowledge and the neutrality and objectivity of the researcher are rejected (Greenwood and Levin 1998, pp. 64–66); the distinction between researcher and participant serves only as a provisional statement. Finally, the action researcher inevitably becomes a participant as well (Greenwood and Levin 1998, pp. 64 f). Based on these ontological and epistemological assumptions, credibility is introduced as an alternative criterion for evaluating action research. Greenwood and Levin (1998, p. 67) claim that 'only knowledge generated and tested in practice is credible'. Altered patterns of social action are the ultimate test of credibility of the new knowledge. Greenwood and Levin (1998, p. 68) point out that it is not a community of similarly trained professionals that is deemed to be able to decide issues of credibility. Instead, it is the stakeholders' willingness to accept and act on the results.

Greenwood and Levin (1998, p. 68) consider two test procedures: (1) the workability test: the researcher has to establish whether the actions taken in the action research process result in a solution to the problem; (2) the test of credible knowledge making: the researcher has to ascertain whether the new knowledge is being accepted as a legitimate truth. Drawing on the constructionist framework, Greenwood and Levin (1998, p. 68) argue that meaning is constructed through deliberative processes and that there are two theoretical models that help the researcher in her test. In the communicative settings with participants, chains of arguments evolve and can be analysed. The credibility of the final argument can be evaluated by comparing the communicative procedures to a discourse and to Habermas' (1984) criteria of the ideal speech situation or to a dialogue and Gadamer's (2002[1960]) model of the fusion of horizons.

17.2.3 The Use of Simulation Models in Action Research and the Subject Matter of Their Validation

In some action research projects simulation models are used, in particular in the environmental sciences (natural resource management) and management sciences (group decision support systems). These simulation models do not necessarily involve a social component; sometimes the social simulation component may be rather implicit. Stakeholders may be involved in several stages of the modelling process or simply for means of data collection.[2] In the following, I use the case of socio-ecological simulation models as an exemplar which allows the problematization of important aspects that will be addressed in subsequent sections.

A socio-ecological simulation model is used as a shared representation of a socio-ecological system and its future (see the example in the next paragraph). It is used to facilitate communication of the stakeholders, whose (tacit) knowledge it makes explicit. Depending on the stage of development, the model will be considered a representation of the shared knowledge of the stakeholders. Initially, there may be no shared knowledge with respect to the problem. It involves a sample of the stakeholders who represent the heterogeneity of the (tacit) knowledge. The sample seeks to reproduce the diversity of possible viewpoints and behavioural patterns, bracketing the statistical representativeness. The simulation model is further developed in several cycles throughout the processes of social learning in the action research project. The stakeholders are supposed to be empowered to circumvent the worst-case scenario and create a shared vision for the solution of the problem in the socio-ecological system they belong to.

The Larzac companion modelling exercise (Simon and Étienne 2010) will illustrate the use of socio-ecological simulation models in action research. The Causse du Larzac is located in France southeast of the Massif Central. It is a karstic plateau

[2] Each action research project that uses simulation models can be subsumed under participatory modelling approaches. Please note that the opposite does not apply. Many projects of participatory modelling have a commercial, not a scientific background like action research.

with a long land use history of grazing and cereal cropping. Rural migration and mechanization are some of the deep socio-economic changes that have also changed the land use in the area. The objective of the Larzac companion modelling exercise was to develop alternative forest management plans by supporting forest owners and livestock farmers while they deliberated on solutions to their forest management problems. A bottom-up process was initiated. The project leadership decided to develop a computer simulation model rather than use role-playing games, based on an impression that forest owners and farmers preferred to discuss computer simulations. First, researchers and local professionals built an agent-based model which was validated via a participatory process. Second, this simulation model was used by researchers, forest owners and farmers to explore and discuss alternative forest policies and management strategies. The socio-ecological model included a representation of the territory and its land uses (pastures, silvopastures, forests, cultivated areas), as well as ecological dynamics such as tree encroachment dynamics (called 'ecological part' for short). In addition, the model included a significant number of agentic forest owners and farmers and their grazing and harvesting practices—representing almost all of the participants and their behaviour ('social part'). Note that model input for the social part was collected essentially by interviewing forest owners and farmers. Only the individual participants know their own grazing and harvesting practices, as well as changes in the land use pattern and their consequences in recent decades. Technically speaking, not only data for calibrating the social part of the model was collected; rather, all parameters and all variables, all behavioural equations pertaining to the forest owners and farmers were derived from these interviews. There are no prior behavioural equations of the social sphere (as opposed to dynamical equations in physics, or behavioural equations in some rational choice-based social simulations). This also means that prior to the interviews with the forest owners and farmers there is no conceptual model of the social part of the socio-ecological simulation model.

How may such a simulation be validated? Before addressing this question in more detail, the question will be further elaborated: what exactly is the subject matter of validation? In the following, I distinguish three subject matters: (1) V-S1: the simulation model that is used in the action research project; this task is addressed by Barreteau et al. (2013: 213) who state that 'validation is the compulsory stage where stakeholders will have the opportunity to check the effectiveness of the computer model in representing correctly their behaviours and ways of acting'; (2) V-S2: the results of the simulations; here, the task is to show that the results of the simulations are valid; and (3) V-S3: the results of the action research project; although this handbook focusses on the validation of computer simulation models, it has to be discussed whether the results of the simulations can be valid without the results of the action research project as a whole being valid. Irrespective of the answer to this question, the focus will be on subject matters V-S1 and V-S2.

In the following section, I argue that the constructionist ontology and the interpretivist epistemology, which are based on the intended use and domain of applicability of the simulation model, reveal that the SCS definition of simulation validation is not adequate for validating such socio-ecological simulation models.

17.3 The Logical Empiricist Versus the Post-positivist Understanding of Validity

Socio-ecological simulation models from action research are hybrids bridging the natural science–social science divide. As far as the representation of the ecological part of the simulation model (e.g. the representation of interactions between groundwater dynamics and surface water in a water management model) is concerned, the modeller may refer to a realist position. However, with respect to the groups of stakeholders (e.g. water users, landowners, local community representatives, water management agency) and their different views on social reality, the social part of the simulation model has to be based on a constructionist ontology. The validation of such a socio-ecological simulation is discussed in three steps: (i) validation of the simulation model, (ii) validation of the simulation results; and (iii) validation of the results of the action research project—leading from a logical empiricist understanding of validity to the requirement of a post-positivist concept to a general definition of simulation validity and its application.

17.3.1 The Logical Empiricist Understanding of Validity

Data and stylized facts will serve as points of reference for the validation of the ecological part of such a simulation model. Logical empiricism combined with a realist ontology justify the assumption that empirical data and stylized facts connected with an appropriate mathematical framework (see the chapters in Part V) are suitable for checking that the ecological part represents what it purports to represent. This epistemology postulates the empirical and logical values of accuracy, robustness, and consistency as epistemic standards (see Chap. 40 by Hirsch Hadorn and Baumberger in this volume). Validation is seen as a 'strictly formal, algorithmic, reductionist, and "confrontational" process' and validity 'becomes a matter of accuracy' (Barlas and Carpenter 1990, p. 157). A more moderate claim states that empirical accuracy is at the heart of the standard conception of simulation validation (see Chap. 40 by Hirsch Hadorn and Baumberger in this volume). By and large, 95% of the chapters of this handbook on simulation validation are dedicated to the question of how simulations can be validated from the point of view of a realist ontology and an epistemology that is based on logical empiricism—basically positions that dominate in the natural sciences and the quantitative social sciences (see also the overview by Feinstein and Cannon 2003). Following Scheurich's critique (1996, p. 49), this view and the related practices can be summarized as a positivist understanding of validity.

For the purpose of this chapter, I assume that there is no particular challenge in validating the ecological part of such a simulation model. For example, the accuracy with which the simulated groundwater dynamics represents the empirical groundwater dynamics can be quantified. In particular, this task nicely fits the SCS definition of simulation validation as substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the models's intended application (Schlesinger et al. 1979).

17.3.2 The Need for a Post-positivist Understanding of Validity

If the validity of the social part of such a simulation model is checked, the concept of accuracy proves problematic. We assume that an agent-based model has been implemented.

As far as the validation of the social simulation model (V-S1) is concerned, it has to be shown that the (tacit) knowledge and the perceptions of the social reality of the participants whose perspectives and views have been collected, for example by way of qualitative interviewing, are indeed represented correctly by the respective agents in the social part of the agent-based simulation model. To illustrate the problem, I return to the Larzac companion modelling exercise. Among other things, the researchers used French sentences that resembled a programming language organized as conditional clauses of the form 'if … then … else …' to check whether the rules integrated into the agent-based model represented the farmer's views correctly—thus, these sentences were part of the validation exercise (Simon and Étienne 2010, p. 1376). Often, the accuracy with which the properties of the agents represent the participants' statements cannot be quantified. Note that, for instance, participants often add further conditions after being confronted with their programmed statements. Note too that participants' statements include words with a cultural significance and meaning that cannot easily be represented in computer simulations. As there are limitations to the representation of contradictory views in a single simulation model even the notion of representation becomes vague. For instance, Simon and Étienne (2010, p. 1376) mention the discussions on pine encroachment in the validation phase. The final agreement was that pine encroachment would be the most difficult natural process to control as they had already experienced it on their farms. The researchers then discussed with the farmers how their respective agents should be modelled. Farmers considered them to be simple representations of their actual behaviour but relevant enough to be useful in discussing the problems at stake.

The concept of accuracy does not seem to be appropriate for validating the results of the simulations (V-S2) either. Assume that the simulation runs of such a socio-ecological model have produced five qualitatively different results. Let us call these qualitatively different results scenarios. Each scenario describes one possible future state of the socio-ecological system. My first argument here is that for a scenario to be considered valid it is not of major importance that it be accurate. First of all, to be valid a scenario has to be considered by the stakeholders. The emphasis is on the future being created by the stakeholders, *not on the model predicting the future to the stakeholders*. To be valid, a scenario must help us communicate about possible futures. Assume scenario O_1 makes the best prediction for harvesting fish of type F_1. If fish of type F_1 is rejected by stakeholders S_1 and S_2 and if their agreement is needed, scenario O_1 will not be considered—irrespective of how accurate the prediction may have been. My second argument refers to what is discussed as a self-destroying or self-fulfilling prophecy in the philosophy of the social sciences. If scenario O_2 predicts that all fish are harvested in a 5-year period under certain circumstances, then stakeholders will most probably consider other scenarios and scenario O_2 will not become true. In other words, there is a moving target. And the movers are the stakeholders. Only stakeholders' willingness to accept and act on a scenario makes the scenario valid. The validity of the model's results is better assessed with Greenwood and Levin's (1998) concept of credibility-validity (see Sect. 17.2.2 above).

Note that I do not argue that there are no quantities in the models that allow calculation of some sort of accuracy. I do argue that (a) accuracy is of minor importance—it is not appropriate *to decide* issues of validity—and (b) the point of reference to calculate accuracy is highly questionable as there is a moving target that depends on stakeholders' decisions. Note also that this is an empirical statement, not a prescriptive statement or value judgement.

Altogether, the concept of accuracy does not seem to be suitable for the intended application and in the domain of applicability of this simulation model. I suggest broadening the SCS definition of simulation validation and replacing the concept of accuracy because it is not adequate for simulation models which are based on a constructionist ontology and interpretive epistemology. In the methodology of the social sciences, qualitative researchers have developed substitutes for the positivist concept of validity—called post-positivist[3] concepts of validity (Scheurich 1996; recent trends in the discourse on validity in qualitative social research are discussed by Cho and Trent 2006; Koro-Ljungberg 2008). It is the goal of the general definition of simulation validity I present in the following section to be sufficiently general as to allow a logical empiricist and/or post-positivist specification.

17.4 A General Definition of Simulation Validity

Referring to the above discussion, I propose the following general definition of simulation validity that is suitable for the natural and the social sciences and constitutes an argument-based approach to validation based on Toulmin's account (1950, 1958). Toulmin proposes a five-part model for an argument: (1) claims: the position or claim being argued for; (2) grounds: reasons or supporting evidence that bolster the

[3] Please note that I follow Scheurich's (1996) broad definition of the term post-positivist, embracing all validity concepts that question the positivist concept. This deviates from Lincoln et al. (2011) more narrow use of the term. They distinguish five paradigms—positivist, post-positivist, constructivist, critical and participatory—and subsume under the post-positivist paradigm all those researchers who recognize and support validity, look for a qualitative equivalence for establishing validity and employ respective procedures.

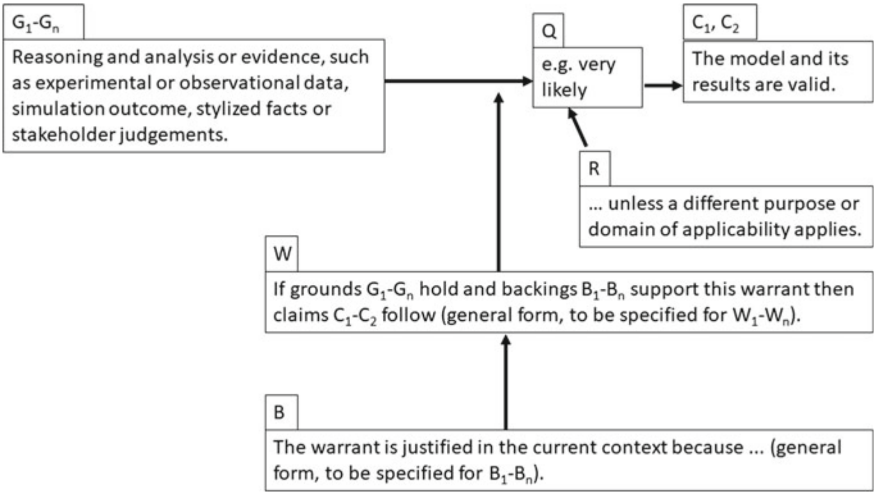

G1-Gn
Reasoning and analysis or evidence, such as experimental or observational data, simulation outcome, stylized facts or stakeholder judgements.
Q
e.g. very likely
C1, C2
The model and its results are valid.
R
... unless a different purpose or domain of applicability applies.
W
If grounds G1-Gn hold and backings B1-Bn support this warrant then claims C1-C2 follow (general form, to be specified for W1-Wn).
B
The warrant is justified in the current context because ... (general form, to be specified for B1-Bn).

Fig. 17.1 Full claims(C)-grounds(G)-warrant(W)-backing(B) structure of a scientific argument to support the claim that a simulation model and its result are valid, limited by a qualifier (Q) and a rebuttal (R)

claim; (3) warrant: the principle, provision, or chain of reasoning that connects the grounds/reason to the claim; (4) backing: support, justification, reasons to back up the warrant; and (5) rebuttal/reservation: exceptions to the claim; description and rebuttal of counter-examples and counterarguments. Additionally, Toulmin considers that an argument may be sound and yet not be completely convincing. In this case, we are interested in the strength of the argument. This is expressed by a qualifier.

If applied to the validation of a computer simulation model, the *claim* is the conclusion that the model and its results are valid (claims C_1 and C_2). The *grounds* include all kinds of reasoning and analysis or evidence, such as experimental or observational data, stylized facts, simulation outcome, and expert and stakeholder judgements that bolster both claims (see Fig. 17.1). Note that arguments can be chained to support the claim more convincingly.

17.4.1 Definition

Def. V-Mod: A simulation model is valid to the extent to which scientific argumentation supports the claim that the model represents what it purports to represent for the proposed purpose and domain of applicability.

Def. V-SR: A simulation model's results are valid to the extent to which scientific argumentation supports the intended use of the results for the proposed purpose and domain of applicability.

Table 17.1 Structure of the warrants connecting the grounds to the claims

Warrants
W_1: If grounds $G_i (i = 1, \ldots, n)$ hold and backings support this warrant, then claim C_1 follows
W_2: If grounds G_j $(j = 1, \ldots, k)$ hold and backings support this warrant, then claim C_2 follows

This argument-based approach to simulation validation encourages the validator to use multiple kinds of evidence to support the use of a computer simulation for a particular purpose, but it is not overly prescriptive. The approach refrains from setting strong, prescriptive rules. The task for the validator is to build an argument. The approach acknowledges that validation can never be established absolutely. It requires evidence that the simulation model (a) represents what it purports to represent, (b) is consistent with an appropriate ontological framework assumed for the intended domain of applicability, (c) meets the requirements of the proposed purpose (cf. Parker 2009), and that the simulation results (d) are reliable and supported by reasons and empirical evidence, and (e) their intended interpretations are reasonable and consistent with an appropriate ontological and epistemological framework assumed for the intended domain of applicability. Two basic warrants (W_1 and W_2; see Table 17.1) connecting the grounds to the claims C_1 and C_2 derive from this requirement. A sound argument integrates various strands of reasons and evidence into a coherent account of the extent to which (i) the model represents what it purports to represent or (ii) the results can be used for the proposed purpose and domain of applicability. The task for validators is to create a body of reasons and evidence sufficient to inform potential users of the strength and limitations of a particular simulation model and its results for particular purposes. At that level of generalization, it is not clear when sufficient evidence has been gathered. Nevertheless, even this general formulation allows the rejection of an intended use, e.g. if the assumed epistemological or ontological framework is not suitable for the domain of applicability. Simulation scientists have to elaborate and further specify criteria (a)–(e) for particular domains of applicability.

17.4.2 Application to Socio-Ecological Simulation Models in Action Research

The general definition of simulation validity requires that the validator develop a scientific argument that supports the intended use of her socio-ecological simulation model and its results for the proposed purpose and domain of applicability. The argument-based approach allows the natural science–social science divide to be bridged by using a logical empiricist framework for the ecological part of the simulation model and a post-positivist framework for the social part.

For reasons of space, the full claims-grounds-warrant-backing-rebuttal/reservation structure cannot be elaborated in this chapter. In the following, the focus is on the backings, i.e. support, justification, and reasons for backing up the warrants.[4] On the level of the backings, significant differences in the ontological and epistemological frameworks have to be made explicit.

17.4.2.1 Validation of the Ecological Part

Backings are elaborated to support warrants W_1-W_2 referring to a logical empiricist framework in which data, stylized facts and simulation output will serve as major grounds together with an appropriate mathematical framework. I do not go into details here and refer in particular to Chaps. 15, 16, 17 and 19, 20, 21, 22 of this volume.

To be able to present the full structure of backings in Sect. 17.4.2.3 (see Table 17.2), I assume—without any further explanation—the following backings B_{1E1} to B_{1E3} in relation to warrant W_1 and backing B_{2E} in relation to warrant W_2 where E denotes the ecological part.[5]

B_{1E1}: the ecological part represents what it purports to represent because it is based on a well-founded theoretical model.[6]

B_{1E2}: the ecological part is consistent with an appropriate ontological framework because the assumptions on the nature of ecological reality are plausible.

B_{1E3}: the ecological part meets the requirements of the proposed purpose because it helps communication about possible futures and finding a solution to the problem.

B_{2E}: the ecological results are reliable and supported by empirical evidence because they are congruent with relevant data and/or stylized facts.

Recently, Baumberger et al. (2017) have proposed an argument-based framework for the validation of climate science models. It appears they do not refer to Toulmin. However, their approach seems appropriate for providing further detail on some steps in the validation of the ecological part of socio-ecological models.

17.4.2.2 Validation of the Social Part

Backings are elaborated to support warrants W_1-W_2 referring to a post-positivist framework in which the stakeholders' judgements, simulation output and related reasoning serve as major grounds. The argument is outlined in some detail here. For reasons of space, I concentrate on two aspects: the choice of a particular post-positivist concept of validity and the justification of the stakeholders' judgements as appropriate grounds for evaluating the post-positivist validity claim.

[4] Note that we can think of many other backings which are not relevant for my argument here, e.g. backings related to the methodological frameworks applied in the validation.

[5] For reasons of space, a thorough explanation cannot be given in this chapter. I acknowledge that there are diverse ways to specify these backings.

[6] This is meant to be a necessary, not a sufficient condition.

Table 17.2 Structure of the backings

Backing
Ecological part (E) of the socio-ecological model
B_{1E1}: The ecological part represents what it purports to represent because it is based on a well-grounded theoretical model
B_{1E2}: The ecological part is consistent with an appropriate ontological framework because the assumptions about the nature of ecological reality are plausible
B_{1E3}: The ecological part meets the requirements of the proposed purpose because it assists communication about possible futures and finding a solution to the problem
B_{2E}: The ecological results are reliable and supported by reasons and empirical evidence because they are congruent with relevant data and/or stylized facts
Social part (S) of the socio-ecological model
B_{1S1}: The social part represents what it purports to represent because the stakeholders state that the relevant features of the implemented model are congruent with their views
B_{1S2}: The social part is consistent with an appropriate ontological framework because the assumptions on the nature of social reality are plausible
B_{1S3}: The social part meets the requirements of the proposed purpose because it assists communication about possible futures and finding a solution to the problem
B_{2S}: A simulated scenario is supported by reasons and evidence because the stakeholders consider this scenario
Overall (O) model
B_{1O}: Warrant W_1 is justified in the current context because B_{1E1}-B_{1E3} and B_{1S1}-B_{1S3} hold
B_{2O}: Warrant W_2 is justified in the current context because B_{2E} and B_{2S} hold

Backings Pertaining to the Simulation Model

Post-positivist Concept of Validity. Based on the constructionist ontology and interpretive epistemology (that relate to backing B_{1S2}, see Table 17.2; see also Sect. 17.2.1) as well as the purpose of solving pertinent problems in a given context through a democratic inquiry where professional researchers collaborate with participants in order to seek and enact solutions to problems of major importance to the local people (that relate to backing B_{1S3}, see Table 17.2), a concept of validity has to be specified that refers to warrant W_1. The simulation model is used to facilitate communication by the stakeholders whose (tacit) knowledge it makes explicit. It is reasonable to argue that the representation is considered valid because it makes explicit the stakeholders' tacit knowledge and because it facilitates the stakeholders' communication about the problem. Qualitative social researchers have been reluctant to apply the concept of validity, which they relate to a positivist understanding (for a short overview of that discourse see Lather 2007). A suitable post-positivist concept of validity has to be selected and justified. For the social part of the socio-ecological model, the concept of credibility as elaborated by Greenwood and Levin (1998; see Sect. 17.2.2 above) can be applied fruitfully.

The Stakeholders' Judgements. Based on the concept of credibility, appropriate standard practices have to be selected and justified. Lincoln and Guba (1985, p. 314) describe member checks as 'the most crucial technique for establishing credibility'.[7] A member check (also called respondent validation) is a communicative process whereby a researcher provides the people on whom he or she has conducted research with an account of his or her findings and requests feedback on that account. Examples of techniques include the semi-standardized interview using the structure laying technique and the focus group (see Sects. 17.5.1 and 17.5.2). With member checking, the validation procedure shifts to the stakeholders' judgements as points of reference. The backing B_{1S1} related to warrant W_1 then reads: 'The social part represents what it purports to represent because the stakeholders state that the relevant features of the implemented model are congruent with their views'. Thus, the act of validating, which is the task of the scientist, becomes an act of providing or denying legitimacy—a task for the stakeholder (Barreteau et al. 2013, p. 203). This relationship is also supported by action researchers Greenwood and Levin (1998, p. 68), who points out that it is not a community of similarly trained professionals that is deemed to be able to decide issues of credibility, but the stakeholders—an argument which relates to pragmatist philosophy.[8]

As pointed out above, the simulation model is further developed throughout the processes of social learning in the action research project. Validating the simulation model becomes a continuous task of quality assurance.

Backings Pertaining to the Simulation Results

Post-positivist Concept of Validity. This procedure has to be repeated for the warrant W2 pertaining to the simulation results. Again, an appropriate post-positivist concept of validity has to be selected and justified. Remember that rather than being an accurate representation of the possible future of a socio-ecological system, the simulation results are generated to assist communication about possible futures and finding a solution to a problem. This purpose is even supported by a simulated worst-case scenario with no empirical probability of realization. On the other hand, the factual future of the socio-ecological system that results from the action research project may not have been simulated at all. The simulated scenarios may just have served as initial scenarios for a joint discussion and decision by the participants.

The Stakeholders' Judgements. Typically, simulation experiments provide a large number of different scenarios as results—too many to be useful as a means of com-

[7] Other techniques include prolonged engagement, persistent observation, peer debriefing, negative case analysis, and progressive subjectivity (Guba and Lincoln 1989). Lincoln and Guba (1985) also recommend the triangulation of sources, methods and investigators, which they deemphasize in 1989 as 'too positivist' (Guba and Lincoln 1989, p. 240).

[8] Note that there is a significant difference to the logical empiricist concept of credibility, which defines it as the warranted degree of (professionals') belief in or confirmation of the simulation results (see Chap. 40 by Hirsch Hadorn and Baumberger in this volume).

munication among the participants. The intended application of the simulation results requires that the number of scenarios be small and be confined to those scenarios which are relevant to the stakeholders. Remember that in the action research framework, a (re)construction of social reality by the researcher is just another socially constructed reality with no epistemic priority or superiority. Applied to the choice of scenarios, this means that the stakeholders make the decision on the relevance of the scenarios that will be simulated. The empirical evidence for a scenario arises from the willingness of the stakeholders to consider this scenario—an argument which again relates to pragmatist philosophy. As Schubert (2015, p. 5) summarizes, quoting James (1907, p. 142): 'The predictions are "made true"'. This leads to backing B2S relating to warrant W2: 'A simulated scenario is supported by reasons and evidence because the users consider this scenario'. The emphasis is on the future being created by the stakeholders, not on the model predicting the future to the stakeholders.[9] Examples of techniques that can be used by the action researcher to obtain the stakeholders' judgements include the focus group or role-playing games and computer games.

17.4.2.3 Validation of the Overall Model

Ultimately, the backings have to be discussed for the overall model. At this stage of the validation process, the accumulated reasoning and evidence supporting the ecological and the social part has to be evaluated. The reasoning and evidence for the ecological part are based on a logical empiricist framework, while the evidence for the social part relates to a post-positivist framework. This leads to backings B_{1O} ('warrant W_1 is justified in the current context because B_{1E1}-B_{1E3} and B_{1S1}-B_{1S3} hold') and B_{2O} ('warrant W_2 is justified in the current context because B_{2E} and B_{2S} hold', see Table 17.2). Note that both backings rather hide conflicts from incommensurable ontologies and epistemologies. In the end, the constructionist ontology is superior to the realist ontology because the purpose of the socio-ecological simulation model is to mediate the communication between participants. The (from a realist perspective) best ecological part will not be useful if the stakeholders do not accept it as relevant to their discussions. The intended domain of applicability is the social sphere as constructed by the stakeholders and their tacit knowledge and its future as created by the stakeholders (*not* the social sphere as represented from an objectivist point of view and its future as predicted to the stakeholders). In practice, this means that a socio-ecological simulation model which is not accepted by the stakeholders has to be considered to be invalid, even if the ecological part may be considered to be valid from a realist perspective. The same holds for simulation results.

[9] Although the simulation model may calculate a specified probability for the realization of a certain scenario, this probability cannot be trusted. In the social world, there are always relevant actors and context conditions that are not represented in the simulation model. They have the power to dramatically change the probability of realizing the scenarios favoured by the participants of the action research project. See the study by Worrapimphonga et al. (2010), who report about the change of a governor, terminating any opportunity for implementing the favoured fishery management scenario.

An overview on the structure of backings is given in Table 17.2.

17.4.2.4 Overall Validation of the Action Research Project

This section somewhat transgresses this handbook's focus on the validation of simulations. However, the simulation model was used in an action research project, and finally, the latter has to be validated. Therefore, this subsection will conclude with a brief examination of the overall validation of the action research project.

The framework of action research requires that the participants must be able to use the knowledge that emerges. As Greenwood and Levin (1998, p. 68) point out, only the stakeholder's willingness to accept and act on the results arrived at collectively decide issues of credibility. The researcher has to ascertain whether the actions taken in the action research process result in a solution to the problem (workability test), and whether the new knowledge is accepted as legitimate truth (test of credible knowledge making; see Sect. 17.3.2).

While Greenwood and Levin (1998) refer to the concept of credibility, there is also an alternative. The concept of pragmatic validity as proposed by Kvale (1995) constitutes a terminological alternative. Essentially, Greenwood and Levin (1998) have presented a pragmatic concept of credibility which is consistent with Kvale's concept. In critically oriented action research on the other hand, Lather's (1986) concept of catalytic validity may apply. In an anti-foundationalist motivation, this concept relocates validity in achieving social justice, deeper understanding, and broader visions. Defined as the degree to which the research process reorients, focuses, and energizes participants towards knowing reality in order to transform it (Lather 1986, p. 67), catalytic validity considers unequal power relations or invisible structures that hamper thinking and action—a dimension which is not included in Greenwood and Levin's (1998) pragmatic concept of credibility.

It has to be discussed whether this overall concept of validity, applied to the whole action research project, feeds back into the concept of validity applied to the validation of the simulation model and its results. A thorough philosophical analysis cannot be provided in this chapter. However, one significant argument shall be raised that deserves further discussion. This argument relates to the requirement of representation and may rather be one of validation semantics. To be credible, the simulation model and its results have to fulfil this requirement to a larger extent than the results of the whole action research project. It is emphasized that the simulation model represents what it purports to represent, the socio-ecological system—at least to the degree to which it helps communication among the participants. Action researchers are not particularly concerned whether the inferences from the simulation model represent the solution to the problem at hand. They simply consider whether the solution works and the new knowledge is accepted as legitimate truth. Here, the aspect of (social) change is emphasized. The idea that the new knowledge and the change are represented by the inferences is neglected.

In the following, I consider techniques from qualitative social research which serve to obtain the stakeholders' judgements.

17.5 Validation Techniques Related to the Stakeholder's Judgements

Validation techniques to obtain the stakeholders' judgements refer to two tasks which derive from backings B_{1S1} and B_{2S}: the action researcher has to explore (1) whether the stakeholders consider the relevant features of the implemented model to be congruent with their views, and (2) which scenarios are considered by them. Validation techniques are applied repeatedly in a sequence of workshops with the stakeholders. There is no standard procedure for this. The step of validation is one (albeit repeated) task in a long process of participatory modelling which proceeds in cycles of developing and validating the simulation model.

The following techniques are suitable as member checks and can be combined. Finally, all of these techniques provide what Kvale (1995, pp. 30–32) has called 'communicative validity'—and would better be called 'communicative validation', meaning validation based on communication—which involves 'testing the validity of knowledge claims in a dialogue'. Basically, the validity of an interpretation (relating to the simulation model) is worked out in a dialogue between participants and action researchers. For analytical reasons, in the following, these techniques are presented separately.

17.5.1 Qualitative Interviewing

Qualitative interviewing constitutes the most important technique to explore, whether the stakeholders consider the relevant features of the implemented model to be congruent with their views. The participants are asked whether the model captures their assumptions about their own behaviour and their environment, whether some important dynamics related to the problem are missing or are not well represented compared to their assumptions about social reality. These questions address the participant's perceptions of (causal) relations (such as, 'If *y* happens, I do *x*' and '*A* is a precondition for *B*').

The semi-standardized interview (Flick 2014, pp. 217–223) is particularly suitable for validation purposes. In social research, this method is usually applied to reconstruct an interviewee's subjective theory about an issue under study. Applied to simulation validation, two phases can be distinguished. In a first meeting, the interviewee's view is obtained and recorded. Then, the interview is transcribed, roughly content analysed and transformed into model assumptions. In a second meeting, the structure laying technique is applied. The essential model assumptions—ranslated into everyday language—are presented to the interviewee as concepts on small cards for two purposes: (i) to assess the content: the interviewee is asked to recall the interview and check if its content is correctly represented on the cards. If this is not the case, she may reformulate, eliminate, and/or replace assumptions with other more appropriate ones; and (ii) to structure the remaining concepts from the interview that could not be transformed into model assumptions by the action researcher: the interviewee is asked to structure these concepts in a form similar to (causal) statements by applying the structure laying technique rules. The result of such a structuring process using the structure laying technique is a graphic representation of the stakeholder's statements. Finally, the action researcher compares the interviewee's view as represented by the cards to the assumptions implemented in the simulation model. Corresponding assumptions are perceived as being communicatively validated. Differing assumptions are changed according to the reconstructed view of the interviewee based on the graphic representation of the stakeholder's statements. Variations of the structure laying technique can also be used in games.

See Barnaud et al. (2008) for a sample study in which interviews were applied and Dray et al. (2006) for a study in which a structure laying technique was applied. Flick (2014, Chap. 16) provides a short introduction to qualitative interviewing and the structure laying technique. Gubrium and Holstein (2001) provide an extensive introduction to interviewing. Structure laying techniques are explained by Scheele (1992) and Scheele and Groeben (2010), and applied in diverse contexts by Scheele and Groeben (1986), Wagner (2003) and Weidemann (2009).

17.5.2 Focus Groups

In action research, focus groups are applied as a form of group discussion, rather than group interview, with an emphasis on a particular topic. On the one hand, the group is confronted with the researchers' interpretations of the individual statements in the interviews. Here, the group discussion becomes a tool for reconstructing the stakeholders' views more appropriately. Corrections by the group concern individual statements that are perceived as being not correct, not socially shared, or too extreme. On the other hand, the discussion may generate new knowledge on the group level. Finally, the group's view may converge to a community consensus on the questions of 'what is "real"', 'what is useful', and 'what has meaning', three criteria put forward by Lincoln et al. (2011, p. 116) for judging 'reality' or validity in qualitative inquiry in the social sciences. 'The meaning-making activities themselves are of central interest to social constructionists and constructivists simply because it is the meaning-making, sense-making, attributional activities that shape action (or inaction)' (Lincoln et al. 2011, p. 116). The criteria of the ideal speech situation (Habermas 1984) or the fusion of horizons model (Gadamer 2002 [1960]) serve as quality criteria for assessing the credibility of the group's final opinion.

See Barnaud et al. (2008) for a sample study in which focus groups were applied, Flick (2014, Chap. 17) for an introduction to group discussion methods including the focus group, and Barbour (2007) for an extensive discussion of this technique.

17.5.3 Role-Playing Games

Role-playing games, gaming simulation or even computer games constitute the most demanding technique with which to explore whether the stakeholders consider the relevant features of the implemented model to be congruent with their views. The technique provides—and requires—cross-validation of three models: the conceptual model, the simulation model, and the role-playing model (or gaming simulation model, or computer game model). In the following, I concentrate on role-playing games. On this level of abstraction, the procedure can be adapted to gaming simulation or computer games without significant changes.

The validation (and development) of the agent-based model proceeds in a cyclical way, combining the multi-agent system with a role-playing game. In each cycle, the role-playing game used with the stakeholders is a simplified version of the agent-based model. For example, the agents in the model correspond to the players in the game, the spatial interface of the model corresponds to the gaming board, and a time step of the model corresponds to a gaming round. The main difference is that while in the role-playing game decisions are made by human beings—the players/stakeholders/users –, in the agent-based model the corresponding decision-making processes are modelled. Thus, the game is a way to make transparent the model's assumptions for the stakeholders. The role-playing game allows the players to understand and therefore criticize and validate the model suggested by the action research team. In this cyclical process, the validation of the simulation model and the role-playing game mutually refer to each other (for the validation of games, see Klabbers 2009, Chap. 7) and to the target, the pertinent problem in a given context. Playing the game triggers lively and germane discussions among the stakeholders concerning their real situation. See Castella et al. (2005), Barnaud et al. (2008), and Naivinit et al. (2010) for sample studies. The cyclical procedure can be variated. Dumrongrojwatthana and Trébuil (2011) describe a model in which only the ultimate version of the (in this case computer-assisted) role-playing game validated by the players was converted into an agent-based model. Barreteau et al. (2001) describe the general approach behind the joint use of role-playing games and computer simulations. See Klabbers (2009) for an introduction to gaming simulation.

Additionally, role-playing games, gaming simulation and computer games constitute appropriate techniques to identify scenarios considered relevant by the stakeholders. In the gaming sessions, the stakeholders discuss diverse scenarios and may reach a collective agreement on those scenarios that should be further explored, this time by way of computer simulation. See Barnaud et al. (2008) for a sample study in which role-playing games were used in this way.

17.5.4 Inappropriate Techniques and Related Consequences

Techniques known from checking face validity, such as diagrams or figures visualizing the simulation output (mentioned by, e.g. Simon and Étienne 2010, p. 1376, and Barreteau et al. 2013, p. 208) have to be questioned as a suitable validation technique. Although used by some action researchers, they do *not* constitute member checks in the sense of qualitative social inquiry. They do *not* establish communicative validity in the sense of Kvale (1995). Instead, they reflect a logical empiricist approach. They are particularly prone to judgemental biases of stakeholders as outlined by Irvine et al. (1998).

Inappropriate techniques are not just inadequate for supporting the validity of the simulation model and its interpretation—they are detrimental to the whole action research project. See the study by Simon and Étienne (2010), who report how data collection at the request of one stakeholder distorted the cooperation between scientists and participants.

17.6 Discussion

This article claims that the stakeholders' judgements are indispensable points of reference in simulation validation in a comparatively small field,[10] namely constructionist action research using simulation models that include a social component. The discussion will focus on four questions: is the term 'validation' correct? Is the term 'judgements' correct? Could the stakeholders' judgements be suitable for validation purposes in other research fields as well? Do the stakeholder's judgements ultimately provide some sort of face validity?

Validation Versus Evaluation. You can evaluate simulations or their results on the basis of, e.g. whether they find consensus, whether they prove applicable, whether they are in fact applied, whether they help to solve problems, etc. But why call a related evaluation 'validation'? The answer of the action researcher with a pragmatic orientation is that the 'credibility-validity' of action research knowledge (Greenwood and Levin 1998, p. 63; see Sect. 17.2.2) depends on whether actions that arise from it solve problems (workability) and increase participant's control over their own situations. The framework of action research requires that the participants must be able to use the knowledge that emerges. Accordingly, simulation results obtained from action research using simulations are considered to be *invalid* knowledge if they are not used by the stakeholders. The predictions of the scenarios are made true only by being used. Based on Greenwood and Levin's pragmatist definition of credibility-validity, this is not a matter of evaluation but a matter of validation. This pragmatist perspective infers from the willingness of the stakeholders to use the

[10] Action research is applied in the social sciences, and has 'by and large […] not been a popular form of social research' (Bryman 2012, p. 393).

knowledge (a) its likely truth, (b) the warranted confidence in the simulation results, and (c) the correctness of the underlying model assumptions.

The Stakeholder's Judgements. What is the nature of the stakeholder's feedback obtained from communicative validation? In particular, is there a principled difference to data and stylized facts? Or does this feedback constitute some kind of data? In particular, is there a similarity to experimental data? I claim that there is a principled difference. In the cyclical processes of model development and validation, qualitative interviews, focus groups, and role-playing games provide two types of feedback: validation statements and corrections. The validation statements are unique for simulation validation. In all the data and stylized facts-related techniques, the validation statement comes from a scientist. In action research, the validity statement itself comes from the participant. The action researcher will critically reflect on this validity statement. However, ultimately the action researcher lacks an independent (objective) perspective. It is the judgements which constitute the unique point of reference. Only the corrections can be interpreted as some kind of data collected to improve the model in the next cycle of model development.

The Domain of Applicability. Could the stakeholders' judgements be suitable for validation purposes in other research fields too? Barlas and Carpenter (1990) have argued this. Based on a relativist position, they highlight the need for a dialogue between the modeller and other model stakeholders, and they claim that validation is 'a matter of social conversation rather than objective confrontation' (Barlas and Carpenter 1990, p. 163). Kleindorfer et al. (1998, p. 1098) put forward the claim that the model builders are free to establish and increase the credibility of the model through any reasonable means. Based on a hermeneutic position, they suggest involving 'other model stakeholders, such as model users and referees of journal articles' to increase the credibility of the model. They put forward the idea of a dialogue about a model's warrantability. There is not sufficient space here to discuss this idea extensively. In short, I wish to point out that the validity claims of Barlas and Kleindorfer's users' judgements are inferior to those of the participants in action research. In a social constructionist framework, the action researcher has lost her epistemic superiority vis-à-vis the stakeholder in accessing social reality. The stakeholder's epistemic priority is based on her exclusive access to her social reality, which justifies the trust in her judgements. Concerning Barlas and Kleindorfer's users, there is no such epistemic priority vis-à-vis the simulating scientist, which sets limits on their judgements' contribution to validating the simulation model. However, Schubert's (2015) statement that the predictions of the simulation model are *made* true holds for Barlas and Kleindorfer's users and their societal futures.

A less general claim would be that the stakeholders' judgements are suited for validating *all* social science simulations. Isn't there always an epistemic priority of the stakeholder vis-à-vis the simulating scientist, emerging from her exclusive access to her social reality? Here, my answer is that simulation models do not necessarily aim to represent people's everyday lives. There is a long, ongoing, debate on objectivity and subjectivity in social research. This distinction is reflected in the development of nomothetic and idiographic models in social simulation (see Ahrweiler and Gilbert 2009). The majority of social science simulations is developed from a rather objectivist point of view connected to a critical realist ontology. In these simulation studies, a logical empiricist framework is applied referring to data and stylized facts as points of reference for simulation validation (see Chap. 31 by Fagiolo et al. and Chap. 35 by Mäs in this volume). In this framework, the stakeholders' judgements are dispensable additional points of reference because they are perceived as being by largely inferior to data and stylized facts. From my point of view, the essential question addressing the validity of social science simulations is not whether these models can be validated by the stakeholders' judgements. Rather the question is when is a realist and when is a constructionist position adequate for modelling the social world. For example, Ahrweiler and Gilbert (2015) argue that in the case of a policy modelling exercise, only a constructivist position is appropriate and a user community view of model evaluation is recommended. It can be assumed that the ontological framework of several social science simulations is not well-founded.[11]

A last claim would be that the stakeholders' judgements are indispensable points of reference for *all* action research studies that use simulation models, i.e. including those studies whose model does *not* have a social component. Let us assume that a (purely) ecological model is developed. I question the suitability of such a model for action research. Such a design violates the requirement of democratic inquiry in which professional researchers collaborate with participants to seek and enact solutions to problems of major importance to the local people. A democratic inquiry simply cannot be established when for instance the interactions between groundwater dynamics and surface water have to be represented in a simulation model. The modelling exercise is asymmetric. There is no need for a democratic process. If such a model were developed and used in a (badly designed) action research project, we would expect the stakeholders to be unable to understand the model and its interpretation. It would not help them communicate about their problem nor would it help them enact solutions. Any helpless effort by the action researcher to obtain the stakeholder's judgements of such a model will fail because the stakeholders are simply over-challenged by this, and this fact is known by the researcher and by the stakeholders themselves. Such a design will lead to action research projects which fail to achieve their aims.

Communicative validity based on group consensus. In computer simulation, the concept of face validity has been reformulated as referring to individuals knowledgeable about the target who are asked whether the model and/or its behaviour are reasonable (Sargent 2013, p. 16). As Murray-Smith (2015, p. 95) notes this technique is especially helpful in early stages of the development of simulation models when there is no prototype system available for testing. In action research, there is typically no such prototype system. The stakeholders' judgements provide expert statements in the sense of the face validation approach. It has been objected that

[11] See, for instance, the critique put forward by Barreteau et al. (2013, p. 210) on the use of Bayesian belief networks: 'The translation of participant-provided information into probabilities is mediated by the modeller and is rather opaque, as in many participatory modelling approaches'.

'different stakeholders might well have different views and understandings of their own behaviour, the behaviour of others and the ways in which stakeholders interact with one another. These are not matters of some objective truth' (Moss 2008, Sect. 5.11). This objection ignores the fact that action research considers and models multiple (social) realities and rejects the idea of an objective truth. So does communicative validation. Moss's objection also ignores the numerous built-in credibility checks: the validation procedure involving qualitative interviews, focus groups, and role-playing games requires each stakeholder to first validate the relevant part of the simulation model from his or her perspective. Next a common view of social reality is established, supported by the simulation model. When debating and deliberating in the focus groups, the stakeholders may criticize each other and develop new and shared views together. Finally, the validity of the simulation model and its results is based on group consensus which overcomes initially different views of the stakeholders. The sequence of judgements includes numerous model falsifications (and subsequent model improvements by the action researcher) establishing a high degree of credibility on the part of the stakeholders which is also approved by the action researcher. This is face validity as defined by Sargent (2013). However, the cyclical validation procedure involving qualitative interviews, focus groups, and role-playing games is a much more sophisticated procedure than the concept of face validity suggests. And, while the intuition of the concept of face validity directs the attention to the visual sense, the key aspect is instead the human capacity for communicative action. Therefore Kvale's (1995) concept of communicative validity captures the essence of this validation procedure much better than the concept of face validity.

17.7 Conclusions

This article claims that the stakeholders' judgements are indispensable points of reference in simulation validation in constructionist action research, if the model includes a social component. Obtaining the stakeholder's judgements in an action research framework is an iterative communicative procedure which requires a strong background in qualitative methods of empirical social research as well as gaming simulation and establishes communicative validity. Beyond this rather narrow field, the stakeholders' judgements may be obtained as additional points of reference if a hermeneutic approach to simulation validation is preferred.

17.8 Outlook

What are important questions for further research? (1) Validation techniques. The presented techniques are well-known in the social sciences, but not for the purpose of validation however. Procedures and principles for applying these techniques in simulation validation have to be specified and problems in their application addressed.

The practitioners already apply such procedures and principles. However, we lack a related methodological discourse. Quality criteria related to these techniques ultimately help the action researcher evaluate the validity of the stakeholders' judgements. (2) The proposed argument-based approach to validation has to be elaborated. Basic assumptions and the philosophical foundation have to be further specified and juxtaposed with alternative accounts of simulation validation. (3) Different concepts of credibility from logical empiricist (see Chap. 40 by Hirsch Hadorn and Baumberger in this volume) and post-positivist (Greenwood and Levin 1998; Guba and Lincoln 2005; Lincoln and Guba 1985) perspectives need to be clarified.

Acknowledgements The author thanks Claus Beisbart for helpful discussions concerning this manuscript.

References

Ahrweiler, P., & Gilbert, N. (2009). The epistemologies of social simulation research. In F. Squazzoni (Ed.), *Epistemological aspects of computer simulation in the social sciences* (pp. 12–28). Heidelberg: Springer.

Ahrweiler, P., & Gilbert, N. (2015). The quality of social simulation—an example from research policy modelling. In M. Janssen, M. Wimmer & A. Deljoo (Eds.), *Policy practice and digital science—integrating complex systems, social simulation and public administration in policy research* (pp. 35–55). Springer, Heidelberg.

Barbour, R. (2007). *Doing focus groups*. London: Sage.

Barlas, Y., & Carpenter, S. (1990). Philosophical roots of model validation two paradigms. *System Dynamics Review, 6,* 148–166.

Barnaud, C., Bousquet, F., & Trebuil, G. (2008). Multi-agent simulations to explore rules for rural credit in a highland farming community of Northern Thailand. *Ecological Economics, 66,* 615–627.

Barreteau, O., Bousquet, F., & Attonaty, J. M. (2001). Role-playing games for opening the black box of multi-agent systems. Method and lessons of its application to Senegal river valley irrigated systems. *Journal of Artificial Societies and Social Simulation, 4*. http://jasss.soc.surrey.ac.uk/4/2/5.html.

Barreteau, O., et al. (2013). Participatory approaches. In B. Edmonds & R. Meyers (Eds.), *Simulating social complexity. understanding complex systems* (pp. 197–234). Heidelberg: Springer.

Baumberger, C., Knutti,R., & Hirsch Hadorn, G. (2017). Building confodence in climate model projections. An analysis of inferences from fit. *WIREs Climate Change*, *8*, e454.

Bhaskar, R. (1978). *A realist theory of science*. Brighton, England: Harvester Press.

Berger, P., & Luckmann, T. (1966). *The social construction of reality*. New York: Doubleday.

Bryman, A. (2012). *Social research methods* (4th ed.). Oxford: Oxford University Press.

Castella, J.-C., Trung, T. N., & Boissau,S. (2005). Participatory simulation of land-use changes in the Northern Mountains of Vietnam. The combined use of an agent-based model, a role-playing game, and a geographic information system. *Ecology and Society, 10*, 27.

Cho, J., & Trent, A. (2006). Validity in qualitative research revisited. *Qualitative Research, 6,* 319–340.

Coghlan, D., & Brydon-Miller, M. (2014). *The SAGE encyclopedia of action research* (Vol. 1 and 2). London: Sage.

Dewey, J. (1900). *The school and society*. Chicago: Chicago University Press.

Dewey, J. (1902). *The child and the curriculum*. Chicago: Chicago University Press.

Dewey, J. (1991a). *Logic: The theory of inquiry*. Carbondale: Southern Illinios University Press.

Dewey, J. (1991[1927]). *The public and its problems*. Athens: Ohio University Press.

Dray, A., Perez, P., Jones, N., Le Page, C., D'Aquino, P., & Auatabu, T. (2006). The atollgame experience: from knowledge engineering to a computer-assisted role playing game. *Journal of Artificial Societies and Social Simulation, 9*. http://jasss.soc.surrey.ac.uk/9/1/6.html.

Dumrongrojwatthana, P., & Trébuil, G. (2011). Northern Thailand case: Gaming and simulation for co-learning and collective action. Companion modelling for collaborative landscape management between herders and foresters. In A. van Paassen, J. van den Berg, E. Steingröver, R. Werkman & B. Pedroli (Eds.), *Knowledge in action. The search for collaborative research for sustainable landscape development* (pp. 191–219). Wageningen Academic Publishers.

Étienne, M. (Ed.). (2014). *Companion modelling. A participatory approach to support sustainable development*. Dordrecht: Springer.

Feinstein, A. H., & Cannon, H. M. (2003). A hermeneutical approach to external validation of simulation models. *Simulation & Gaming, 34,* 186–197.

Flick, U. (2014). *An introduction to qualitative research* (5th ed.). London/Thousand Oaks, CA/ Dehli: Sage.

Gadamer, H.-G. (2002 [1960]). Truth and method. Joel Weinsheimer. (2 revised ed.). New York: Continuum.

Glasersfeld, E. V. (1982). An interpretation of piaget's constructivism. *Revue Internationale de Philosophie, 36,* 612–635.

Greenwood, D. J., & Levin, M. (1998). *Introduction to action research: Social research for social change*. Thousand Oaks: Sage.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park: Sage.

Guba, E. G., & Lincoln, Y. S. (2005). Paradigmatic controversies, contradictions, and emerging confluences. In N. Denzin & Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research* (pp. 191–215). Thousand Oaks: Sage.

Gubrium, J. F., & Holstein, J. A. (Eds.). (2001). *Handbook of interview research*. London: Sage.

Habermas, J. (1984). Theory of communicative action. In T. McCarthy (Ed.), *Reason and the rationalization of society* (Vol. 1). Boston: Beacon Press.

Heron, J., & Reason, P. (1997). A participatory inquiry paradigma. *Qualitative Inquiry, 3,* 274–294.

Hershberg, R. M. (2014). Constructivism. In D. Coghlan & M. Brydon-Miller (Eds.), *The SAGE encyclopedia of action research* (Vol. 1 and 2, pp. 182–186). London: Sage.

Houston, S. (2014). Critical realism. In D. Coghlan & M. Brydon-Miller (Eds.), *The SAGE encyclopedia of action research* (Vol. 1 and 2, pp. 219–222). London: Sage.

Irvine, S. R., Levary, R. R., & McCoy, M. S. (1998). The impact of judgemental biases on the validation of simulation models. *Simulation & Gaming, 29,* 152–164.

James, W. (1907). Pragmatism's conception of truth. *The Journal of Philosophy, Psychology and Scientific Methods, 4,* 141–155.

Johansson, A., & Lindhult, E. (2008). Emancipation or workability? Critical versus pragmatic scientific orientation in action research. *Action Research, 6,* 95–115.

Kemmis, S. (2008). Critical theory and participatory action research. In P. Reason & H. Bradbury (Eds.), *Handbook of action research. Participative inquiry and practice* (2nd ed., pp. 121–138). London: Sage.

Klabbers, J. H. G. (2009). *The magic circle. Principles of gaming & simulation* (3rd ed.). Rotterdam: Sense Publishers.

Kleindorfer, G. B., O'Neill, L., & Ganeshan, R. (1998). Validation in simulation: Various positions in the philosophy of science. *Management Science, 44,* 1087–1099.

Koro-Ljungberg, M. (2008). Validity and validation in the making in the context of qualitative research. *Qualitative Health Research, 7,* 983–989.

Kvale, S. (1995). The social construction of validity. *Qualitative Inquiry, 1,* 19–40.

Lather, P. (1986). Issues of validity in openly ideological research. *Interchange, 17,* 63–84.

Lather, P. (2007). Validity, qualitative. In G. Ritzer (Ed.), *The blackwell encyclopedia of sociology* (pp. 5161–5165). Oxford: Blackwell Publishing.

Lincoln, Y. S., & Guba, E. (1985). *Naturalistic inquiry*. Beverly Hills: Sage.

Lincoln, Y. S., Lynham, S. A., & Guba, E. G. (2011). Paradigmatic controversies, contradictions, and emerging confluences, revisited. In N. Denzin & Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research revisited* (pp. 97–128). Thousand Oaks: Sage.

Moss, S. (2008). Alternative approaches to the empirical validation of agent-based models. *Journal of Artificial Societies and Social Simulation*, *11*, http://jasss.soc.surrey.ac.uk/11/1/5.html.

Murray-Smith, D. J. (2015). *Testing and validation of computer simulation models: Principles, methods and applications*. Cham: Springer.

Naivinit, W., Le Page, C., Trébuil, G., & Gajaseni, N. (2010). Participatory agent-based modeling and simulation of rice production and labor migrations in Northeast Thailand. *Environmental Modelling and Software, 25,* 1345–1358.

Nicholas, M. C., & Hathcoat, J. D. (2014). Ontology. In: D. Coghlan & M. Brydon-Miller (Eds.), *The SAGE encyclopedia of action research* (Vol. 1 and 2, pp. 570–572). London: Sage.

Parker, W. S. (2009). Confirmation and adequacy-for-purpose in climate modeling. *Proceedings of the Aristotelian Society, Supplementary, 83,* 233–249.

Reason, P., & Bradbury, H. (Eds.). (2008). *Handbook of action research. Participative inquiry and practice* (2nd ed.). London: Sage.

Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation, 7,* 12–24.

Scheele, B. (Ed.). (1992). *Struktur-Lege-Verfahren als Dialog-Konsens-Methodik*. Münster: Aschendorff.

Scheele, B., & Groeben, N. (1986). Methodological aspects of illustrating the cognitive-reflexive function of aesthetic communication. *Poetics, 15,* 527–554.

Scheele, B., & Groeben, N. (2010). Dialog-Konsens-Methoden. In G. Mey & K. Mruck (Eds.), *Handbuch qualitative Forschung in der Psychologie* (pp. 506–523). Wiesbaden: VS.

Scheurich, J. J. (1996). The masks of validity: A deconstructive investigation. *International Journal of Qualitative Studies in Education, 9,* 49–60.

Schlesinger, S., Crosbie, R. E., et al. (1979). Terminology for model credibility. *Simulation, 32,* 103–104.

Schubert, C. (2015). Situating technological and societal futures. Pragmatist engagements with computer simulations and social dynamics. *Technology in Society, 40,* 4–13.

Shotter, J. (2014). Social constructionism. In D. Coghlan & M. Brydon-Miller (Eds.), *The SAGE encyclopedia of action research* (Vol. 1 and 2, pp. 704–707). London: Sage.

Simon, C., & Étienne, M. (2010). A companion modelling approach applied to forest management planning. *Environmental Modelling and Software, 25,* 1371–1384.

Steinberg, S. R. (2014). Critical constructivism. In D. Coghlan & M. Brydon-Miller (Eds.), *The SAGE encyclopedia of action research* (Vol. 1 and 2, pp. 203–206). London: Sage.

Toulmin, S. (1950). *An examination of reason in the place of ethics*. Cambridge: Cambridge University Press.

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Voinov, A., & Bousquet, F. (2010). Modelling with stakeholders. *Environmental Modelling and Software, 25,* 1268–1281.

Wagner, Rudolph F. (2003). Clinical case formulation in the research program "subjective theories". Patients with chronic diseases. *European Journal of Psychological Assessment, 19,* 185–194.

Weidemann, D. (2009). A cultural psychological approach to analyze intercultural learning. Potential and limits of the structure formation technique. *Forum Qualitative Social Research*, *10*, 43. http://www.qualitative-research.net/index.php/fqs/article/view/1246.

Worrapimphonga, K., Gajaseni, N., Le Page, C., & Bousquet, F. (2010). A companion modeling approach applied to fishery management. *Environmental Modelling and Software, 25,* 1334–1344.

Chapter 18
Validation Benchmarks and Related Metrics

Nicole J. Saam

Abstract This chapter proposes benchmarking as an important, versatile and promising method in the process of validating simulation models with an empirical target. This excludes simulation models which only explore consequences of theoretical assumptions. A conceptual framework and descriptive theory of benchmarking in simulation validation is developed. Sources of benchmarks are outstanding experimental or observational data, stylized facts or other characteristics of the target. They are outstanding because they are more effective, more reliable or more efficient than other such data, stylized facts or characteristics. Benchmarks are set in a benchmarking process which offers a pathway to support the establishment of norms and standards in simulation validation. Benchmarks are indispensable in maintaining large simulation systems, e.g. for automatic quality checking of large-scale forecasts and when forecasting system upgrades are made.

Keywords Validation benchmarks · Touchstone · Yardstick · Engineering reference standard · Benchmarking · Benchmarking metrics

18.1 Introduction

According to its most simple definition, a benchmark is a point of reference or standard against which things may be compared (Oxford Dictionaries). Benchmarking is a comparative method for performance evaluation of systems using benchmarks, and it is known from management science and computer science. In the former, the quality of an organization's policies and products is measured, compared and evaluated using standard measurements, or similar measurements of its peers, e.g. industry bests. In the latter, benchmark programmes are developed and run on computers to compare and evaluate the performance of different processors and computer architectures. Benchmarking is also known from the stock market, where the performance

N. J. Saam (✉)
Institute for Sociology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
e-mail: nicole.j.saam@fau.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_18

of investment funds is measured against the performance of a point of reference, the stock market itself.

In computer simulation, benchmarking is predominantly used for engineered systems (which are not the focus of this volume). In recent years, however, some empirical sciences have adopted benchmarking with respect to modelling and computer simulations (e.g. hydrology and biogeosciences, see references in Sect. 18.2.2). In particular, hydrological models are evaluated using benchmarks. Nevertheless, validation benchmarks are comparatively rare in simulation validation. In this chapter, I argue that the potential of benchmarks for validating simulations is yet to be sufficiently recognized, and propose broading the use of benchmarks in validating simulations. Validity I understand as the degree to which a scientific argument supports the intended interpretation of the simulation model and its results for the proposed purpose and domain of applicability (see Chap. 17, Sect. 17.4 by Saam in this volume). I thus put forward an argument-based approach to simulation validation. The benchmarking exercise provides results which support the establishment of such a scientific argument. In particular, it serves to quantify the fit between simulation outcome and that part of reality that the simulation model is meant to enable us to understand or explain. It provides a statement on how close the simulation outcome is to the benchmark. However, the results of the benchmarking process do not indicate whether the fit is good enough for the intended application, and we do not know whether the results are good for the right reason(s). Benchmarking can therefore only be a first step on the long road that is validating a simulation model and its results.

This chapter proposes benchmarking as an important, versatile and promising method in the overall process of validation of simulation models with an empirical target. This excludes simulation models which only explore consequences of theoretical assumptions. A conceptual framework and descriptive theory of benchmarking in simulation validation is developed. Sources of benchmarks are outstanding experimental or observational data, stylized facts or other characteristics of the target. They are outstanding because they are more effective, more reliable and more efficient than other such data, stylized facts or characteristics. Benchmarks are set in a benchmarking process which offers a pathway to support the establishment of norms and standards in simulation validation. They are indispensable for maintaining large simulation systems, e.g. for automatic quality checking of large-scale forecasts and when forecasting system upgrades are made.

In the following, I define the basic concepts of this theory (benchmark variables, benchmarks proper, benchmarking; Sects. 18.2 and 18.3), develop two typologies (types of benchmarking and types of validation benchmarks, Sects. 18.3.1 and 18.4), explain why validation benchmarks are used (Sect. 18.2.2), which criteria apply for selecting a benchmark (Sect. 18.3.2) and why which kind of metrics are used (Sect. 18.5), and finally discuss strengths and weaknesses of the method (Sect. 18.6). Questions for future research are outlined in the conclusion.

18.2 The Concept of Validation Benchmarks

18.2.1 Defining Validation Benchmarks

Avoiding a definition that is too narrow, I define a *validation benchmark* in computer simulation as point of reference or standard against which the simulation results may be assessed according to validity characteristics, such as accuracy or credibility. Validation benchmarks have to be distinguished from verification benchmarks. The latter refers to the process of determining that a model implemented on the computer accurately represents the developer's conceptual description of the model as well as the solution to the model (see Oberkampf and Trucano 2008), while the first assesses the degree to which a model represents the target.

While Schlesinger et al. (1979) focus in their definition of model validation ('*the substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model*'; Schlesinger et al. (1979, p. 104)) on only one validity characteristic—accuracy—my definition is more open and recognizes that simulation results may be evaluated according to credibility or other criteria. This means that the concept of validation benchmark can also be used to assess the credibility of a computer simulation model and its results. In the natural sciences, when introducing the idea of benchmarks for validating simulations it may be wise to concentrate on the presently predominant characteristic for the validity of simulation results, i.e. accuracy. In the social sciences, error often cannot be quantified, setting limits to the knowledge gained from calculating accuracy statistics.[1] In this case, the model's output is better assessed according to other or additional criteria. In the long run, particularly if the use of simulations should spread into many fields in technology and society, it may be necessary to consider alternative or additional characteristics to accuracy. This is, at least, what we can learn from the use of computer benchmarks, a mature field where performance benchmarking was dominating and *dependability* benchmarking has since gained recognition (Vieira and Madeira 2009). For the moment, I propose concentrating on the development of *accuracy benchmarks* for simulation validation (please note, however, the more differentiated view that results from developing the typology of validation benchmarks in Sect. 18.4).

To improve terminological clarity, I introduce a distinction that has not yet to be made in the literature—I distinguish the benchmark variable from the benchmark *proper*—both of which are presently referred to in the literature as benchmarks. The benchmark variable denotes that variable in a computerized model (and antecedent conceptual model) for which the decision is made that a benchmark *proper* will be sought and defined. The benchmark *proper* denotes one or more points of reference or determinate values of the variable on a scale in the benchmark *variable's* range

[1] In some cases, the measure of agreement between the simulation result and the benchmark can have a large tolerance. In other cases, the benchmark itself is less quantitative (or not quantitative) and, therefore, a measure of agreement cannot be accurately defined.

of values (or sometimes a common characteristic of several benchmark *variables*) against which the simulation results will be assessed (which was my initial definition of validation benchmarks at the beginning of this section). As will be elaborated in Sect. 18.4, there are different classes of benchmarks *proper*. Note that in the interests of readability, I will explicitly refer to benchmark *variables* and benchmarks *proper* wherever necessary, but adhere to the concept of validation benchmarks if both aspects are addressed.

From a merely technical perspective, benchmarks may qualify as (a) one variable, several variables, or functional relationships of two or more variables, (b) specified in time and/or space, (c) measured on all available scales (nominal, ordinal, metric) (d) whose value(s) may have the mathematical form of a scalar, a time series or a matrix and (e) the statistical form of a distribution, probability distribution etc., or any combination thereof.

In the empirical sciences, validation benchmarks are useful for different purposes:

(i) To support simulating scientists in the conception of the validity of their model's results. The validation results feedback into the further development of models, especially when models are applied regularly, such as in forecasts of simulating surface water flooding events.

(ii) Help public agencies and non-governmental organizations to choose the model that best fits their requirements by comparing the accuracy and other features of alternative models. This requires that there are competing models on issues of public interest, a precondition that is often not fulfilled (an exception is climate modelling).

The latter purpose being rather an exception, validation benchmarks will predominantly serve science and the evaluation and further development of models.

18.2.2 Motivations for Using Validation Benchmarks

There are various reasons for using validation benchmarks. Some important motivations will be illustrated, without being exhaustive. The intention is to sensitize the reader to the versatility of the method. For some phenomena that are modelled, no observational data is available to validate the simulation results. For instance, microsimulation models that forecast population growth are faced with a no data situation in which official population projections serve as an external benchmark (see e.g. the study by Harding et al. 2010).

For some other phenomenon, there may be observational data to validate the simulation results. However, scientists do not merely want to compare observations and simulation results. Instead, they seek to assess model performance as good or poor (Seibert 2001). In intercomparison projects, validation benchmarks are used to set minimum levels of acceptability for model performance (see e.g. the study by Nicolle et al. 2014). Skill analysis provides an answer to the question as to how close a forecast was to the observations compared to how close a benchmark was. For example, in hydrological ensemble prediction, climatological or meteorological data may serve as benchmarks (see the benchmark intercomparison study by Pappenberger et al. 2015).

Some other simulation models are too complex to expect validation of each simulated process. Luo et al. (2012, p. 3858) for instance argue that a land model typically simulates hundreds or thousands of biophysical, biogeochemical and ecological processes on regional and global scales over hundreds of years. Even if observations were available, it is unrealistic to expect that so many processes on all spatial and temporal scales are validated independently. One way to evaluate such models has been a holistic assessment, as performed in intercomparison projects. A second way is benchmarking: a systematic evaluation against data from a range of carefully selected observations and experiments. A benchmarking process can be organized as a project (see e.g. the International Land Model Benchmarking project, Hoffman et al. 2017, and https://www.ilamb.org/), which is recommended if one intends to develop of a community-based model evaluation system that is open source and modular, allowing contributions by many different modelling and measurement teams.

In terms of practical use, a quick warning is sometimes needed. For example, in urban flood risk management, the computational cost of highest-resolution data is too expensive, rendering a timely warning difficult. Here, a compromise between computational cost and accuracy is needed and provided by lower resolution benchmark data (see e.g. the study by Fewtrell et al. 2011).

The use of computer simulation for practical applications such as in medicine or nuclear reactor safety necessitates a justified confidence that the models are adequate representations of the target. There is a need to establish standards for the validation of these models that typically evolve in an iterative process. Validation benchmarks provide users with benchmark examples and benchmark experiments that can be run with the previous and new versions of a simulation package (see e.g. the study by Lund et al. 2012). Such validation benchmarks can also be understood as test problems (Oberkampf and Trucano 2008, p. 718). These test problems may not only allow quantification of the accuracy of the computational model by comparing its results with experimentally measured data. They may also enable the interpolation or extrapolation of the computational model to conditions corresponding to the model's intended use or determination if the estimated accuracy of the computational model—for the conditions of its intended use—satisfies the accuracy requirements specified (Oberkampf and Trucano 2008, p. 724).

This diversity of motivations for using validation benchmarks raises the question of whether there are different types of validation benchmarks—a question that will be addressed in Sect. 18.4.

18.2.3 Sources of Benchmarks

Four sources of benchmarks (benchmark *variables* and benchmarks *proper* in this chapter's terms) are distinguished by Luo et al. (2012, p. 3862): direct observations; experimental results; observation proxies ('data-model products'); and functional relationships or patterns. These categories seem to overlap, e.g. experiments can involve direct observations. Patterns are patterns in observational data.

Validation against observational data has been the most common approach to model evaluation in science (Oreskes 2003). Although observational and experimental data are generally accepted to be the most reliable benchmarks for model performance, there are also reasons to question them as benchmarks. It is recognized that even the most direct measurements need some level of processing, up-scaling and assumptions to yield the final estimates (Luo et al. 2012, p. 3862). Pappenberger et al. (2015) discuss the issue of observations or proxies as a basis for benchmarks. They consider a river discharge proxy, a reference river discharge that is not observed but simulated by a hydrological model using observed meteorological inputs. They argue that observation-based benchmarks are easier to construct than proxy-based ones and might seem to be more suitable than the latter. They also argue that observational data can contain errors, e.g. observed discharge data captured during floods usually contains errors from when measurement devices stop functioning or are destroyed or bypassed by flood waters. A second reason for proxy-based benchmarks is the lack of long historic data series (Pappenberger et al. 2015, p. 708). Another suggestion is made by Luo et al. (2012, p. 3862), who discuss interpolation and extrapolation of data according to some functional relationships in order to extend the data's spatial and temporal scales. They argue that related errors may be well-quantified. Nevertheless, extrapolation functions may introduce artefacts, especially outside the observation ranges.

In the social sciences, the use of survey data raises the same question of suitability as a source for benchmarks. The error in survey data often cannot be quantified; for instance, Harding et al. (2010, p. 57) report 'noise' in survey estimates due to small sample size in some groups and the distorting effect of survey data being collected in a period of economic prosperity. In this case, proxy-based benchmarks such as official population projections or cross-sectional data from national bureaus of statistics are more suited. However, caution is required when comparing a simulation's output with that of external projections, depending heavily on the assumptions made. A model output missing the projection does not necessarily reflect a bad model (Harding et al. 2010, p. 55).

Oberkampf and Trucano (2008, p. 726) emphasize the role of validation experiments to obtain experimental data suited for simulation validation. Validation experiments constitutes a new type of experiment conducted for the primary purpose of determining the predicative accuracy of a computational model. Such an experiment should be jointly designed by experimentalists, model developers, code developers and code users. Validation experiments should not only be distinguished from scientific discovery experiments, model calibration experiments and experiments serving as system performance tests. They should also be designed, executed and analysed separately.

Functional relationships or patterns in the data can be used as benchmarks to evaluate the model's results, in particular when uncertainties in data due to both random and systematic error are unknown. For example, correlations between El Nino-related climate anomalies and growth rate of atmospheric CO_2 can be used to examine the

Table 18.1 Types of benchmarking

Object compared Number of objects	Simulation validation process	Simulation outcome
One	–	Performance benchmarking
Several	Process benchmarking	Best-in-class benchmarking

consistency between the observed and the simulated ecosystem responses to climate change (Luo et al. 2012, p. 3862f.).

Benchmark levels of performance that a certain model can be targeted to achieve have also been defined based on standard simulation results of a well-accepted model, the model ensemble mean, and statistically based model results (see the references in Luo et al. 2012, p. 3863).

18.3 The Benchmarking Process

Validation benchmarks are selected and used in a process called benchmarking. Benchmarking is defined as a practice in and through which the improvement of a simulation model's performance is sought through comparison with a reference point.

I distinguish *benchmarking in the narrow sense*, which requires a social process and agreement on benchmark variables, and benchmarks proper from *benchmarking in the broader sense*, which relaxes this assumption. In the latter case, it is only required that the scientist connects the choice of a certain benchmark variable and related benchmark proper to the claim that both are submitted to the relevant scientific community in order to achieve an agreement on the use of these benchmarks in all future models in this domain of applicability and with this intended application of the model. The social dimension of benchmarking distinguishes this validation practice from alternative practices in which simulation results are compared to any sort of points of reference (see Sect. 18.6.2 for further discussion).

18.3.1 Types of Benchmarking

Several types of benchmarking can be distinguished depending on the object compared (simulation outcome versus validation process) and the number of simulation models (see Table 18.1). All of the types may be used by different disciplines or industries. Note that these types are not on the same level:

Performance benchmarking requires that the outcome characteristics of a single simulation model are compared to its benchmarks (level 1).

Best-in-class benchmarking[2] requires that the outcome characteristics of several models on the same phenomenon are compared to each other (level 2).

Process benchmarking requires that the whole validation process, and not just the outcome characteristics, of several models are compared to each other (level 3).

Levels 1 and 2 can be classified as benchmarking *in* validation and level 3 as benchmarking *of* validation.

18.3.1.1 Performance Benchmarking

Luo et al. (2012, p. 3857) note that we are in an infant stage of benchmarking analysis. In those disciplines in which we perceive an enduring effort in developing simulation models on a particular phenomenon, benchmarking analysis may evolve into a standard technique for validating new versions of already existing simulation models.

Six major steps are proposed as a framework for benchmark analysis for performance benchmarking: (1) *Develop benchmarking validation concept*. This step includes identification of key aspects of the simulation model that require evaluation. (2) *Select benchmark variables*. Potential (candidate) benchmark variables—a set of benchmark variables—are identified. Note that this may be a long and arduous process, in particular for simulation models with many output variables. Several tasks are hidden behind this step: (a) identification of benchmark variables; (b) choice of the scale on which each benchmark variable will be measured; and (c) choice of the error (or resolution) with which each benchmark variable will be measured (see Sect. 18.3.2 for more details). At the same time, (d) a decision on the quality of the point of reference has to be made that will be associated with each candidate benchmark variable (see Sect. 18.4 for more details on different types of validation benchmarks and their related qualities). Only after these decisions have been made are data collected or validation experiments run to obtain data on candidate benchmark variables. Pros and cons of different candidate benchmark variables are discussed, and finally, a decision is made on which ones to use. (3) *Select metrics*. Appropriate validation or benchmarking metrics (see below) are selected on the basis of the characteristics of the benchmarking data and simulation outcome, e.g. their scale and error characteristics. (4) *Define evaluation criteria and procedure*. Evaluation criteria (such as a level that is still acceptable or a test that has to be passed) and related procedures are chosen according to the accuracy requirement of the intended application. (5) *Evaluate simulation model*. Simulation results are compared to benchmarking data and assessed using the metrics and the evaluation criteria. (6) *Improve simulation model*. Model deficiencies are identified and rectified (as an example, see the studies by Luo et al. 2012, and Pappenberger et al. 2015).

[2] Please note that this definition deviates from the use of best-in-class benchmarking in business benchmarking, where it is also known as process benchmarking.

In a more narrow interpretation, benchmark analysis would only include the comparison exercise, whereas the step of improving the simulation model would be considered a response, but not a part of the analysis.

18.3.1.2 Best-in-Class Benchmarking

Best-in-class benchmarking compares to rival models and their performance. The strengths and weaknesses of the peer models are identified in order to determine each model's strengths and weaknesses. This helps to prioritize specific areas for improvement. Intercomparison projects (see also Chap. 34 by Knutti et al. in this volume) are examples of best-in-class benchmarking (Sundberg (2011, p. 120).

Six major elements are proposed as a framework for best-in-class benchmarking: (1) *Develop benchmarking concept*. This step includes identification of the model domain that requires evaluation. (2) *Select benchmarking partners*. A number of simulation models in the domain are identified and modellers contacted and asked for their cooperation in the intercomparison project. (3) *Develop common framework*. A common evaluation framework, including a common set of criteria, a common set of benchmark variables and their related benchmarks proper, appropriate metrics, and a testing scheme are negotiated by the partners. (4) *Implement common framework*. Each partner implements the common framework to their simulation model. (5) *Evaluate simulation models*. Benchmarking results are compared among models and models are assessed. (6) *Improve simulation models*. Model deficiencies are identified and rectified (as an example, see the study by Nicolle et al. 2014).

Best-in-class benchmarking is not just an extension of benchmarking analysis to several models. Instead, it establishes a social process characterized as coopetition (see below) and a moving target which cannot be reached once and for all. Once best-in-class benchmarking has been chosen, the peers have to be identified and they have to be involved in a benchmarking project as benchmarking partners providing their simulation results and model details. This project establishes a hybrid process of cooperation and competition (Wolfram Cox et al. 1997), also known by the generic term coopetition (Brandenburger and Nalebuff 1998). The norm of competitiveness is endogenous to the endless competition in which best-in-class benchmarking engages the project partners. This makes the benchmarks proper, i.e. the peers' model results that serve as points of reference, a moving target—a target which cannot be reached because continuous improvement of the models moves it ahead.

18.3.1.3 Process Benchmarking

Process benchmarking looks across multiple disciplines or industries to identify successful validation practices—regardless of their source. It is reminiscent of the search for best practices in business benchmarking and supports continuous improvement, increased performance levels and movement towards best practices. This type of benchmarking may involve models on different phenomena in one discipline (internal benchmarking) or from different disciplines (external/interdisciplinary benchmarking). In companies in the simulation software industry, process benchmarking may be an element of a comprehensive quality assurance strategy. Like best-in-class benchmarking, process benchmarking can be characterized as a form of coopetition, but the element of competition is weaker than in best-in-class benchmarking.

Six major elements are proposed as a framework for process benchmarking: (1) *Develop benchmarking concept*. This step includes identification of the model domains and/or disciplines that desire evaluation. (2) *Select benchmarking partners*. A number of simulation models in the relevant domains or disciplines are identified and modellers contacted and asked for their cooperation in the process benchmarking project. (3) *Develop common framework*. A common evaluation framework is negotiated by the partners. However, this framework does not include a common set of validity criteria, a common set of validation benchmarks (benchmark variables and the related benchmarks proper), appropriate metrics, or a testing scheme as in best-in-class benchmarking. Instead, partners agree to collect data or provide information (e.g. by way of interviewing) on their practices in model development, testing, and validation as well as on the performance of their models. (4) *Implement common framework*. Each partner collects data or provides information on their simulation practices. (5) *Evaluate simulation processes*. Benchmarking results are compared among partners and models and best practices are identified. (6) *Improve simulation validation processes*. Process deficiencies are identified and improved based on best practices.

Process benchmarking is not the focus of this chapter and will not be elaborated further here.

18.3.2 Criteria of Benchmark Selection

As explained above (see Sect. 18.3.1.1), benchmark selection includes three distinct tasks: identification of benchmark variables, choice of the scale on which each benchmark variable will be measured, and choice of the measurement precision/error (or resolution) with which each benchmark variable will be measured (in some areas of application, there may be no such choice). First of all, the identification of benchmark variables is specific to the problem addressed by the simulation model. For example, in hydro-meteorological forecasting, the suitability of benchmark variables was found to depend on the model structure used in the forecasting system, the season, catchment characteristics, river regime and flow conditions with little consensus on which benchmark variables are most suited to which application (Pappenberger et al. 2015, p. 698). Nevertheless, general criteria of benchmarks exist, e.g. from theoretical (here, physics), formal and practical points of view. A review of benchmarking studies reveals that the selection of benchmark variables depends on methodological (validity, independent observations) and practical criteria (maintenance and efficiency).

18.3.2.1 Validity Criteria

Benchmarks has to meet the criteria of objectivity, effectiveness, and reliability. Luo et al. (2012, p. 3861f.) argue that an objective benchmark likely derives from data or data products because data can objectively reflect natural processes in the real world. They allow that in some instances, models of previous versions or statistical models be used as benchmarks to gauge improvement of model performance. To be effective, a benchmark should reflect fundamental properties of the systems modelled. They argue that the more variable a data set, the less reliable the benchmark.

18.3.2.2 Independent Observations Criterion

Different model structures and different model parameter sets might lead to rather a similar model outputs, and in particular, rather similar performance when compared with any available observational data—a feature that has been called equifinality (Beven 2006; see also Chap. 32 by Beven in this volume). To mitigate the equifinality issue, benchmarks should draw upon a broad set of independent observations spanning, e.g. multiple temporal and spatial scales (Luo et al. 2012, p. 3862).

18.3.2.3 Maintenance and Efficiency Criteria

Pappenberger et al. (2015) explain that each benchmark variable that was tested had to be easy to maintain and computationally inexpensive. Benchmark variables have distinctly varying computational costs (see e.g. Pappenberger et al. 2015, Fig. 10) and this would often be the main barrier to implementing a benchmark operationally. Some benchmark variables require a high degree of human supervision or optimization. For example, many post-processing methods have a component of forecasting errors, which could be used to formulate a benchmark. Requiring supervised fitting, the implementation of such benchmark variables was excluded. High-quality benchmarking data may be too expensive to collect. For instance, Fewtrell et al. (2011) benchmark urban flood models using coarse scale topographic data with high and lower resolution. The most accurate (high-resolution) data set was not proposed as standard benchmarking data, because it was too expensive to collect. Rather, a less expensive but sufficiently accurate data set was advocated. The choice was complicated by the inconsistent robustness of single variables to the resolution; for instance, water depth estimates at grid scales coarser than 1 m appeared robust while velocity estimates were not.

As Pappenberger et al. (2015, p. 708) note, the ‘most useful and honest benchmark’ is one that is tough to beat. But what does it mean to beat a benchmark? The statement relates to a particular interpretation and use of the benchmarking data in the comparison. I suggest a broader view in which the benchmarking data is used to assess performance. There are different types of validation benchmarks.

18.4 A Typology of Validation Benchmarks

In this section, I propose a typology of validation benchmarks for different domains of applicability and intended applications of simulation models. The type construction is based on Weberian ideal types (Weber 1978 [1921]). To form an ideal type, a conceptual extreme is constructed from empirical reality. In its conceptual purity, the ideal type does not exist in empirical reality. An ideal type is a heuristic tool and its value depends on its utility. Thus, ideal types do not require that all of the requirements are met for each single case. My aim is to highlight the versatility of validation benchmarks and performance benchmarking.

Four types of validation benchmarks are distinguished: touchstone benchmarks, yardstick benchmarks, standard benchmarks and strong-sense benchmarks. The distinction arises from different uses of the reference point in the comparison, e.g. as a means for validating explanations, forecasts, or newly developed engineered systems in which different requirements apply. If the point of reference is interpreted as a touchstone, the benchmark is used to distinguish superior models from inferior ones. If the point of reference is interpreted as a yardstick, the benchmark constitutes a hypothetical best-practice result that the simulation model's results should approach at its best. If the point of reference is interpreted, in contrast, as an engineering reference standard, the benchmark establishes requirements for the simulation model's results. The interpretation leads to different decisions if the benchmark is not reached by the simulation result. Models that do not reach their yardstick can be published if the results are ambitious while models that do not meet their touchstone or standard will be refuted and their interpretation or engineering application denied.

These types depend on the domain of applicability and the intended applications of the model which determine the data and accuracy requirements, setting limits on the source of the benchmark. The latter influences the error characteristics, which is connected to the feasibility of uncertainty quantification of the benchmarking data. Different metrics apply, the benchmarking process being endorsed by different types of actors such as individual scientists, scientific associations or standardization bodies (see Table 18.2).

Given the complexity of simulation models, the validation of a single model may include several types of benchmarks. For example, the study by Luo et al. (2012) uses yardstick and touchstone benchmarks.

18.4.1 Strong-Sense Benchmarks

The concept of strong-sense benchmarks has been introduced by Oberkampf et al. (2004). Considered an engineering reference standard, they should be of a very high quality in order to support, e.g. the development and maintenance of high-consequence, engineered systems in fields such as nuclear reactor safety, underground nuclear waste storage, and nuclear weapon safety. Oberkampf et al. (2004)

Table 18.2 A typology of validation benchmarks

	Touchstone benchmark	Yardstick benchmark	Standard benchmark	Strong-sense benchmark
Interpretation	Touchstone	Yardstick	Engineering reference standard	Engineering reference standard
Confidence	Qualitative	Quantitative	Quantitative	Quantitative
Assessed dimension of validity	Reasonableness, robustness, (accuracy)	Accuracy	Accuracy, reliability	Accuracy, reliability
Domain of applicability	Research, in particular basic research, inter-comparison studies	Research	Engineered systems	High-consequence, engineered systems
Intended applications of model	No true application; explain and understand phenomena	Make forecasts, in particular maintain large-scale forecasting systems	Develop and maintain engineered systems	Develop and maintain high-consequence, engineered systems
Data requirements	No-data situations, (observational data)	Historical (observational) data	Experimental data	Experimental data
Source of benchmark	Stylized facts, functional relationships or patterns, (observations)	Observations or observation proxies, experiments	Case examples, experiments	Validation experiment
Accuracy requirement	Low	Depends on application	High	Very high
Error characteristics	Unknown	Depends on application	Small error	Small error
Uncertainty quantification of benchmarking data	Infeasible	Depends on application	Required	Required
Metrics	May be irrelevant	Relevant	Central role	Central role
Benchmarking process endorsed by	Single scientist, group of scientists	Group of scientists, scientific association	Standardization body or corporation	Standardization body or corporation

and Oberkampf and Trucano (2008) claim that strong-sense benchmarks can be used in any field of simulation, not just high-consequence simulations. The authors contend that strong-sense benchmarks in the sense they define them do not presently exist in computational physics or engineering.[3] Note that whether the system of interest, e.g. a component of a nuclear power plant, meets its performance or safety requirements is a topic separate from the question of model validation (Oberkampf and Trucano 2008, p. 725).

For this domain of applicability and intended use, the accuracy requirement is very high, and there is also a requirement of high reliability. Validation experiments (as a distinct, new type of experiment) is proposed to obtain benchmark data. They are conducted for the primary purpose of determining the predictive accuracy of a computational model, involving experimentalists, mathematical model builders, simulation analysts, code developers and code users. Methodological guidelines and procedures for designing and conducting validation experiments are summarized in Oberkampf and Trucano (2008). The goal of these experiments is to establish quantitative confidence in the code being used for its intended application. Therefore, high-quality validation metrics based on statistical procedures are required. Uncertainty quantification of benchmark measurements requires that estimates are provided of experimental uncertainty for all quantities measured, as well as uncertainty estimates of all the quantities that could be used as possible inputs for the computational simulation (e.g. boundary conditions, initial conditions, material properties). The benchmarking process is endorsed by a standardization body or corporation, such as NAFEMS (National Agency for Finite Element Methods and Standards).

18.4.2 Standard Benchmarks

I introduce the concept of standard benchmarks for engineered systems that do not cause tremendous damage if the systems fail. While most features are congruent to the case of strong-sense benchmarks, the major difference is that the accuracy requirement being high, the source of the benchmark is case examples or experiments, but not distinct validation experiments. For an illustrative example, see the study by Lund et al. (2012) on the validation of multibody musculoskeletal models.

Even if the damage in the case of invalid models is not tremendous, it may nevertheless be severe. This distinguishes engineered applications from research in the natural and social sciences, where errors, uncertainties and unqualified use are unwelcome, but typically do not have such serious consequences. There is a dividing line between the domains of research and real applications, offering the opportunity for a type of benchmark that differs from a standard.

[3] The fact that the authors contend that strong-sense benchmarks do not presently exist in computational physics or engineering makes me hesitant to take over their claim that strong-sense benchmarks can be used in any field of simulation (see Table 18.2, where I specify high-consequence, engineered systems as the domain of applicability).

18.4.3 Yardstick Benchmarks

Interpreted as a yardstick, a benchmark constitutes a hypothetical best-practice result that the simulation model's results should approach at its best. The goal is to establish quantitative confidence in the code being used for making forecasts, in particular when maintaining large-scale forecasting systems. Such benchmarks are particularly useful for automatic quality checking of large-scale forecasts and when forecasting system upgrades are made.

The sources of yardstick benchmarks are diverse, ranging from observations to observation proxies and experiments. As a consequence, the accuracy requirement, error characteristics and the feasibility of uncertainty quantification of the benchmarking data vary with the data and the intended application. In the social sciences and the life sciences, there are not only a large number of parameters but also often an intra and intersubjective variability, difficult to describe in terms of probability distributions, often making uncertainty quantification infeasible, while in physical systems, uncertainty quantification is certainly feasible (see Chap. 22 by Dalton et al. in this volume). The benchmarking process is endorsed by a group of scientists or a scientific association. For illustrative examples, see the studies by Hoffman et al. (2017), Pappenberger et al. (2015), Luo et al. (2012), and Fewtrell et al. (2011).

18.4.4 Touchstone Benchmarks

Benchmarks is interpreted as touchstones in models explaining and understanding phenomena constituting no application in the strict sense of the word, such as toy models. They provide qualitative confidence in the code, and their validation does not focus on accuracy. Accordingly, Schlesinger et al. (1979, p. 104) definition of simulation validation that relies on a '*satisfactory range of accuracy*' is not suitable here. Rather et al. (2000, p. 202f.) definition may apply. Validation is conceived as a proactive, diagnostic effort to ensure that the model's results are reasonable and credible. Used in no-data situations (e.g. counterfactual simulation experiments) or in disciplines or applications with high errors in observational data and with high uncertainty with respect to model parameters, variables and structure, the reasonableness and robustness of the simulation results is more important than their accuracy (see also Chap. 4 by Murray-Smith in this volume). The source of touchstone benchmarks are qualitative characteristics of the target, as represented in stylized facts (see Chap. 16 by Meyer in this volume) or functional relationships or patterns. Uncertainty quantification of this sort of benchmarking data is infeasible. The benchmarking process is endorsed by an individual scientist or a group of scientists. For illustrative examples, see the study by Harding et al. (2010).

In addition, there is a second use of touchstone benchmarks. In intercomparison projects, a benchmark may be used to set a minimum level of acceptability for model performance of the models compared. In this case, Schlesinger et al. (1979, p. 104) or Caldwell and Morrison (2000) definition of simulation validation may be applied, depending on the error characteristics of the benchmarking data and the relevance of unknown model parameters and variables. See Nicolle et al. (2014) for an illustrative example relying on Schlesinger et al. definition.

18.5 Metrics Related to Benchmarking

While the term benchmarking metrics is used by some scientists (e.g. Luo et al. 2012; Hoffman et al. 2017), there seems to be no distinct class of benchmarking metrics from a statistical point of view. Rather, a wide variety of procedures are used that are known from descriptive statistics to forecast verification.[4] Basically, what is referred to as benchmarking metrics seems to belong to the badly arranged field of validation metrics (see Chap. 13 by Marks in this volume)—being identical with rather than a subclass of the latter class. As Oberkampf and Trucano (2008, p. 738) note, validation metrics is not only in an early stage of development. In particular, there is no overview—spanning all disciplines and their applications—of the criteria for choosing a distinct metric. One reason may be that validation metrics are also applied beyond the field of simulation validation, with different presuppositions.

In this subsection of the validation benchmarks chapter, I cannot achieve such an overview. Even after this volume has been published, this seems to remain a desideratum. Instead, I introduce some major distinctions and justify the selection of some metrics in this subsection. To improve terminological clarity, I introduce a distinction that has yet to be made in the literature—I distinguish validation metrics from benchmarking metrics. *Validation* metrics denotes a metric calculated from two types of variables, e.g. simulated temperature and observed temperature. *Benchmarking* metrics denotes a metric calculated from three types of variables with one variable being a benchmark variable, e.g. simulated temperature, benchmark temperature and observed temperature. Put differently, only benchmarking metrics imperatively require a benchmark variable. Defined in this way, skill scores (see below) are the most important class of benchmarking metrics, while most of the metrics actually used for calculating the fit between simulation outcome and benchmark data are validation metrics.

The choice of a validation metric seems to depend on (i) typical statistical features of the variables in the model, such as scale level; (ii) the simulation approach: dynamic model (see the definition by Hartmann 1996, p. 83) or Monte Carlo simulation study; (iii) the preferred definition of validity: highlighting accuracy or reasonableness; (iv) the accuracy requirement for the intended application of the model; (v) the error and uncertainty characteristics of the benchmarking data; (vi) the empirical relevance of omitted and/or unknown parameters and variables. Several of these criteria are

[4] The term forecast verification is preferred in the atmospheric sciences, however, it is explained that synonyms are evaluation or validation.

interrelated. I do not claim that the list is complete or even includes all of the important criteria.

In the following, I concentrate on those metrics that are actually used in benchmark studies in the empirical sciences. The overwhelming majority of these studies are based on dynamic models as defined by Hartmann (1996). Disciplines cover hydrology, meteorology, climatology and biogeosciences. Two comprehensive books or chapters are Jolliffe and Stephenson (2011) and Wilks (2011, Chap. 8). Validation metrics for Monte Carlo simulation models are the focus of Marks (see Chap. 13 in this volume). A short overview on validation metrics in engineering sciences—which are not in the focus of this volume—is presented in Liu et al. (2011) and Oberkampf and Barone (2006).

18.5.1 Basic Concepts

Basic concepts suited to comparing simulation results and benchmarking data include (definitions adapted to benchmarking):

Accuracy. The level of agreement or average correspondence between individual pairs of simulation results and benchmarking data. The difference between the simulation results and the benchmarking data is the error. The lower the errors, the greater is the accuracy (cf. Chap. 5 by Roy in this volume).

Bias (or unconditional bias, or systematic bias). The correspondence between the average simulation results and average benchmarking data. This concept is different from accuracy. For instance, simulated temperature that is consistently too warm exhibits bias whether or not the simulations are otherwise reasonably accurate or quite inaccurate.

Association. The strength of the linear relationship between simulation results and benchmarking data (e.g. Pearson's correlation coefficient measures such a linear association, see Sect. 18.5.2; cf. also Chap. 19 by Robinson in this volume).

Although accuracy measures measure accuracy, they do not give an answer to the question of what constitutes *acceptable* accuracy. To better assess model performance, it is suggested that one compare one's simulation results with results obtained in some other way. The concept of skill is based on this idea:

Skill. The performance of the simulation results relative to some 'unskilful' reference results. For example, in weather forecasts common choices for the reference forecasts are climatological values of the predicand (mean values over some recent reference period, typically of 30 years length), persistence forecasts (values of the predicand in the previous time period) or random forecasts (see also Sect. 18.5.3 for more details).

Pappenberger et al. (2015, p. 709) recommend that during the step of benchmark selection a range of performance measures and skill scores be calculated before a decision is made on the most suitable benchmark variables and their characteristics.

18.5.2 Measures of Accuracy

Scalar measures of accuracy are suited to summarizing, in a single number, the overall performance of a set of simulations. Metrics have been developed for dichotomous, multi-category and continuous variables, as well as for deterministic or probabilistic models (including the special case of ensemble forecasts). For reasons of space, only a short introduction into metrics for benchmarking can be given in this subsection of our chapter on validation benchmarks. Therefore, this subsection will concentrate on methods for continuous variables. In forecast verification (see Wilks 2011, p. 324), it is common for scalar performance and skill measures, computed using individual simulation/benchmark pairs, to be used in evaluating continuous non-probabilistic simulation results.

Common measures include bias, mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE) and correlation. Definitions, questions addressed, and features such as range and perfect score are summarized in Table 18.3. Mean error (also called the additive bias) and (multiplicative) bias do not measure the magnitude of the errors, and therefore, they are not accuracy measures in and of themselves. They do not measure the correspondence between simulation results and benchmark data, i.e. it is possible to obtain a perfect score for a bad forecast if there are compensating errors. Bias is best suited to quantities that have 0 as a lower or upper bound. MAE, MSE and RMSE are common coefficients which measure the magnitude of the errors. MAE and MSE do not indicate the direction of the deviations. The MSE will be more sensitive to larger errors than will the MAE, and hence it will also be more sensitive to outliers. It can be decomposed into component error sources following Murphy and Winkler (1987). Putting more weight on large errors than smaller errors is appropriate if large errors are especially undesirable. Often the MSE is expressed as its square root, RMSE, which has the same dimension as the simulation results and the benchmarking data.

The correlation coefficient r can be used to measure the linear association of simulation results and benchmarking data (see Table 18.3). r does not take bias into account—it is possible for simulation results with large errors to still have a good correlation coefficient with the benchmarking data. r is sensitive to outliers.

For continuous probabilistic simulation results such as with ensemble forecasts, the continuous ranked probability score (CRPS, Hersbach 2000; see Table 18.3) is a well-suited measure. The CRPS compares the distribution of the simulation results with the benchmarking data. It ranges from 0 to infinity with lower values representing a better score. It collapses to the MAE for deterministic forecasts.

Table 18.3 Overview of basic validation metrics and benchmarking metrics (indicated by asterisk) for continuous variables

Name of measure	Definition	Range	Perfect score	Question addressed
Mean error (additive bias)	$MeanError = \frac{1}{N}\sum_{i=1}^{N}(S_i - B_i)$	$[-\infty, \infty]$	0	What is the average error of the simulation results?
Bias (multiplicative bias)	$Bias = \frac{\frac{1}{N}\sum_{\mathrm{i}=1}^{N} S_i}{\frac{1}{N}\sum_{i=1}^{N} B_i}$	$[-\infty, \infty]$	1	How does the average magnitude of the simulation compare to the average magnitude of the benchmark?
Mean absolute error (MAE)	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert S_i - B_i\rvert$	$[0, \infty]$	0	What is the average magnitude of error of the simulation results?
Root mean square error (RMSE)	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - B_i)^2}$	$[0, \infty]$	0	What is the average magnitude of error of the simulation results?
Mean squared error (MSE)	$MSE = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}(S_i - B_i)^2$	$[0, \infty]$	0	What is the average magnitude of error of the simulation results?
Correlation coefficient	$r = \frac{\sum(\mathrm{S}-\bar{\mathrm{S}})(\mathrm{B}-\bar{\mathrm{B}})}{\sqrt{\sum(\mathrm{S}-\bar{\mathrm{S}})^2}\sqrt{\sum(\mathrm{B}-\bar{\mathrm{B}})^2}}$	$[-1, 1]$	1	How well did the simulated values correspond to the benchmark values?
Continuous ranked probability score (CRPS)	$CRPS = \int_{-\infty}^{\infty}(P_s(x) - P_b(x))^2 dx$	$[0, \infty]$	0	How well did the simulated probabilities predict the benchmark values?
Skill score* (generic form)	$\mathrm{SS}_{\mathrm{ref}} = \frac{(A_f - A_{ref})}{(A_{perf} - A_{ref})} * 100\%$	[depends on score used, 1]	1 (0 indicates no improvement over the reference results)	What is the relative improvement of the simulation results relative to some ‘unskilful’ reference results?

18.5.3 Skill Scores

The skill of a simulation study can be assessed by how close the simulation results are to the observations compared to how close a benchmark was. This implies a significant shift in perspective: skill analysis does not examine the difference or relationship between simulation results and benchmarking data. Benchmarking data is not a substitute of observational data. Instead, three sources of data are assumed: artificial data (simulation results), observational data, and benchmarking data. The relationship between skill, performance of the simulation and a benchmark can be generalized as: skill ~*f*(simulation, observations)/*f*(benchmark, observations). Here, *f* denotes a function (i.e. validation metric) which expresses the difference between quantities, the simulation or benchmark values and the observed values for some phenomenon. Skill scores for different simulation models can only be compared for the same underlying benchmark as well as a given validation function or metric.

Skill analysis allows a simulation model to be classified as having (a) *no skill*: the simulation model's results are consistently worse than a set benchmark; (b) *naïve skill*: the simulation model is skilful against too simplistic a benchmark. More challenging (difficult to beat) benchmarks could be designed; or (c) *real skill*: no benchmark which can be implemented at a lower cost than the operational system can beat the simulation model. Naïve skill should be avoided. This is the situation in which Pappenberger et al. (2015, p. 708) above-quoted statement applies: the most useful and honest benchmark for use in simulation model validation is one that is tough to beat. Benchmarks that are too naïve can easily result in a high skill being calculated. Thus, the importance of using benchmarks that are known and understood is essential in assessing how 'good' a simulation model's results are.

The skill (also called forecast skill) is usually presented as a skill score. In generic form, the skill score for forecasts *f* characterized by a particular measure of accuracy *A*, with respect to the accuracy A_{ref} of a set of reference forecasts is given by $SS_{ref} = (A_f - A_{ref})/(A_{perf} - A_{ref}) * 100\%$, where A_{perf} is the value of the accuracy measure that would be achieved by perfect forecasts and A_f the value of the accuracy measure that would be achieved by the forecasts f. If $A_f = A_{perf}$ the skill score attains its maximum value of 100%. If $A_f = A_{ref}$ then $SS_{ref} = 0\%$, indicating no improvement over the reference forecast. If the forecasts being evaluated are inferior to the reference forecast with respect to the accuracy measure *A*, $SS_{ref} < 0\%$ (Wilks 2011, p. 305).

The skill score has to be specified for a suitable measure of accuracy *A*, e.g. by the MSE, RMSE or the CRPS, for all *Ai*. Note that perfect forecasts have RMSE, MSE or MAE = 0, which allows the rearrangement of the skill score, e.g. to $SS_{ref} = 1 - MSE_f/MSE_{ref}$ (in proportion rather than percentage terms).

The skill score thus is suited to indicating the improvement of a forecast based on simulations (A_{sim}) relative to forecasts based on benchmarking data (A_{bench}). Specified for MSE, we obtain $SS_{bench} = 1 - MSE_{sim}/MSE_{bench} = 1 - (1/n\sum(S_i - O_\mathrm{i})^2/1/n\sum(B_i - O_\mathrm{i})^2)$, with S_i denoting the simulation results, B_i the benchmarking data, and O_i the observational data (Perrin et al. 2006).

18.5.4 Murphy–Winkler Framework and Beyond

Traditionally, forecast verification has emphasized accuracy and skill, concepts which have been amended by Murphy (1993). In particular, the Murphy–Winkler framework (Murphy and Winkler 1987) adds the concepts of reliability, resolution, discrimination and sharpness. The choice of scores depends on which of these attributes is most important to the scientist. This framework, which is based on the joint distribution of the forecasts and observations, may be transferred to the joint distribution of simulation results and benchmarking data. It can be used as a diagnostic tool for decomposing error sources.

A concise overview of metrics suitable for validation benchmarks for variables on all scales, for deterministic and probabilistic models as well as for use with models that have a spatial component can be found in Jolliffe and Stephenson (2011) and Wilks (2011). Granger and Jeon (2003) have extended the concept of distance to distances other than vertical distances indicating distances between two time series. 'Horizontal' distances can be calculated for one series leading or lagging the other—known as time–distance. A similar distance can be calculated for spatial features if a movement is faster or slower than expected.

18.5.5 Holistic Measurement

The challenge for validation metrics and benchmarking metrics is to holistically measure model performance. As a comprehensive benchmarking study usually considers many variables, a suite of metrics across several variables has to be synthesized on the relevant scales (e.g. spatial, temporal) on which the model operates. Coefficients used include, e.g. Taylor skill to represent the degree to which simulations matched the temporal evolution of variables, NMAE (normalized mean absolute error) to quantify bias (i.e. the 'average distance' between observations and simulations in units of observation—the spatial dimension), and reduced chi-squared statistics to quantify observational uncertainty (Schwalm et al. 2010). How to combine coefficients to holistically represent model performance skill is an issue for further research in validation metrics. For instance, coefficients could be weighted according to the intended use of the model (see also Chap. 24 by Liu and Yang in this volume).

To facilitate holistic assessment, new types of diagrams have been developed. For example, Taylor diagrams have been developed to visualize benchmarking results and plot several statistics at one time (Taylor 2001).

18.6 Discussion

As a method, benchmarking unfurls a social dynamic which is unique. It produces normalizing knowledge (see below) even in the absence of standardization bodies. In the discussion, I explain the underlying social mechanism and I highlight the pitfalls of unfocussed benchmarking.

18.6.1 Normalizing Simulation Validation

Drawing on Michel Foucault's (2008) work on governmentality, benchmarking has been described as a general technology of performance in and through which various subjects, such as companies or states, and spaces are constituted and acted upon as governable objects. In particular, benchmarking has been conceived as a normalizing governmental technology (Triantafillou 2004, p. 496), normalization being understood in terms of 'the procedures and processes through which a norm is brought into play and informs the practices that it seeks to regulate'. Benchmarking encourages or stimulates self-governance via the production of normalizing knowledge.

We should expect then, that the use of benchmarking in simulation validation will not only improve a simulation model's performance but support the establishment of norms and standards in simulation validation. Benchmarking imposes competitive pressure on simulating scientists without directly coercing them to benchmark or seeks to improve their performance according to the norms and standards in question. Benchmarking activities will indirectly coerce scientists into improving the performance of their simulations.

Governmentality studies have repeatedly emphasized that benchmarking is not neutral with respect to the selected benchmarks (e.g. Fougner 2008; Bruno 2009). Rather, they reflect the theoretical frameworks informing those who select the benchmarks. In this respect, benchmarking in simulation validation supports the production of normalizing knowledge in different theoretical frameworks put forward by simulating scientists in diverse fields.

18.6.2 The Social Character of Validation Benchmarks

When one is selecting validation benchmarks (benchmark variables and the related benchmarks proper), a social aspect has to be discussed: Benchmarks are set in a benchmarking process. A benchmark fundamentally represents an agreement of a community. For instance, in business benchmarking and computer benchmarking (Vieira and Madeira 2009, p. 69), benchmarks are agreed upon by the companies involved, and are sometimes even endorsed by a standardization body or corporation, e.g. the Transaction Processing Performance Council (TPC). The agreement may be explicit or tacit. If there is a formal agreement, the benchmark turns from a point of reference to a standard. The need to develop a consensus by experts on defining and selecting benchmarks is also emphasized by simulating scientists (e.g. Luo et al. 2012, p. 3864; see also Chap. 23 by Schlünzen in this volume).

Basically, most benchmarks are data of some sort. As Murray-Smith has pointed out (see Chap. 15 in this volume), data are the major points of reference for the validation of a simulation model's results. But while most benchmarks are some kind of data, the opposite does not hold. The process of benchmarking assumes a key position. The benchmarking process makes some kind of data a benchmark (while the overwhelming amount of data just remains data). The feature of agreement is thus an important one. Here I wish to critically discuss its significance for simulation validation.

A short review of benchmarking practices in diverse fields shows that agreement on the benchmarks is no general requirement for benchmarking. Fougner (2008), for instance, discusses competitiveness indexing and country benchmarking by the World Economic Forum and the International Institute for Management Development. While there is consensus on the benchmarks on the part of the benchmarking agencies, there is no agreement needed from the benchmarked countries. This example suggests distinguishing between an agreement required of prospective benchmarking partners in a benchmarking project, and consensus on the pertinent benchmarks. This distinction clarifies that benchmarks in simulation validation need not be agreed upon right from the beginning. In principle, a single scientist can evaluate simulation outcomes according to a point of reference that she thinks is important. Of course, her evaluation may then not matter for other people. There may be a gradual process starting from performance benchmarking with benchmarks that are not yet agreed upon by a community of researchers, eventually advancing in the direction of best-in-class benchmarking with consensus on the pertinent benchmarks. This gradual process prevents a scenario in which only disciplines with large simulation communities make use of benchmarks. It will help those empirical sciences without a significant community of simulating scientists to start working with benchmarks. At present, in disciplines which adopt computer simulation as a new scientific method, there are many single simulation models and a significant scientific community often does not exist yet (and may only develop slowly in the near future).[5]

The potential danger arising from the use of benchmarks that do not result from an agreement have been described by the concept of bench-marketing. In the computer industry, companies developed their own, highly biased benchmarks and misused their computer's good performance measurements vis-à-vis these benchmarks for marketing purposes. Companies created configurations that maximized performance against the benchmark, not against real-world applications. This gave rise to the establishment of standardization bodies, e.g. the Transaction Processing Performance Council (Nambiar et al. 2014, p. 2).

[5] The fact that many authors created their own benchmarks out of necessity has been reported even for computer benchmarks, see Stratton et al. (2012, p. 1).

18.6.3 Between Validation and Comparison—the Limitations of Benchmarking

Benchmarking aims to improve a simulation model's performance by comparing the simulation outcome with a reference point. As explained above, the related metrics denote, e.g. the magnitude of the simulation error, the linear association of simulation results and benchmarking data, or the skill of a forecast. While these metrics are suited to measuring the improvement of consecutive versions of a simulation model, there is often no definition of a level that is still acceptable or a test that has to be passed. This holds particularly for yardstick benchmarks. As a consequence, the decision on the validity of a simulation model and its results rests on the evaluation of the obtained error, association or skill, etc. The challenge of holistically measuring model performance makes this evaluation even more difficult—making it prone to subjective judgments and bias. If there is some agreement between the simulation output and the point of reference, then the model is declared 'validated'—meaning that one trusts the model because it reproduces some important features of the target sufficiently well. At this stage, a judgment is made by the simulating scientist that the obtained error, association or skill, etc. either possesses a satisfactory range of accuracy consistent with the intended application of the model (referring to Schlesinger et al. 1979 SCS definition of model validation) or is satisfactory as to indicate that the model's outputs are reasonable for their intended purposes (referring to Caldwell and Morrison 2000 definition). Only standard benchmarks and strong-sense benchmarks require the definition of a level that is still acceptable or a test that has to be passed. There is no consensus in the validation benchmarks literature whether validation metrics should include or be supplemented with, e.g. hypothesis testing. There, the validation assessment is formulated as a decision problem to determine whether or not the computational model is consistent with the benchmarking data. This allows acceptance of a model and its results as valid or its rejection. For instance. Oberkampf and Barone have argued that 'validation metrics should be measures of agreement, or disagreement, between computational models and experimental measurements; issues of adequacy or satisfaction of accuracy requirements should remain separate from the metrics' (Oberkampf and Barone 2006, p. 12).

This statement leads to the discussion of recommended features of validation or benchmarking metrics. This discussion is split and scattered in different disciplines, and cannot be outlined or summarized here. To illustrate the questions addressed, I again refer to Oberkampf and Barone (2006, pp. 11f.), who discuss, e.g. whether a metric should include an estimate of the numerical error in the simulated output variable of interest, the measurement errors in the benchmarking data, or the error resulting from post-processing of the benchmarking data, or whether a metric should depend on the number of experimental measurements of the benchmark variable.

Next, the limitations of benchmarking are discussed based on the example of model-to-model comparison. While this special case of best-in-class benchmarking fulfils the defining criterion of benchmarking, because each model's results serve as points of reference for the validation of the other model's results, there seems to be no general argument making one of them the superior model. This example seems to be the limit case of a validation that ends up as a mere comparison if no theoretical argument is found in favour of the superiority of one model. Only theoretical argumentation makes a model-to-model comparison an exercise in model validation. Without empirical or experimental data as points of reference, only sound arguments can establish that one of the models is the most credible one.

18.6.4 The Price of Efficient Benchmarks

A further aspect is the choice of a benchmark to reduce computational cost (see e.g. Fewtrell et al. 2011). Although constituting a legitimate goal, in particular for automatic quality checking of large simulation models and for when system upgrades are made, or if in the case of emergency, e.g. expected floods, quick warning is needed, this practice establishes a second criterion—validation at low cost. From an epistemological point of view, this practice reduces the standards for simulation validation. Typically, calculating metrics is followed by benchmarking diagnostics, i.e. the analysis of errors. Often, more cost-efficient benchmarks have lower precision, resulting in a loss of knowledge from benchmarking diagnostics.

18.6.5 The Devaluation of Benchmarks Proper

A critical issue that damages the value of benchmarks in the evaluation of simulations is related to the parameterization of simulation models. Adjusting parameters is a necessary step for obtaining sound simulation results even in areas where theoretical knowledge is very strong. Parameters are adjusted using concepts such as calibration or tuning, but with a different tone. As Lenhard (see Chap. 39 in this volume) puts it, 'calibration' is commonly used in the context of preparing an instrument, like calibrating a scale one time for using it very often in a reliable way'. Another concept used for adjusting parameters is tuning. Tuning is related to achieving a fit with artificial measures, or to a particular case. The term is more pejorative. In practice, however, calibration and tuning are not always easily differentiated. Nevertheless, I will ignore this in the following, where I adhere to the notion that calibration is performed once and then finished. Tuning, however, is done in many consecutive steps in which validation metrics are calculated to conceive of the success of the tuning exercise. There is no restriction on the number of adjustments to be made. As the tuning of a parameter is performed according to the overall behaviour of the model, the success is measured on the same global level on which the metrics are calculated. Basically, the same metrics are calculated. Thus, tuning improves the agreement of the simulation outcome and the related benchmarks proper. Typically, the parameters which are tuned are not theoretically well motivated. The result may be that a model's output shows an overfit to the related benchmarks, rather than the 'true' fit. The problem is that with further tuning, the agreement between simulation output and benchmarks proper can technically be increased whereas there is no 'true' increase in the validity of the model and its results.

18.7 Conclusions

This chapter has presented a conceptual framework and descriptive theory of benchmarking in simulation validation. Validation benchmarks and the related techniques of benchmarking are a flexible method and key to the improvement of simulation validation. Sources of benchmarks are outstanding experimental or observational data, stylized facts or other characteristics of the target. They are outstanding because they are more effective, more reliable and more efficient than other experimental or observational data, stylized facts or characteristics of the target. Offering starting points even for toy models, benchmarking will promote the normalizing of simulation validation. In some disciplines, the development of a community-wide benchmarking system has just begun (e.g. for land models in the biogeosciences). From a normative philosophy of science perspective, benchmarking is valued particularly for supporting the epistemic value of accuracy and the social value of efficiency (see Chap. 40 by Hirsch Hadorn and Baumberger in this volume). Other epistemic values such as robustness of results and coherence with background knowledge are ignored by the benchmarking exercise.

For the further establishment of benchmarking in simulation validation, two topics should be of major concern: the development of a prescriptive benchmark theory and an integrative view on related metrics:

Benchmark theory. The weakness of the descriptive theory of benchmarking in simulation validation is obviously that it does not make prescriptions. A prescriptive theory is needed for the further development of the method. For instance, practitioners need to know when and why they should use which type of benchmark. The prescriptive theory is functional in providing some orientation. However, it does not provide methodological rules. Practitioners should also know how standards are derived (see Chap. 23 by Schlünzen in this volume) from benchmarks. Remarkably, we do not find such a benchmark theory in other fields, such as business benchmarking, computer benchmarking or benchmarking in the stock market either. However, the core of the method seems to be comparison, which is applied and theorized in many disciplines. This is a topic for interdisciplinary research by fields including methodology, philosophy of science, and mathematics.

Inventory. The development of suitable validation/benchmarking metrics is an important task. We lack an important precondition for doing this in a more systematic, rather than an ad hoc way—an integrative view spanning the scattered measures and their diverse applications in different disciplines. An inventory is recommended first and foremost. We need to know the available metrics as well as the criteria for choosing a distinct metric.

The development of suitable validation/benchmarking metrics has to address challenges on two levels:

New metrics. As Hoffman et al. (2017, p. 14) state, 'developing metrics that make appropriate use of observational data remains a scientific challenge because of the spatial and temporal mismatch between models and measurements, poorly characterized uncertainties in observationally constrained data products, biases in reanalysis and forcing data, model simplifications, and structural and parametric uncertainties'. In their epilogue, Jolliffe and Stephenson (2011, Chap. 12) note that meteorologists seem to have made much of their forecast evaluation without the collaboration of statisticians (ibid., p. 223)—with the exception of the Murphy–Winkler framework. They consider scoring rules (Gneiting and Raftery 2007; Gneiting 2011) as well as discriminant analysis (McLachlan 1992) as useful starting points for further studies in the subject. They review the critique of economists on measures of accuracy, e.g. the MSE, the development of formal tests of skill and of density forecasting (Tay and Wallis 2000). Recognizing diagnostic medical tests, they argue that atmospheric science has more measures but is less sophisticated (Jolliffe and Stephenson 2011, p. 227) than medical studies. They conclude that there is a need to move beyond purely descriptive sample statistics; in particular, sample scores should only be considered as finite sample estimates of the true scores of the system. Inference should be incorporated when calculating validation metrics (ibid., p. 228). Some of the newly developed metrics are introduced in Marks (see Chap. 13 in this volume).

Coefficients of coefficients. One of the major challenges that have to be addressed for the further establishment of validation benchmarks is the development of comprehensive aggregate or second-order metrics ('coefficients of coefficients') that summarize the benchmarking results for numerous model variables. This second-order metrics is needed to quickly inform us about the validity of competing complex models, a situation we will face much more often with the growing use of simulation models in all sciences.

Finally, I suggest researching decision theory to justify the use of benchmarks in philosophical terms. Since the development of prospect theory (Kahneman and Tversky 1979), there has been the notion of a benchmark of sorts in expected utility theory—here, a reference point, that is, an outcome that partitions the set of decision outcomes into perceived gains and losses. Wedgwood's (2013) benchmark theory offers a model that considers benchmarks as points of reference in rational decision-making. The theory ranks actions according to the desirability of an outcome produced in some state of affairs compared to a standard—a benchmark—for that state of affairs. Note, however, that Wedgwood (2017) explicitly rejects the idea that the value of an option is its utility. Robert (2018) applies Wedgwood's benchmark theory assuming—on the contrary—that the value of an option is understood to be its utility. At the centre of both authors' account is the idea of expected comparative value or expected comparative utility. These basic assumptions have to be specified for justifying benchmark selection as well as justifying the use of an appropriate metrics in simulation validation. Due to its philosophical origin, this theory will rather not provide formal criteria of benchmark selection for practitioners.

Acknowledgements The author thanks Claus Beisbart and William Oberkampf for helpful discussions concerning this manuscript.

References

Beven, K. J. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology, 320,* 18–36.

Brandenburger, A. M., & Nalebuff, B. J. (1998). *Co-opetition: A revolutionary mindset that combines competition and co-operation*. New York: Currency Doubleday.

Bruno, I. (2009). The 'indefinite discipline' of competitiveness benchmarking as a neoliberal technology of government. *Minerva, 47,* 261–280.

Caldwell, S., & Morrison, R. J. (2000). Validation of longitudinal dynamic microsimulation models. Experience with CORSIM and DYNACAN. In L. Mitton, H. Sutherland & M. J. Weeks (Eds.), *Microsimulation modelling for policy analysis. Challenges and innovations* (pp. 200–225). Cambridge: Cambridge University Press.

Fewtrell, T. J., Duncan, A., Sampson, C. C., Neal, J. C., & Bates, P. D. (2011). Benchmarking urban flood models of varying complexity and scale using high resolution terrestrial LiDAR data. *Physics and Chemistry of the Earth, 36,* 281–291.

Foucault, M. (2008). *The birth of biopolitics: Lectures at the College de France, 1978–1979*. Basingstoke: Palgrave Macmillan.

Fougner, T. (2008). Neoliberal governance of states: The role of competitiveness indexing and country benchmarking. *Millennium: Journal of International Studies, 37*, 303–326.

Gneiting, T. (2011). Evaluating point forecasts. *Journal of the American Statistical Association, 106,* 746–762.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association, 102,* 359–378.

Granger, C. W. J., & Jeon, Y. (2003). A time-distance criterion for evaluating forecasting models. *International Journal of Forecasting, 19,* 199–215.

Harding, A., Keegan, M., & Kelly, S. (2010). Validating a dynamic population microsimulation model: Recent experience in Australia. *International Journal of Microsimulation, 3,* 46–64.

Hartmann, S. (1996). The world as a process: Simulation in the natural and social sciences. In R. Hegselmann, U. Müller, & K. G. Troitzsch (Eds.), *Modelling and simulation in the social sciences from the philosophy of science point of view* (pp. 77–100). Dordrecht: Kluwer.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecasting, 15,* 559–570.

Hoffman, F.M., et al. (2017). International land model benchmarking (ILAMB) 2016 Workshop Report. DOE/SC-0186, U.S. Department of Energy, Office of Science, Germantown, Maryland, USA. https://doi.org/10.2172/1330803.

Jolliffe, I. T., & Stephenson, D. B. (Eds.). (2011). *Forecast verification: A practitioner's guide in atmospheric science*. Sussex/Oxford: Wiley-Blackwell.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–329.

Liu, Y., Chen, W., Arendt, P., & Huang, H. -Z. (2011). Towards a better understanding of model validation metrics. *Journal of Mechanical Design, 133*.

Lund, M. E., de Zee, M., Andersen, M. S., & Rasmussen, J. (2012). On validation of multibody musculoskeletal models. *Journal of Engineering in Medicine, 226,* 82–94.

Luo, Y. Q., et al. (2012). A framework for benchmarking land models. *Biogeosciences, 9,* 3857–3874.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley.

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather Forecasting, 8,* 281–293.

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review, 115,* 1330–1338.

Nambiar, R., et al. (2014). TPC state of the council 2013. In R. Nambiar & M. Poess (Eds.), *Performance characterization and benchmarking*, *TPCTC 2013* (pp. 1–15). Cham: Springer.

Nicolle, P., et al. (2014). Benchmarking hydrological models for low-flow simulation and forecasting on French catchments. *Hydrology and Earth System Sciences, 18,* 2829–2857.

Oberkampf, W. L., & Barone, M. F. (2006). Measures of agreement between computation and experiment: validation metrics. *Journal of Computational Physics, 217,* 5–36.

Oberkampf, W. L., & Trucano, T. G. (2008). Verification and validation benchmarks. *Nuclear Engineering and Design, 238,* 716–743.

Oberkampf, W. L., Trucano, T. G., & Hirsch, C. (2004). Verification, validation and predictive capability in computational engineering and physics. *Appl. Mech. Review, 57,* 345–384.

Oreskes, N. (2003). The role of quantitative models in science. In C. D. Canham, J. J. Cole, & W. K. Lauenroth (Eds.), *Models in ecosystem science* (pp. 13–31). Princeton University Press: Princeton.

Pappenberger, F., et al. (2015). How do i know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology, 522,* 697–713.

Perrin, C., Andreassian, V., & Michel, C. (2006). Simple benchmark models as a basis for model efficiency criteria. *Arch. Hydrobiol. Suppl., 161,* 221–244.

Robert, D. (2018). Expected comparative utility theory. A new theory of rational choice. *The Philosophical Forum, 49,* 19–37.

Schlesinger, S., et al. (1979). Terminology for model credibility. *Simulation, 32,* 103–104.

Schwalm, C.R., et al. (2010). A model-data intercomparison of CO_2 exchange across North America: Results from the North American Carbon program site synthesis. *Journal of Geophysical Research, 115*, G00H05, https://doi.org/10.1029/2009jg001229.

Seibert, J. (2001). On the need for benchmarks in hydrological modelling. *Hydrological Processes, 15,* 1063–1064.

Stratton, J.A., et al. (2012). Parboil: A revised benchmark suite for scientific and commercial throughput computing. IMPACT Technical Report. IMPACT-12-01. University of Illinois at Urbana-Champaign: Center for Reliable and High-Performance Computing.

Sundberg, M. (2011). The dynamics of coordinated comparisons: how simulationists in astrophysics, oceanography and meteorology create standards for results. *Social Studies of Science, 41,* 107–125.

Tay, A. S., & Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting, 19,* 235–254.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research, 106,* 7183–7192.

Triantafillou, P. (2004). Addressing network governance through the concepts of governance and normalization. *Administrative Theory and Practice, 26,* 489–508.

Vieira, M., & H. Madeira (2009). From performance to dependability benchmarking: A mandatory path. In R. Nambiar & M. Poess (Eds.), *Performance evaluation and benchmarking, TPCTC 2009* (pp. 67–83). Heidelberg: Springer.

Weber, M. (1978[1921]). Economy and society. Tr. by G. Roth and C. Wittich. Berkeley: University of California Press.

Wedgwood, R. (2017). Must rational intentions maximize utility? *Philosophical Explorations, 20,* 1–20.

Wedgwood, R. (2013). Gandalf's solution to the newcomb problem. *Synthese, 190,* 2643–2675.

Wilks, D. (2011). *Statistical methods in the atmospheric sciences*. Oxford: Elsevier.

Wolfram Cox, J. R., Mann, L., & Samson, D. (1997). Benchmarking as a mixed metaphor. disentangling assumptions of competition and collaboration. *Journal of Management Studies, 34,* 285–314.

Part V
Methodology—Mathematical Frameworks and Related Techniques

Chapter 19
Testing Simulation Models Using Frequentist Statistics

Andrew P. Robinson

Abstract One approach to validating simulation models is to formally compare model outputs with independent data. We consider such model validation from the point of view of Frequentist statistics. A range of estimates and tests of goodness of fit have been advanced. We review these approaches, and demonstrate that some of the tests suffer from difficulties in interpretation because they rely on the null hypothesis that the model is similar to the observations. This reliance creates two unpleasant possibilities, namely, a model could be spuriously validated when data are too few, or inappropriately rejected when data are too many. Finally, these tests do not allow a principled declaration of what a reasonable level of difference would be considering the purposes to which the model will be put. We consider equivalence tests, and demonstrate that they do not suffer from the previously identified shortcomings. We provide two case studies to illustrate the claims of the chapter.

Keywords Equivalence testing · Null hypothesis significance testing · Statistical models · Model validation

19.1 Introduction

Validation is an essential step in the construction and assessment of computer models that are intended for application to scientific and operational challenges (see e.g., Caswell 1976; Gentil and Blake 1981; Reynolds et al. 1981; Mayer and Butler 1993; Oreskes et al. 1994; Rykiel 1996; Loehle 1997; Vanclay and Skovsgaard 1997; Robinson and Ek 2000; Sargent 2012). Model validation is typically carried out for one of two reasons, namely, (i) to determine whether predictions from the model can be used (fit for purpose) and (ii) to determine whether the model sufficiently faithfully represents the processes it is designed to represent (goodness of fit). However,

A. P. Robinson (✉)
CEBRA, School of BioSciences, The University of Melbourne, Parkville, Victoria 3010, Australia
e-mail: apro@unimelb.edu.au

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_19

there is no consensus on what is the uniformly best way to proceed, in part due to the variety of models, model applications, and candidate tests.

The purpose of this chapter is to review ways of approaching the challenge of model validation using Frequentist statistical tools. The Frequentist approach to model validation is to expose the model to some statistical test that compares the model outputs against data that is the outcome of the process that the model is intended to represent. If the model survives the test, then it is validated. In order for this to be possible, the model must produce predictions or estimates that have a measurable analog. For example, a model may be intended to represent the physiological responses of an idealized forest to climate, and one of its outputs may be the growth of the forest. We could try to validate this model by comparing field measurements of forest volume growth against the predicted volume growth obtained from a suitably tuned execution of the model (Sect. 19.4.2).

Furthermore, the body of the chapter is devoted to a statistical framework for validation for deterministic models, that is, models for which the outcomes are derived from known functions of the inputs. We extend the class of models to include stochastic models, for which the outcomes are random, in Sect. 19.5.1.

Although the examples that we will examine in this chapter are reasonably simple, to keep our exposition brief, the principles that we outline can be applied to models of any level of complexity. The sole prescription is that the model produces predictions or estimates that have a measurable analog, so that the outputs of the model can be compared statistically with measures of the process that the model is to represent. These situations may require greater or lesser amounts of statistical modeling in order to provide resilient statistical inference—for example, accounting for temporal correlation, or measurement errors, or hierarchical data—and such contingencies are not covered here. Nonetheless, the principles in this chapter can be applied when a model produces quantitative outputs that should align with measurable real-world processes if the model is valid.

There are two important distinctions between the Frequentist statistical tools that have been used to validate simulation models. We briefly cover them here but will provide more detail in the chapter. First, some tools *estimate* the goodness of fit of a model, e.g., the root mean squared error (*RMSE*), as opposed to *testing* the goodness of fit. Analysts tend to report the estimates but often omit reporting their underlying variability. They might claim, for example, that the *RMSE* is less than a given number, but fail to acknowledge that the true value of the *RMSE* is uncertain, and the reported value is only an estimate of the true value. Goodness of fit tests provide a framework to capture this uncertainty, but lead to the second distinction, namely in the disposition of the hypotheses.

As we shall see below, these goodness of fit tests requires the specification of two hypotheses, a null (or default), and an alternative. Often, analysts will apply tests for which the *default* hypothesis equates to the outcome that the model is valid. This equation leads to counter-intuitive and undesirable behavior. A set of tests, called *equivalence tests*, reverse the usual polarity—in them, the null hypothesis equates to the outcome that the model is not valid—and they do not suffer from the disadvantages of the other tests.

Some prefatory remarks are in order. Our goal is not to start arguments about different techniques. Therefore, we do not indulge in naming and shaming authors that have approached the problem from different points of view. However, some referencing of prior work is essential to provide context. The reader is not assumed to have detailed knowledge of the vocabulary or conceptual apparatus that is applied in the Frequentist statistics framework; rather, this material is briefly and selectively introduced.

The chapter is structured as follows. We begin by defining the key concepts that are used for inference in Frequentist applied statistics, with some simple examples. Then we briefly describe and critique some of the ways that authors have used some of these statistics to validate models. Next, we introduce tests of equivalence, and demonstrate how these tests improve on the shortcomings of the traditional methods. We provide some examples of the different approaches before drawing our conclusions.

19.2 Frequentist Statistics

19.2.1 Important Background

Statistics, like most other disciplines, has developed a specialized vocabulary that is used to express its key concepts. The following material is intended to provide enough information to help the reader follow the chapter; it is not comprehensive, we will take shortcuts where possible. Those who are conversant with statistical thinking and tools may elect to skip this section and begin at Sect. 19.3.

Frequentist statistics is a branch of statistics that includes a collection of statistical tools that build upon a very specific interpretation of probability, namely that probability is used to represent only the outcomes of infinitely repeatable instances of experiments. This is in contrast to Bayesian statistics, in which probability can play a much more extensive role, being used to capture apparently non-repeatable instances such as the degree of belief of a rational individual as to whether it will rain on New Year's Day in 2050.

Informally, we could interpret Frequentist statistics as being a set of concepts and tools for learning about the world by repetition. Archetypes of Frequentist statistics include classic devices such as coin tosses, the throwing of dice, shuffling of cards, and drawing of colored balls from suitable urns. These devices can be generalized to match a bewildering array of different and more complex circumstances, forming the superstructure of Frequentist statistics. The interested reader can find more detail in Casella and Berger (1990), among many others.

19.2.1.1 The Basic Framework

We use statistics to study *processes* and *populations* based on some assumptions and a (hopefully) representative sample of data. For example, we might want to know the weights of units being produced by a manufacturing process, or the average height of all the trees in a forest. There are subtle but important differences between processes and populations but they are not relevant for our purposes. Here, we will assume that the interest is about a process, and that observations have been made from an arguably representative sample of units that have been taken from that process.

The assumption of representativeness can rely on the specification of the sampling process, for example, that it was *simple random sampling*. This approach is the basis of *design-based*, as opposed to *model-based*, Frequentist statistics. In any case, the assumption of representativeness can be justified by using graphical diagnostics to assess the fidelity of the sample to whatever is known or suspected about the process, in the sense that the sample shares statistical characteristics with its process. It is important to keep in mind that having a purely random sample is not essential for Frequentist statistical inference, but that doing so does make it easier to argue that the sample is representative.

19.2.1.2 The Random Variable

We use the *random variable* as a bridge between the useful ideas and tools of probability and things that we want to measure and observe. Here are two definitions. Informally, a random variable is a variable that we use to represent the outcome of a random process. More formally, a random variable is a function that creates a relationship (called a *mapping*) between the space of events and a probability line (that is, the line 0–1).

For example, we can represent the outcome of a single roll of a fair six-sided die using a random variable; we think that each of the six possible outcomes has an equal probability, so the random variable is a function that maps the rolls into one of the 1, 2, …6 and we associate equal probabilities to the 1, 2, …6. This device enables us to use probability to describe outcomes of the process of the fair die roll. For example, we can now say that the probability of rolling a 1 is 1/6, or the probability that the upper face is odd is 1/2.

Many kinds of outcomes can be represented by random variables, for example, the outcome of a coin toss, or the number of heads thrown in five coin tosses, or a height measurement of a randomly selected person, or even the average of the height measurements of a group of 10 randomly selected persons.

Importantly, functions of a random variable can also be random variables, and functions of multiple random variables can also be random variables, so we can work with probability for both of these function types. Here, we will focus on random variables as described by their empirical probability distributions and by the summary measures of the empirical distributions which are calculated as functions of the data.

19.2.1.3 Distributions

So, the reader imagines, each of these random variables should be quite different—the outcome of the coin toss, the die roll, and the height measurement. How do we represent the differences? Without going too much into the weeds, each random variable is described by a *distribution*. The distribution dictates the relationship between all possible values that the random variable can take and the values of the probability line. The way that the distribution is expressed depends on the nature of the random variable. If the random variable is discrete, then the distribution is usually expressed as the probability of each of the possible outcomes, called the *probability mass function*. If the random variable is continuous, then the distribution is usually captured by the probability that the outcome will be less than or equal to any given value, called the *cumulative distribution function*. Later on, we will refer to quantities called *quantiles*, which can be computed from this function, and are the values of the distribution below which a given proportion of the distribution lies. For example, the 0.75 quantile is that point of the distribution that has 75% of the distribution below it.

As an example, the random variable that we use to represent the outcome of a fair coin toss is simply the allocation (mapping) of 0.5 probability to a head and 0.5 probability to a tail. Furthermore, the random variable that we use to represent X as the count of heads from five fair coin tosses is distributed according to a special device called the Binomial distribution. For 5 experiments each having probability of success 0.5 we write $X \sim Bi\ (5, 0.5)$. The random variable that we use to represent the average height of a random sample of ten adults could be captured by one of a number of different distributions, but commonly one of the *Normal* family, also called *Gaussian* family of distributions, is used.

These examples suggest that identifying the distribution of a random variable is more than just a matter of picking a name; indeed, the names identify families of distributions, which are themselves distinguished by quantities that are called *parameters*. We choose the distribution from the family of distributions by selecting the needed parameters. A substantial part of statistical theory is focused on the estimation of these parameters using functions of observed data. For example, the Normal distribution is indexed by its mean and variance, we might write $H \sim N(\mu_H, \sigma_H^2)$. The binomial distribution, mentioned above, is indexed by two parameters, namely, the number of observations n (considered fixed and known) and the probability of success (or failure) p. We shall refer to the combination of a random variable, distribution, and parameters as a *statistical model*. So, $H \sim N(\mu_H, \sigma_H^2)$ is a statistical model for the random variable H, and it says that H follows the Normal distribution with mean μ_H and variance σ_H^2. As shorthand, we shall commonly say that H is Normal or Normally distributed.

To sum up, statistics is concerned with modeling data, usually either for discovery or decision-making. We assume that these data come from a process, and that they can be represented by a random variable with a given distribution and appropriate parameters. Generally, we use a capital letter to denote a random variable, and the same letter in lower case to denote a sample of data taken from the process, $x = (x_1, \dots x_n)$.

19.2.2 Estimation

When we know the values of the parameters, we can use the model to make predictions or inference. But how do we know what the parameter values should be? That is, for example, given height measures from a random sample of ten adults, and assuming that we are willing to invoke the Normal distribution, how do we choose values for the mean and variance to best represent the data? This process is *estimation*.

19.2.2.1 Point Estimation

This challenge hints at a huge body of statistical work, but to choose two intuitive examples, the sample mean and the sample variance are almost always used as *estimators* for the process mean and the process variance. There is a lot of solid theory that suggests that this intuitively reasonable idea is a good idea most of the time.

Often we denote the estimator of a parameter by means of a hat: the estimator of the process mean μ (say) is then $\hat{\mu}$. And, commonly we estimate the process mean by the sample mean, which we often denote using a bar atop the random variable, so we might say that $\hat{\mu}_X = \bar{x}$; in words, the process mean for the random variable X is estimated by the sample mean of the data $x_i,\ i = 1 \ldots n$. We also say that the sample mean is an estimator for the process mean *in general*, and any given sample mean can be an estimate of the mean of the process from which the sample is taken.

19.2.2.2 Interval Estimation

At this point we have to introduce a complication, namely that the parameter estimates are functions of the sample data—for example, the mean and the variance—and since the data are represented by a random variable, then the parameter estimates should also be random variables, and they also have a distribution. So, the data have a distribution, and the parameter estimates that are computed from the data also have a distribution, and the distributions of the parameter estimates are different from the distribution of the data, although they may be related. For example, there is a beautiful device called the Central Limit Theorem that in its simplest form says that if the sample size is sufficiently large then the distribution of the parameter estimate for the population mean is Normal.

To capture this variability, Frequentist statistics provides a special kind of estimate as well as the point estimate introduced earlier—the confidence interval estimate. Confidence interval estimates are the range of numbers that are obtained with a given probability (e.g., 95%) under an infinite repetition of the algorithm of taking a fresh random sample from the same process and computing the estimates. The confidence interval provides valuable information about the uncertainty of the point estimate. The intended coverage of the interval must be stated, and commonly 95% confidence intervals are reported.

19.2.3 Models of Dependence

We now move to the statistical consideration of a specific type of model that allows the parameters of the distribution of a random variable to vary depending on the parameter values of other random variables. This development opens up a very rich class of models for different kinds of processes. A simple example is the linear model,

$$Y = \beta_0 + \beta_1 X + \varepsilon;\ \ \varepsilon \sim N\left(0, \sigma^2\right). \tag{19.1}$$

This model simply states the relationship between a random variable of interest Y (also called the *response variable* or the dependent variable) and another random variable X (called the *predictor variable*, sometimes also called the independent variable), for n observations that we have. It says that the fundamental relationship between X and Y is a straight line, with slope β_1 and y-intercept β_0, and there are some random fluctuations or errors from the straight line represented by the ε, which themselves follow the Normal distribution with mean 0 and variance σ^2.

The slope and intercept can be estimated from data (x, y) in a range of ways, for example, ordinary least squares, or maximum likelihood. It turns out that the parameter estimates for the βs are the same in either case but this is not always true. A great deal of attention has been paid to the different qualities that different kinds of parameter estimates have—how they work in the long run, how they work when there are plenty of data, and so on.

We introduce this particular model because we will test simulation models by comparing predictions that arise from the models against observations that are made of the processes or populations that the models are intended to represent. In the ideal case, the predictions and the observations will line up exactly—on the 1:1 line—so β_0 will be 0 and β_1 will be 1. However, because there is variability in observations, the predictions and the observations will not line up, and the values of β_0 and β_1 will not be exactly 0 and 1, respectively, even if the model is very good. This concern naturally leads to the question: how good is good enough? How close should β_0 and β_1 be to their designated values for us to be confident about the model? This question leads us to another statistical protocol: the hypothesis test.

19.2.4 Null Hypothesis Significance Tests

Having introduced the principles of random variables, distributions, parameters, and models, we can now articulate a process for asking questions about models, and most particularly model parameters, in the light of data. Many questions are possible; here we focus on the following general kind of question: given a process, a model that we believe represents the process, and observations considered representative of the outcome of the process, does it seem likely that the true, process values for the parameters in the model might be or might not be a specific value? These questions can be addressed by a device called a Null Hypothesis Significance Test (NHST).

More concretely, recall from the previous section that if we had a set of observations and a set of predictions from a model, then we might be interested in whether the observations and the predictions match, which can only happen if $\beta_0 = 0$ and $\beta_1 = 1$. We do not know what the true values are of these parameters, we only have estimates, calculated from data. So, given our estimators and their distributions, what are likely and unlikely values that the process-representing mode parameters β_0 and β_1 might take?

Hypothesis testing traditionally proceeds from identification of a so-called null hypothesis, which is supposed to represent the current state or the default state, and is the claim against which we wish to measure evidence. For example, given a representative sample x of a random variable X that represents a process, and a statistical model that says $X \sim N(\mu, \sigma^2)$, we might start with a null hypothesis that the process mean is truly 0; $H_0 : \mu_X = 0$, with the alternative hypothesis being $H_1 : \mu_X \neq 0$.

Briefly, we tackle this problem in the following way. We determine what the distribution of the parameter estimator would be if the null hypothesis were true. We then compute a test statistic from the data. This test statistic also has a known distribution if the null hypothesis is true (we also say "under the null hypothesis"), which is used to determine a *rejection region*, which is a set of values that the test statistic could take that are so unlikely—relative to the distribution under the null—that if we observed them we would become convinced that the null hypothesis cannot be true.

Imagine, for example, that if the null hypothesis were true then it would be reasonable to see values of the test statistic between -2 and 2, and the value we computed was 4. This value being so far away from the reasonable expected range creates doubt for us in the validity of the model for the test statistic, including the hypothesized parameter values. Of course, there is more to the model than simply the null hypothesis, as we shall see, but there are ways of assessing the importance of the other aspects of the model.

A more detailed description follows. We do not know μ_X but we have an estimator for it: the sample mean, $\hat{\mu}_X = \bar{x}$. However, as we noted earlier, the sample mean has a distribution because it is a function of the realizations x of the random variable X. It turns out that if the data are Normal then the sample mean is also Normal, but also as noted above, if the sample size is large enough, then we can treat the sample mean as though it were Normal, even if x is not Normal. Another beautiful device called the Law of Large Numbers tells us that we can assume that the mean of x is the same as the mean of X, and by algebra (not shown here), we can prove that the variance of $\bar{x}$ is just the variance of X (this variance is σ^2, the true population variance) divided by n, the sample size. So,

$$\bar{x} \sim N\left(\mu_X, \frac{\sigma^2}{n}\right). \tag{19.2}$$

Our goal is to test the null hypothesis that $\mu_X = 0$. Intuitively, if $\bar{x}$ is a long way from 0 then μ_X probably is not 0, whereas if $\bar{x}$ is close to 0 then μ_X might or might not be close to 0. The open question is: how far from 0 does $\bar{x}$ have to be before we cannot support the idea that μ_X is zero? The answer to this question lies in its distribution, or more completely, in its distribution assuming the null hypothesis is true. Under the null hypothesis, $\mu_X = 0$, and after some algebra, we arrive at the test statistic z and its distribution under the null hypothesis:

$$z = \frac{\bar{x}}{\sigma/\sqrt{n}} \sim N(0, 1). \tag{19.3}$$

Note that here we are assuming that we know σ, which is a very unlikely state of affairs. We will revisit this assumption. Also, the quantity $\sigma/\sqrt{n}$ is called the standard error.

The hypothesis test simply proceeds by evaluating z and comparing it with nominated *quantiles* of the standard Normal distribution $N(0, 1)$, e.g., ± 1.96, corresponding to the 95% interval delimited by the 2.5 and 97.5% quantiles. The quantiles create the rejection region, which comprises all points outside the interval described by the quantiles. We obtain the quantiles by referring to the cumulative distribution function of the distribution.

So, if z is between the quantiles, then we do not reject the null hypothesis. If z is outside the quantiles, then we reject the null hypothesis. Sometimes we refer to the test result as being *significant*, or, preferably, *statistically significant*. This phrase is shorthand for saying that the test statistic is within the rejection region of the test.

Comfortingly, the rejection region is closely related to the complement of the corresponding confidence interval estimates, although the rejection region surrounds the hypothesized value and the confidence interval surrounds the estimate.

Now to clean up some loose ends. We specifically choose the quantiles to deliver a chosen probability of rejecting the null hypothesis when in fact it is true, that is, when $\mu_X = 0$. This probability is called the *size* of the test and is generally chosen to be small. Such an error—namely, the mistaken rejection of a true null hypothesis—is sometimes called a *Type 1 error*. So, we choose the quantiles that set the rejection region in order to control the probability of a Type 1 error, and that probability is called the size of the test.

It is also possible that the test statistic could be *outside* the rejection region even though $\mu_X \neq 0$, in which case we would mistakenly fail to reject the null hypothesis. This kind of error is called a *Type 2 error*. The probability of that happening depends on μ_X and can be computed from the *power function* of the test. The power function reports the probability of rejecting the null hypothesis as a function of the true, unknown value of the parameter, assuming that the model is otherwise true. For example, the power function is equal to the test size at $\mu_X = 0$. We can also compute the probability that the test statistic will be in the rejection region at other values of μ_X, conditional on the model.

The astute reader will have noticed that calculating z involves the sample size n. In fact, it turns out that z increases as n increases (specifically, z increases proportional to the square root of n). Consequently, any given hypothesis can be rejected if the sample size is large enough.

The rejection region is set based on an assumed distribution for the test statistic assuming that the null hypothesis is true. The test statistic may be in the rejection region because the chosen distribution, which the region is calculated from, is wrong. Generally, careful statisticians will use graphical displays of the data to guard against an inappropriate choice of distribution; these are particularly common in model construction. Note that the statistical characteristics of the estimators are derived from the assumed characteristics of the process from which the sample is drawn, not from the characteristics of the sample itself—although those should be indicative of the process in any case. So, we examine graphical diagnostics of the sample in the hope that they provide information about the process.

Up until now, we have been focusing on rejection regions that are symmetric. Other scenarios may be of interest. Sometimes, all of the inferential interest is in detecting the possibility that the mean is above the hypothesized value (or, equivalently, below). These cases lead to so-called one-sided tests, in which the rejection region is only the interval above (below) a given quantile. The rejection region is then the complement of the one-sided confidence intervals, which we will use to construct simple equivalence tests below.

Finally, we assumed that we knew the population variance, which is almost never true. In practice, we tend to estimate the population variance using the sample variance, which of course is itself subject to variation! There is a simple remedy: we compute the rejection region using quantiles of the so-called Student's t-distribution instead of the standard Normal distribution because Student's t allows for uncertainty in the estimator of the population variance.

At this point, the breathless reader should have a grasp of the basic vocabulary and principles that are used in Frequentist-style estimation and testing. There are many other kinds of tests under this broad umbrella, and many more in different traditions. Now we move to the question of how to apply this framework for validating models.

19.3 Statistical Model Validation: Why and How?

Briefly, we shall say that models are validated when they pass a systematic statistical assessment using appropriate tests, to be discussed, with respect to a dataset other than that with which they have been constructed. As we shall see, there is a distinction between a useful model and a validated model. We now look at the reasons that models are validated, and the implications that these reasons have for the activities that ensue.

19.3.1 Why Validate?

Frequentist model validation usually occurs for one of two reasons. First, a user may wish to validate a model for use in a particular setting, perhaps a different spatial location or extent, or a different time period, to that in which it was constructed. For example, a regional model that predicts tree height from tree bole diameter has been constructed for a number of different species in northern and central Idaho forests. Such models are constructed over a wide spatial extent using data gathered at a particular time. We may be interested whether the model holds for a particular forest within the area, or even outside the area, at some point in time long after the model was constructed. Second, a user may wish to validate the model to assess whether it shows patterns that would be expected from theory. For example, co-authors and I constructed a semi-mechanistic model of tree growth processes called 4-PG (see, e.g., Duursma et al. 2007). It was of interest to try to validate the predictions of forest growth using measures of various physical dimensions of the forest. A systematic comparison of independent observed measures against predictions is a gold standard in modeling. As a further example, Capes et al. (2017) used equivalence tests to try to validate an allometric model using field data.

Several activities that are components of model validation are necessary but not sufficient. First, we may report statistics that summarize the quality of model fit, such as R^2 or root mean squared error (see below), computed from the dataset used to fit the model. This is model checking, it comprises an important set of steps undertaken to help form an opinion of the validity and utility of the model, but it does not count as validation in isolation. Second, we may split the available data into complementary *training* and *testing* sets, and compute testing-based fit statistics of a model that was fit using the training set. This is also model checking, and it provides a useful insight into whether portions of the data differ systematically from one another. It still does not qualify as model validation in isolation.

We now consider some popular measures and tests of goodness of fit, and assess their utility with regards to model validation.

19.3.2 Estimating Goodness of Fit

Although they are not commonly described as estimators of goodness of fit, a number of the popular regression summary statistics are routinely interpreted in this useful way. That is, the statistic is routinely reported as evidence of the goodness of fit but the statistic is rarely compared formally in the sense of a statistical test. An example of such a statistic is the root mean squared error, *RMSE*.

$$RMSE = \sqrt{\sum_{i=1}^{n}(x_{p(i)} - x_{m(i)})^2} \tag{19.4}$$

where $x_{p(i)}$ and $x_{m(i)}$ are the i-th values of the process observation and the model prediction, respectively.

This quantity reports the square root of the average squared distance between the observations and predictions, and is coded in the units of the response variable. We can interpret the *RMSE* of the model as a measure of the residual uncertainty, that is, the uncertainty about y conditional on the model. We can informally compare it with the standard deviation of y; believing in the model and knowing the predictor variables relieves us of uncertainty about the value of y to the extent that the *RMSE* is smaller than the standard deviation of y.

A popular alternative is Pearson's correlation, which can be computed between two random variables. The correlation takes a value within $[-1, 1]$ and is 0 when there is no detectable linear match between the random variables, 1 when there is an exact linear match (regardless of the slope, so long as it is positive), and -1 when there is an exact linear match with some negative slope.

A related estimator of the goodness of fit of a model that is routinely reported by regression software is the R^2 statistic, which spans $[0, 1]$. In linear models, R^2 reports the proportion of the variance in the response variable that is matched by variation in the predictor variables. As a side-note, R^2 is known to increase with the number of predictor variables, so often the *adjusted* R^2 is reported, which is the R^2 penalized by the number of predictor variables.

Each of these statistics has been reported, from time to time, as measures of the validity of a model, however, these summary statistics are useful but not sufficient. In each case, they report some measure of the closeness of the predictions to the observations, but they do not provide any evidence of the closeness of the slope to 1 or the intercept to 0, which are the conditions mentioned in Sect. 19.2.3 for the model to be validated. Therefore it is possible that a model is useful, in that it returns summary statistics that show a substantial reduction in uncertainty about y, but not validated.

19.3.3 Testing Goodness of Fit

We now move to the challenge of testing goodness of fit. Not surprisingly, a number of such statistical tools have been applied to validation problems. For example, Freese (1960) introduced an accuracy test based on the standard χ^2 test, subsequently extended by Reynolds (1984) and Gregoire and Reynolds (1988). Kleijnen (1974) mentions common techniques such as χ^2 tests, Kolmogorov–Smirnov tests, and regression analysis of the actual and simulation output, and comparing the output using parametric or nonparametric statistical tests. Kleijnen et al. (1998) proposed that the means and variances of the model predictions and the process measures be compared by NHST, and the correlation between them tested with the one-sided test, $H_0 : \rho \leq 0$ and $H_A : \rho > 0$ (hereafter, MVS).

Considering applications of these tests, Ottosson and Håkanson (1997) used R^2 and compared with so-called highest-possible R^2, which are predictions from common units (parallel time-compatible sets). Jans-Hammermeister and McGill (1997) used an F-statistic based lack of fit test. Landsberg et al. (2003) used R^2 and relative mean bias. Bartelink (1998) graphed field data and predictions with confidence intervals. Finally, Alewell and Manderscheid (1998) used R^2 and normalized mean absolute error ($NMAE$).

We examine a few of these tests in greater detail. The traditional approach to assessing goodness of fit is Pearson's χ^2 goodness of fit test, which is taught in most introductory statistics classes. In this test, we begin with categorical data (counts in categories) and a model; for example, we may have the outcome of 10,000 die tosses arising from a rainy weekend, and as a model that the die is fair: there should be an equal split between the occurrences of the outcomes 1–6. The null hypothesis is that the model matches the data, and we reject the null hypothesis if the test statistic is too unusual relative to a specific member of the χ^2 family of probability distributions. If we fail to reject the null hypothesis, then we treat the model as a defensible representation of the underlying process. This test is best suited to categorical data, which are not the focus of this chapter.

The primary alternatives that are commonly applied in testing goodness of fit for model validation are (i) an NHST with the null hypothesis being that the process mean is equal to the prediction mean, and (ii) two NHSTs, one on the slope and one on the intercept parameters of the linear model (Eq. 19.1) fit to paired data, with the null hypotheses being that the true intercept $\beta_0 = 0$ and the true slope $\beta_1 = 1$, as suggested by Cohen and Cyert (1961). A variation on the latter is called a *whole-model test* (WMT), which uses the F distribution to test the joint null hypothesis that the true intercept $\beta_0 = 0$ and the true slope is $\beta_1 = 1$ Kleijnen (1995).

As above, outcomes of these tests have been reported, from time to time, as measures of the validity of a model, and they are useful but not sufficient. As (Kleijnen 1995) notes for these and related tests, (i) the bigger the sample size is, the smaller the critical value, so all else being equal a model is more likely to fail to be rejected by a small sample, in short, the fewer data you have the better your chances of acceptance are; and (ii) the test statistic may be statistically significant and yet unimportant, and of course if the sample is very large, then the test statistic is nearly always statistically significant.

The traditional application of hypothesis tests has been shown to be inappropriate for model validation (see, e.g. Mayer and Butler (1993); Kleijnen (1995); Loehle (1997)). This is because, borrowing the parlance from Robinson et al. (2005), the tests are designed to *split*, instead of to *lump*.

19.3.4 Tests for Splitting and Tests for Lumping

The whimsical title for this section sets the scene for thinking about the structure of statistical testing. Most statistical tests are predicated on a null hypothesis of no effect, or no difference between means (for example), and they evaluate the evidence against that null. If the evidence is equivocal, then the null hypothesis is not rejected, and we claim, or declare, or act as though there is no effect or no difference between means. The purpose of such a test is to detect a distinction: to split.

However, what if the inferentially interesting question were instead to prove that some parameters were startlingly similar? Then the traditional splitting test is less interesting: its starting point is the condition for which we wish to assess the evidence, rather than being that against which we assess evidence. The type of test we need is a test of lumping, namely an *equivalence* test. Briefly, an equivalence test is a statistical test that has as its null hypothesis the claim that a parameter does not equal a target value (e.g., 0) and as the alternative hypothesis, the claim that the parameter does equal the target value. So, equivalence tests simply swap the null and alternative hypotheses relative to standard NHST, but still using the same familiar statistical tools.

There is a substantial literature on equivalence testing, we note particularly Berger and Hsu (1996), McBride (1999), Parkhurst (2001), Wellek (2010), and Meyners (2012).

19.3.4.1 What is the Goal: Goodness of Fit or Fitness for Purpose?

Although goodness of fit testing as reviewed above is a popular statistical technique and in common use, it does not satisfy the requirements of model validation. This is for two important reasons. First, as noted, the typical NHST setup takes as the null hypothesis the condition that it wishes to disprove, as opposed to the condition that it wishes to prove. This means that under-powered tests can over-hastily declare a model to be validated by failing to detect a statistically significant difference. Second, the typical NHST makes no allowance for the requirements to be placed upon the model. We detail the latter issue in this section.

The two examples outlined briefly introduced in Sect. 19.3.1 share a common purpose: in each case the goal was to assess whether the model performed to some expected level. This observation leads to a key question: on assessing model performance, how good is good enough? This question encourages us to think about model performance from the point of view of *fitness for purpose* as opposed to its *goodness of fit*. In order to assess fitness for purpose, the tester must assert a benchmark, ideally expressed in units that link to the model application or interpretation (see also Chap. 18 by Saam in this volume). In the case of the height–diameter model noted above, for example, we might say that the benchmark is that the average of the height predictions be within 1 m of the observations. Thinking of the validation of the mechanistic model, we may assert that the benchmark could be $\pm 10\,\mathrm{m}^3\mathrm{ha}^{-1}\mathrm{decade}^{-1}$.

In each case we need to acknowledge that the average differences are random variables, so to apply a benchmark we need to include a statement of probability. We might say, for instance, that we wish the average difference to be within a specific region such as ± 1 m with a given confidence, say 80 or 90%, depending on the use to which the model will be put.

We will label this benchmark the *region of equivalence*. The requirement of a region of equivalence is both a strength and a weakness of this approach. It is a strength because it is the means by which the model validity can be connected to the application or decision context. This means that equivalence testing supports the specification of how accurate the model needs to be in terms of the variable that is being measured. It is also a weakness, however, because it introduces another (possibly) arbitrary aspect to the specification of the test, in addition to the specification of the test size. Setting the region of equivalence should involve careful consideration of how accurate the user needs the model to be.

We recommend the use of equivalence tests for Frequentist-based model validation (Robinson and Froese 2004). We next introduce a few of the key Frequentist tests of equivalence and discuss their application to model validation.

19.3.5 Conceptual Entry Point: TOST

The Two One-Sided Test (*TOST*) algorithm is a simple extension of the traditional NHST, predicated on a different kind of null hypothesis. The null hypothesis in this setting is that the processes that produce the model outputs and the observations have importantly different means, where the magnitude of the important difference is captured using the region of equivalence. The alternative hypothesis is that the means of the processes for the model outputs and the observations are not importantly different.

We provide a verbal description of the *TOST* and then an algorithm. To apply this algorithm, we construct two one-sided rejection regions, each one for tests with size 2α, and ask whether the test statistic occupies both rejection regions simultaneously. If we reject the null hypothesis that the tested parameter is further below the specified value than the benchmark *and* we reject the null hypothesis that the tested parameter is further above the specified value than the benchmark, then we are forced to reject the joint hypothesis that the tested parameter is further away from the specific value than the benchmark. We can interpret this result as being evidence that the tested parameter is significantly close to the target, having declared our equivalence interval and the size of our tests.

The following algorithm assumes that we have n pairs of observations: those that represent the target process, x_p, and those that represent the output of the model, x_m, that is intended to represent the process. If we could prove that $x_p = x_m$ then the model would be validated. The following test is a useful approximation to such a proof exercise.

1. Compute the n values of x_d, which is the difference between the process and the model for each observation, $x_d = x_p - x_m$.
2. Choose a test statistic, e.g., the mean $\bar{x}_d$, and a size for the test, e.g., α. This statistic will be used as follows: if the mean of the difference x_d is demonstrably close to zero, then by the following steps, we can consider the model to be validated.
3. Choose the region of equivalence, $\mathscr{I}$, which is the region that is close enough to the hypothesized value (0) that the difference is practically irrelevant. This

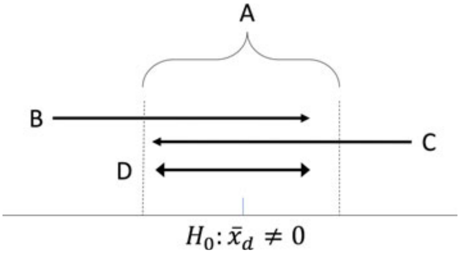


Fig. 19.1 Schematic example of *TOST*. See text for details

interval is analogous to the tolerance of an engineered part, plays the role of the benchmark, and is expressed in the units of x. This is region A in Fig. 19.1.

4. Compute the mean of the difference x_d, and a lower and an upper one-sided $1 - \alpha$ confidence interval (arrows B and C in Fig. 19.1, respectively), the upper and lower limits of which we denote with C_α^- (arrowhead B) and C_α^+ (arrowhead C), respectively. Note that the intervals each have coverage $1 - \alpha$, rather than $1 - \alpha/2$.
5. Form an interval (C_α^+, C_α^-) from the intersection of the two one-sided confidence intervals around the mean difference (arrow D in in Fig. 19.1). This interval should be reported along with the test outcome.
6. Reject H_0, the null hypothesis of dissimilarity, if the interval D is entirely contained within the interval of equivalence A, i.e., $(C_\alpha^+, C_\alpha^-) \subseteq \mathscr{I}$ (this condition is satisfied in Fig. 19.1).
7. If the null hypothesis of dissimilarity is rejected then the model is validated (this is true in Fig. 19.1). If the null hypothesis is not rejected then the model is not validated.

A substantial advantage of the *TOST* is its flexibility; it can be computed from a very wide range of experimental or survey setups. *TOST*s do not require paired data, because the test statistics can be computed directly from means and standard deviations of the observations and predictions. However, a *TOST* computed on the differences of paired data will be more powerful than a *TOST* computed from the summary statistics of the individual variables. Nonetheless, a *TOST* can be computed for any statistic for which a confidence interval can be computed.

A curiosity can occur in that it is possible that the intersection interval D is within the interval of equivalence *and* the usual confidence interval for the mean of the differences does not cover 0. Then, we would reject the null hypothesis that the difference is 0 and also reject the null hypothesis that the difference is not 0. What can we do then? Of course these outcomes are not exactly mutually contradictory, and we could say that the model is fit for purpose but not perfect.

Sargent (2012) and the literature cited within developed a comparable approach to model validation. However, the authors cautioned that the test requires the assumption that the underlying population be Normal, whereas the *TOST* only requires that the sampling distribution of the test statistic be Normal, which may be justified by invocation of the Central Limit Theorem if the sample size is large enough. A cautious statistician can approximate the sampling distribution of the test statistic using a bootstrap, but this detail is beyond the scope of this chapter.

19.3.6 A Uniformly Most Powerful Invariant Test

The *TOST* is useful and flexible but carries no warranty of being the most powerful test. Wellek (2010) presented an equivalence test that is a uniformly most powerful invariant (*UMPI*) test. UMPI tests make at least as good use of the available data as any other test in their class, so in this sense they are the best test for the job. The test is, fittingly, labeled the paired t-test for equivalence (*PTTE*). However, *PTTE* only works for paired observations and for a region of equivalence that is scaled by the population standard deviation, and only enjoys its *UMPI* status when used upon intraindividual distances that are drawn from the Normal distribution.

The algorithm for the test is as follows. We calculate the following quantities for the differences x_d:

- mean ($\bar{x}_d$),
- standard deviation (s_d) and from this,
- the standard error ($s_{\bar{x}_d} = \frac{s_d}{\sqrt{n}}$), which is a measure of the variation of the sampling distribution of the mean.

Here, n is the number of data pairs. We then calculate the t-value corresponding to the observed mean and its standard error by

$$t_d = \frac{\bar{x}_d}{s_{\bar{x}_d}} \tag{19.5}$$

We will compare the absolute value (positive portion) of this value with a cutoff, which is computed as follows. We calculate the noncentrality parameter, $\psi^2 = n \times \varepsilon^2$, where ε is the half-length of the region of equivalence $\mathscr{I}$. Then the cutoff $\tilde{C}_{\alpha;n-1}(\varepsilon)$ that corresponds to a test of size α is the α-quantile of the noncentral F distribution with degrees of freedom $\nu_1 = 1$ and $\nu_2 = n - 1$, and noncentrality parameter ψ^2, as calculated above.

If the absolute value of the t-value is lower than the cutoff then we reject the null hypothesis of dissimilarity, and the model is validated.

19.3.7 More Descriptive: Test of Fidelity

The following material is more technical, and requires an understanding of how to fit simple statistical models such as presented in Eq. (19.1).

Getting the mean right and getting the predictions right are different. The *TOST* and *PTTE* permit validation based on the means, but this does not reassure us that the individual predictions and observations line up, or even are in the right order. Robinson et al. (2005) proposed a combination of *TOST*s that provide a more nuanced test for assessing the goodness of fit of a model given paired observations and predictions. This approach is essentially an equivalence-based version of the test proposed by Cohen and Cyert (1961).

In essence, we perform a *TOST* on each of the intercept and slope estimates, and whereas the usual application of *TOST* tests for population-level agreement on average, our proposed strategy will test for point-to-point agreement as well.

1. Make observations x_p, and calculate predictions from the model x_m.
2. Subtract the mean prediction from the predictions. (We do this so that the estimators of intercept and the slope (conditional on the intercept) are guaranteed to be independent, because the fitted line passes through the (0, 0) point. The slope estimate does not change, and if the model predictions are unbiased, then the intercept will be the same as the mean of the observations.)
3. Establish regions of equivalence: $\mathscr{I}_0$ for the shifted intercept and $\mathscr{I}_1$ for the slope, e.g., (a) $\mathscr{I}_0 = \bar{x}_m \pm 1\,\text{m}$ for the shifted intercept of our tree height model and data, and (b) $\mathscr{I}_1 = 1.0 \pm 0.2$ for the slope. As above, these quantities play the role of a benchmark.
4. Fit a linear regression using the model predictions as the sole predictor variable and the observations as the response variable.
5. Test the intercept for equality to $\bar{x}_m$. Calculate the two one-sided confidence intervals for the intercept using, e.g., the estimate of the standard error of the intercept from the regression output, and determine whether this interval is contained inside the intercept region of equivalence. This is identical to testing that the mean of the observations is equivalent to the mean of the predictions.
6. Test the slope for equality to 1. Calculate the two one-sided confidence intervals for the slope from, e.g., the estimate of the standard error of the slope from the regression output, and determine whether this interval is contained inside the slope region of equivalence.
7. If the null hypothesis that the intercept is different to 0 and the null hypothesis that the slope is different to 1 are rejected, then the model is validated.

The R code to compute and plot these quantities is freely available in the equivalence package (Robinson 2016). The outcome can be represented graphically, as we do below.

19.3.8 Statistical Validation Overview

We conclude this section with provides an overview table (Table 19.1) that summarizes the key information about the estimators and tests of goodness of fit presented in this chapter.

Table 19.1 Summary table of Frequentist statistical tools for model validation. x_p represents the process and x_m the model. Here, "Test is backwards" is used to indicate that the null hypothesis is that the model is valid. Regression modeling is with the model predictions as the predictor variable and the process observations as the response variable. *CLT* stands for the Central Limit Theorem

Name	Notes	Strengths	Weaknesses
Correlation	The correlation of x_m and x_p. High correlation implies a valid model	Familiar	Ignores intercept and slope
RMSE	The square root of the mean of $(x_m - x_p)^2$. Low *RMSE* implies a valid model	Units are the same as x_p	Tester has to define "low"
χ^2 test	The sum of $(x_m - x_p)^2/x_p$ has a known distribution when the model is valid. Low χ^2 implies a valid model	Formally tests validation	Test is backwards
T-test	The t-statistic has a known distribution when the model is valid. Low t implies a valid model	Formally tests validation	Test is backwards
MVS	Compare the means of the model and the process outputs and variances of the model and the process outputs using standard tests and test the null hypothesis that the regression slope is nonpositive. Similarity of means, similarity of variances, *and* nonnegative slope imply a valid model	Formally tests validation. Covers location, scale, and relationship	Test is backwards and Normality is assumed
Regression	Apply standard t-tests to the intercept ($H_0 : \beta_0 = 0$) and the slope ($H_0 : \beta_1 = 1$). Low t statistics imply a valid model	Formally tests validation. Tests the mean and point-to-point agreement	Test is backwards. Normality is assumed or *CLT* is invoked
WMT	Apply standard whole-model style F-test to the model ($H_0 : \beta_0 = 0; \beta_1 = 1$). Low F-statistic implies a valid model	Formally tests validation. Tests the mean and point-to-point agreement	Test is backwards. Normality is assumed or *CLT* is invoked
TOST	Prove that the mean of x_m is at the same time not too far above and not too far below the mean of x_p	Formally tests validation	Tester has to define "not too far". Normality is assumed or *CLT* is invoked
PTTE	Prove that the mean of x_m is at the same time not too far above and not too far below the mean of x_p	Formally tests validation. More powerful than *TOST* but tests different types of limits	Tester has to define "not too far". Normality is assumed
Fidelity	Apply standard *TOST* to the intercept ($H_0 : \beta_0 \neq 0$) and the slope ($H_0 : \beta_1 \neq 1$). High t statistics imply a valid model	Formally tests validation. Tests the mean and point-to-point agreement	Tester has to define "not too far". Normality is assumed or *CLT* is invoked

19.4 Examples

This section provides illustrative examples of the estimators and tests that are outlined in the previous sections. We show examples that exemplify the two primary applications of equivalence tests for model validation, namely the assessment of fitness of purpose, and the assessment of model fidelity, in the sense of how well the model matches the process(es) that it is designed to represent. The R code for all the tests reported here is available from the author. The equivalence tests were performed using functions provided in the equivalence package (Robinson 2016).

19.4.1 Fitness for Purpose

Consider the following hypothetical scenario. A forest manager is interested in using a specific model of tree shape, specifically relating the height of the tree to the cross-sectional diameter as measured at 1.38 m from the ground. Invocation of this model will relieve the manager from the necessity of measuring the heights of the trees in the forest, if the model can be relied upon. The manager measures the diameter and height of a random sample of trees from a particular stand within the forest.

These data are plotted as Fig. 19.2 and available as the “ufc” dataset (Upper Flat Creek, University of Idaho Experimental Forest) in the equivalence package (Robinson 2016) for the open-source statistical environment R Core Team (2017). The model being tested is an empirical multiple regression-style model as documented in Wykoff et al. (1982). Two outliers are ignored for the balance of the analysis; these were dead trees that were measured in error and should not be included in live-tree data for fitting a model.

We can now apply some of the estimators and tests that have been commonly applied in model validation. The correlation between the predictions and observations is 0.84. If we try to use a linear model to link the observations and the predictions, then the point estimate of the intercept is about 1.1 m (95% confidence interval: −0.50, 2.7) and of the slope is about 0.94 (95% confidence interval: 0.87, 1.00). The R^2 value for the regression is 0.70 and the root mean squared error is 4.0 m.

We now use these statistics to apply some traditional model validation tests. For example, the confidence interval for the intercept includes 0 and the confidence interval for the slope includes 1, so we would consider the model to be validated by regression-style tests from Cohen and Cyert (1961). Similarly, the test for positive correlation advanced by Kleijnen (1995) would conclude that the model is valid because the slope is clearly greater than zero.

Next, we apply the various tests of equivalence. The manager will be satisfied with the model if the mean of the predictions is within a meter of the mean of the observations with 95% confidence. The two one-sided confidence intervals for the difference, that is the predictions subtracted from the observations, is (−0.76, −0.07), which is contained completely within ±1 m. According to the *TOST*, the model is valid.

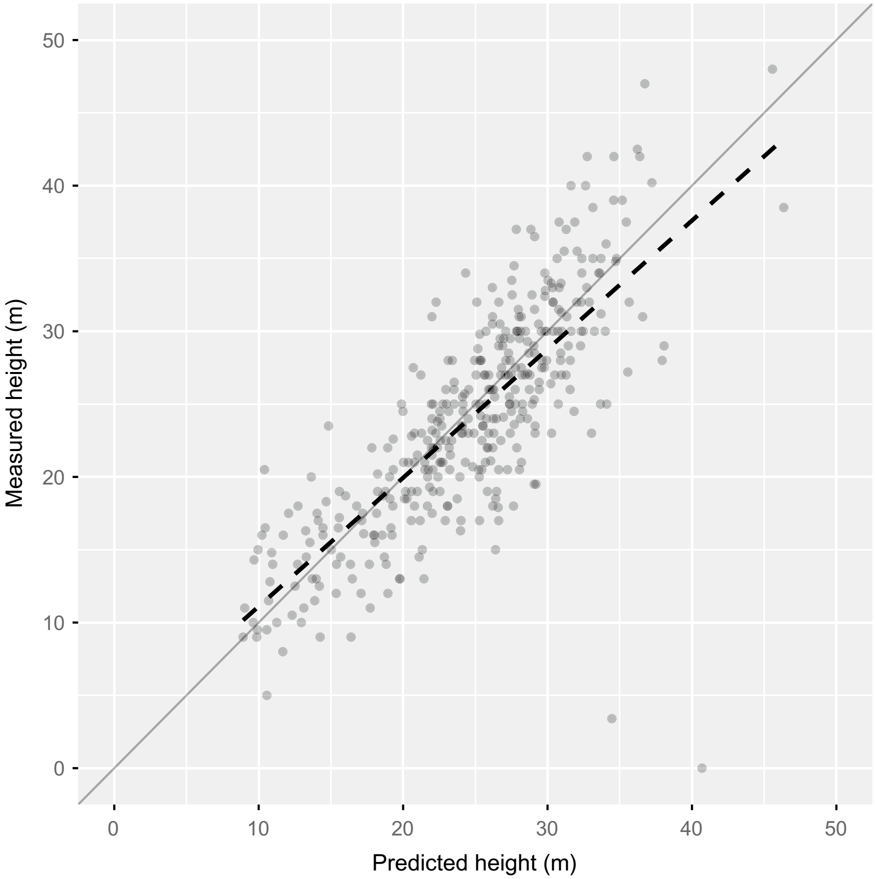


Fig. 19.2 Observed (y-axis) and predicted (x-axis) tree heights for the Upper Flat Creek forest stand, University of Idaho Experimental Forest. The solid line is the 1:1 line and the dashed line shows the linear regression model of the measured height as the response variable and the predicted height as its sole predictor. Two outliers are obvious

This conclusion is conditional on the model assumption that the sampling distribution of the mean of the individual differences is Normal. A resampled estimate of the distribution provided in the left-hand panel of Fig. 19.3 suggests that the assumption is reasonable; the points are close to the straight line that represents the Normal distribution.

We also validate the model using Wellek's *PTTE* as coded in the equivalence package of Robinson (2016). To do this we need a value for ε, which describes the size of the interval of equivalence; Wellek recommends 0.25 as a stringent value, so we adopt that here. The cutoff is 3.187, and the t statistic is -1.97. The absolute value of this statistic is 1.97, which is less than the cutoff 3.187, so the model is validated according to Wellek's stringent test.

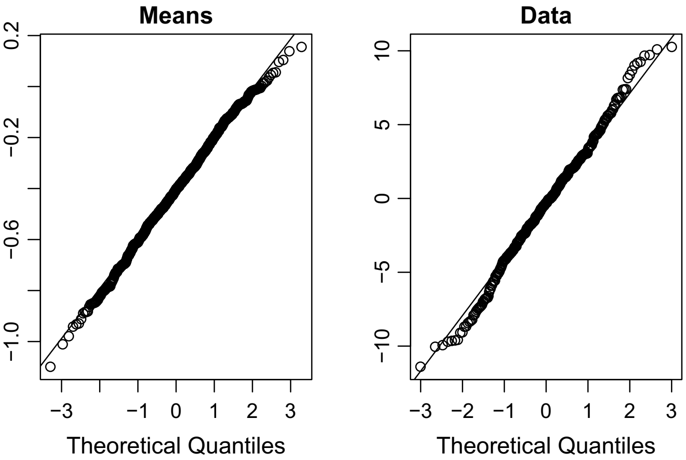


Fig. 19.3 Normal quantile plots for (i) the repeated sample of the mean of the differences (LHS) and (ii) the differences (RHS) of the measured UFC tree heights and predictions. The straight line represents the Normal distribution

This conclusion is conditional on the model assumption that the distribution of the individual differences is Normal. A resampled estimate of the distribution provided in the right-hand panel of Fig. 19.3 suggests that the assumption is reasonable; the points are fairly close to the straight line that represents the Normal distribution.

Now we demonstrate the test of fidelity. Assume that the manager will be satisfied with the model if the mean of the predictions is within 1 metre of the mean of the observations with 95% confidence *and* the slope of the regression line is within 20% of 1 with 95% confidence.

A graphical representation of the regression-based test of Robinson et al. (2005) is presented in Fig. 19.4. The figure is interpreted as follows. The points are the observations as before. The solid black line is the 1:1 line, provided for reference. Two error bars intersect the solid black line at the mean of the predictions (x-axis). The shaded rectangle shows the region of equivalence around the intercept: if the narrower error bar is within the gray rectangle then reject the null hypothesis that the intercept is dissimilar to zero; in the figure, the narrower error bar is inside the rectangle. The dashed lines show the equivalence region around the 1:1 line: if the wider error bar is within the dashed gray lines then we reject the null hypothesis that the slope is dissimilar to 1; in the figure the wider error bar is within the dashed lines. This figure shows that we reject the null hypothesis of dissimilarity at $\alpha = 0.05$ for the intercept at $\mathscr{I}_0 = \pm 1$ m and for the slope at $\mathscr{I}_1 = \pm 0.2$. According to the test of fidelity, the model is valid.

Based on the equivalence tests we believe that the manager can use the height–diameter prediction model with reasonable confidence.

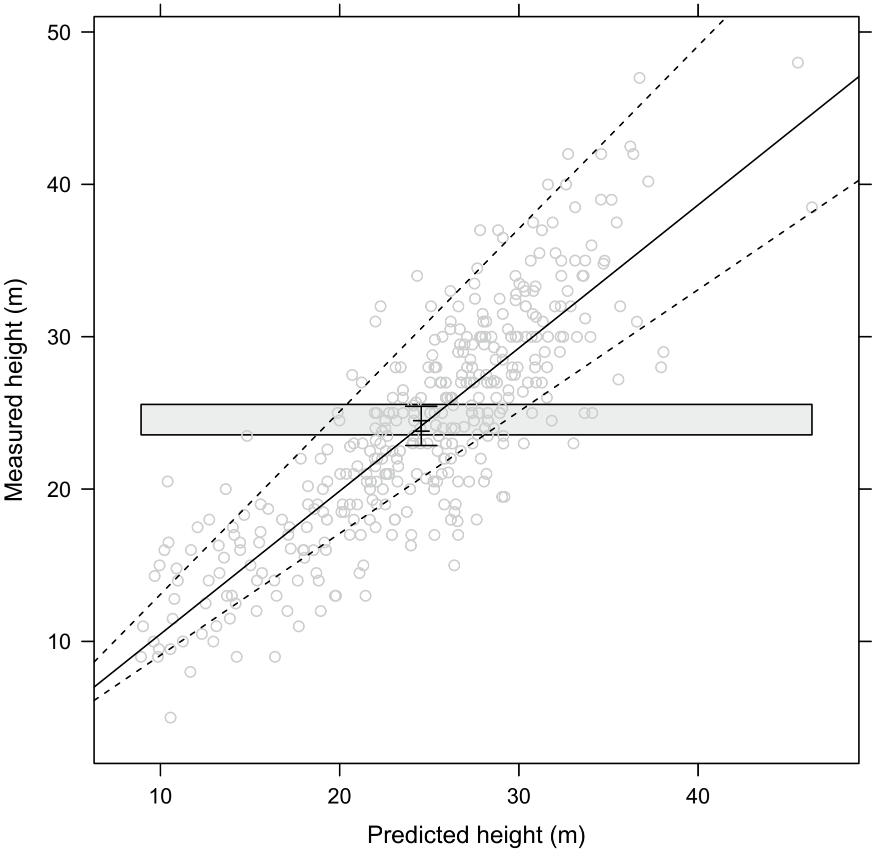


Fig. 19.4 Observed (y-axis) and predicted (x-axis) tree heights for the Upper Flat Creek forest stand, University of Idaho Experimental Forest, with regression-based *TOST* super-imposed. Symbols are interpreted in the text

19.4.2 Validation of a Theoretical Model

In this second case study, we seek to test a process-based model of forest growth, 4-PG (Duursma et al. 2007), using measured forest growth data obtained from field measurements taken in the Priest River Experimental Forest (PREF, for short) in northern Idaho, USA. The model was designed to be applied to mixed-species forests. The PREF is a complex mixture of 12 conifer species, with altitude gradient 1000 m (700–1700 m), mostly west-facing on steep slopes. Three species comprise 75% of the basal area: *Thuja plicata*, *Tsuga heterophylla*, and *Pseudotsuga menziesii*. 35 field plots were located at random across the PREF, stratified by altitude and solar insolation (Pocewicz et al. 2004), to obtain an unbiased sample of the forest structures present within the PREF. Tree volume growth was estimated based on computations

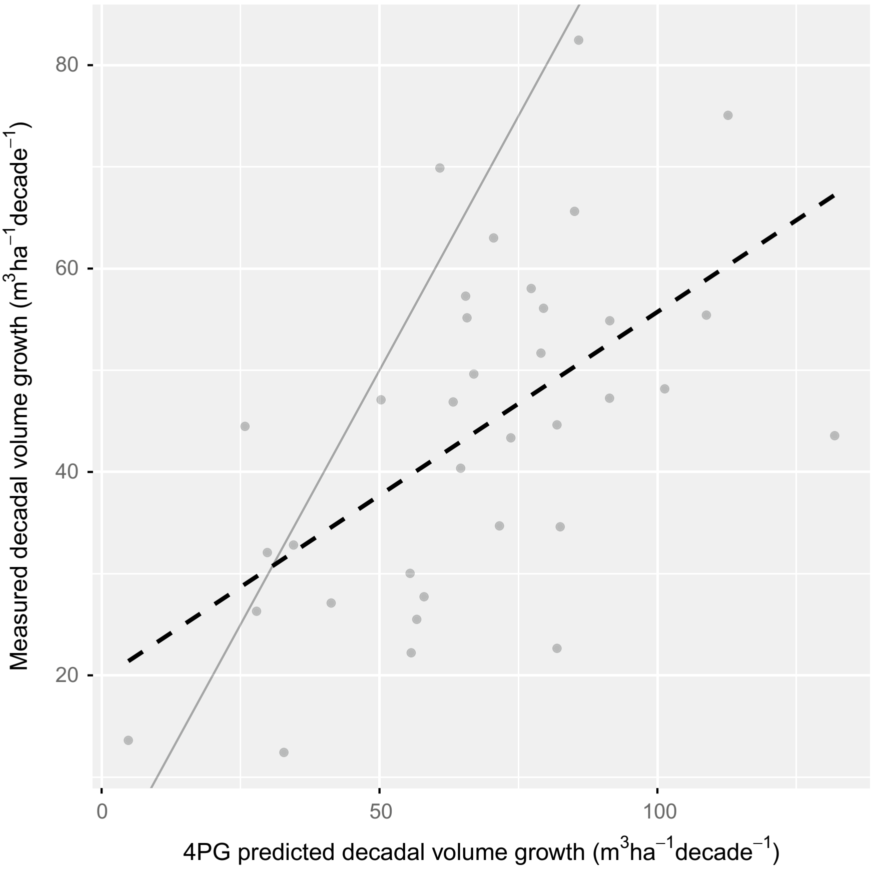


Fig. 19.5 Observed (y-axis) and predicted (x-axis) volume growth amounts for the Priest River Experimental Forest, ID. The solid line is is the 1:1 line and the dashed line shows the linear regression model of the measured height as the response variable and the predicted height as its sole predictor

carried out upon 10-year increment cores. A scatterplot of the predicted volume growth against the observed volume growth is provided in Fig. 19.5.

The correlation between the predictions and observations is 0.57. If we try to use a linear model to link the observations and the predictions, then the point estimate of the intercept is about 19.7 $m^3ha^{-1}decade^{-1}$ (95% confidence interval: 6.4, 32.9) and of the slope is about 0.36 (95% confidence interval: 0.18, 0.54) The R^2 value for the regression is 0.33 and the root mean squared error is 14.0 $m^3ha^{-1}decade^{-1}$.

As above we can use these statistics to apply traditional model validation tests. For example, the confidence interval for the intercept does not include 0 and the confidence interval for the slope does not include 1, so we would consider the model to not be validated by regression-style tests from Cohen and Cyert (1961). However,

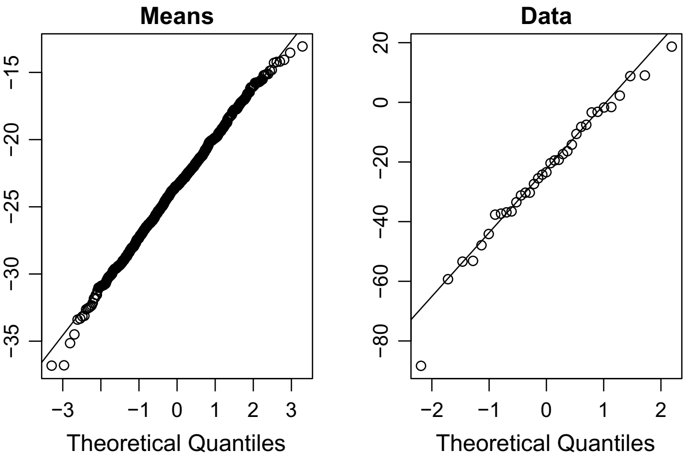


Fig. 19.6 Normal quantile plots for (i) the repeated sample of the mean of the differences (LHS) and (ii) the differences (RHS) of the measured PREF forest volume growth and predictions. The straight line represents the Normal distribution

the test for positive correlation advanced by Kleijnen (1995) would conclude that the model is valid because the slope is clearly greater than zero.

Next, we apply the various tests of equivalence. We will be satisfied with the model if the mean of the predictions is within 10 $m^3ha^{-1}decade^{-1}$ of the mean of the observations with 95% confidence. The two one-sided confidence intervals for the difference, that is the predictions subtracted from the observations, is (−29.9, −17.2), which is not contained completely within ±10 $m^3ha^{-1}decade^{-1}$. Under TOST, the model is rejected as invalid.

This conclusion is conditional on the model assumption that the sampling distribution of the mean of the individual differences is Normal. A resampled estimate of the distribution provided in the left-hand panel of Fig. 19.6 suggests that the assumption is reasonable; the points are close to the straight line that represents the Normal distribution.

Again, we also validate the model using Wellek's *PTTE* as coded in the equivalence package of Robinson (2016). Wellek recommends 0.25 as a stringent value for the interval of equivalence, so we adopt that here. The cutoff is 0.187, and the t statistic is −6.33. The absolute value of this statistic is 6.33, which is greater than the cutoff 0.187, so the model is not validated according to Wellek's stringent test.

This conclusion is conditional on the model assumption that the distribution of the individual differences is Normal. A resampled estimate of the distribution provided in the right-hand panel of Fig. 19.6 suggests that the assumption is reasonable; the points are fairly close to the straight line that represents the Normal distribution.

The regression-based test of Robinson et al. (2005) is presented in Fig. 19.7. The figure is interpreted as was Fig. 19.4. This figure shows that we fail to reject the null

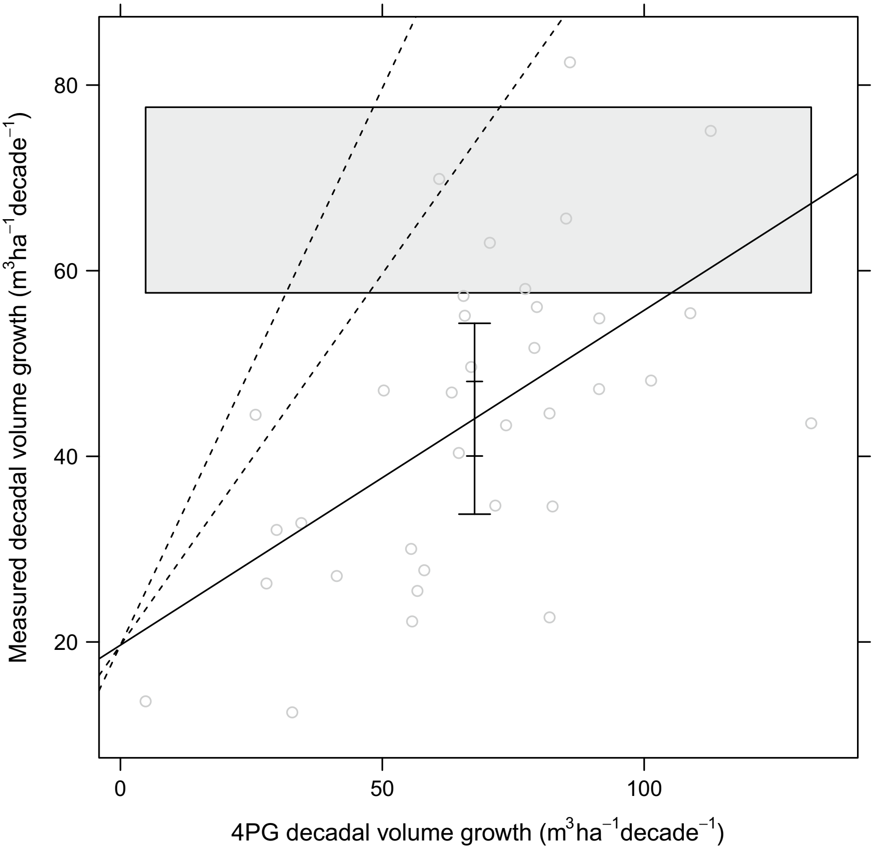


Fig. 19.7 Observed (y-axis) and predicted (x-axis) volume growth amounts for the Priest River Experimental Forest, ID. Explanations of the shapes are in the text

hypothesis of dissimilarity at $\alpha = 0.05$ for the intercept at $\mathscr{I}_0 = \pm 10$ m^{3}ha^{-1}decade^{-1} and for the slope at $\mathscr{I}_1 = \pm 0.2$.

Based on the outcomes of the equivalence tests we believe that 4-PG requires substantial work before it adequately captures volume growth in the Priest River Experimental Forest; it has failed to have been validated.

19.5 Discussion

We have covered three overarching sets of Frequentist tools that can be used for statistical model validation, namely estimates of goodness of fit (correlation, *RMSE*, etc), NHST-based tests (t-tests, linear regression), and equivalence-based tests (*TOST*, *PTTE*, and the test for fidelity).

To some extent, the choice between them will depend on the nature of the validation required. But, if a validation test is required, then it is clear that the shortcomings of the traditional tests identified by Kleijnen (1995) and Robinson and Froese (2004), and as outlined in this chapter, are corrected by equivalence tests, so the latter should be preferred. Another way to consider the difference between traditional NHST and equivalence tests is to note that …"The probability of a type I error in simulation is also called the model builder's risk; the type II error probability is the model user's risk" (Kleijnen 1995). That is, NHST protects the model builder by controlling the model builder's risk, whereas equivalence tests protect the model user by flipping the null hypothesis from being that the model is valid to being that the model is invalid.

19.5.1 Generalizations

An advantage of the regression-style tests, both NHST and equivalence, is that in each case the outcome of the test is based on confidence intervals of the intercept and slope. The means of obtaining the confidence intervals are up to the analyst. If the data have hierarchical structure, or some kind of auto-correlation, then these characteristics can be accommodated in the model used to estimate the slope and intercept, for example using hierarchical statistical models, and therefore model validation can be tested even with data that are derived from more complex circumstances.

Similarly, Aigner (1972) pointed out that the regression approach proposed by Cohen and Cyert (1961) and extended here into the test for fidelity is inappropriate in the case of stochastic simulation models, because the values of the predictor variable, which are the model outputs, are not known exactly. In this case, however, the model parameters can be estimated using a different statistical technique, for example, major axis regression. The confidence intervals on the slope and intercept estimates arising from the model can be interpreted as per the algorithm described in Sect. 19.3.7; all that needs to change is the statistical algorithm used to obtain the estimates.

The attentive reader will recall that in the analysis of the UFC data, we omitted two outliers that were the measures from dead trees. This data-cleaning step is an important one. We were able to remove these two observations with clear conscience because we had external evidence that they were not relevant to the comparison we wanted to make: they were dead trees. In our case, it would not have made any difference to the outcome; the outliers were swamped by the clean data. But what if we had had no other information about the observations, and therefore no principled way to exclude them from the exercise? Then we could have invoked robust statistical techniques. Again, we could compute the one-sided intervals needed for the *TOST*, but using robust statistical techniques to obtain the estimates and intervals, in place of the standard statistical approaches used in this case.

19.5.2 Significant and Important?

One potential criticism that can be leveled at the equivalence tests is the seeming arbitrariness of the establishment of the region of equivalence. Surely it is possible, this criticism imagines, to simply select a region of equivalence that corresponds to our intervals, and thus cook up a spurious validation? And of course this sequence of events is possible, just as it is possible for a disingenuous analyst to choose a convenient value for the size of their NHST. However, the magnitude of the region of equivalence must be declared, and regardless of how it was obtained, if the model user finds the value inappropriate for their own usage, then they can perform their own test if the documentation is complete.

Setting this rejoinder aside, in any case, it is also because of the establishment of this region that the equivalence tests evade the criticism of NHST, namely that arbitrarily small differences can be rendered significant—thereby rejecting the model—if the test has a sufficient sample size. The *TOST* equivalence test, for example, declares the model as validated if the overlap of the one-sided intervals is contained within the region of equivalence. Therefore the model cannot fail to be validated for arbitrarily small differences based on a large sample size.

19.5.3 Nuisance Parameters

A further general concern with the validation of models is in the possibility of tuning of model outputs by changing parameter values to ensure that the model predictions better match the process measures. Even in cases where an analyst takes pains to completely quarantine the model construction from model testing, the parameters may have been inherited from earlier studies that themselves may have reflected some tuning. There is no perfect answer in this scenario. However, there are practices that are clearly less ideal, for example, altering the parameters of a model merely to improve its performance in a validation setting. The risks can be circumnavigated by separating the model testing and the model updating steps: the model can fail to be validated and then the learnings from the validation data can be incorporated into a new version of the model, which then of course must be validated in a separate exercise.

19.5.4 Bayesian or Frequentist Approach?

We conclude the discussion with a few comments about the use of Frequentist tools for model validation as opposed to Bayesian tools, which are described in a companion chapter in this book (see Chap. 20 by Jiang; cf. also Chap. 7 by Beisbart in this volume).

First, any kind of statistical validation is greatly preferable to the sadly all-too-popular practice of plotting the predictions against the observations and commenting that they line up pretty well. Visual inspection and interpretation of data patterns suffer from well known cognitive biases, not least amongst which is overweening optimism, and the statistical validation will inevitably be more reliable regardless of whether it is from the Frequentist or Bayesian foundation. The difference between the Bayesian and Frequentist tools is less important than that reliable statistical tools be used, and that their application be thoughtful and honest, and that the various decisions taken be documented carefully.

Briefly, the difference between the Frequentist and Bayesian approaches can be distilled to a trade-off between resilience and efficiency. It is almost invariably true that Bayesian approaches are as or more efficient than Frequentist approaches in terms of how they use the available data. This means that, all else equal, more precise estimates can usually be drawn from data in the Bayesian setting.

However, this greater efficiency comes at the cost of lower resilience. Bayesian techniques require the analyst to bring more opinion to the analysis, in the forms of probability distributions upon all of the parameters. This reliance upon so-called *prior* information opens the Bayesian analysis to accusations of arbitrariness. In vernacular terms, the Bayesian analysis has more moving parts, and therefore greater vulnerability to breakage.

Often, to try to evade criticisms of arbitrariness, Bayesian analyses will adopt one of a number *uninformative* prior probability distributions, which express (to the best of their mathematical ability) a position of ignorance or indifference about the true value of the parameter. However, even this adoption represents a choice by the analyst, and while it may be the best supported or even the most conservative choice, depending on the context, it is nonetheless a choice, and other choices could be made that may yield different outcomes. Therefore the outcomes of the analysis are contingent on this weightier and fragile apparatus.

Frequentist analysis is not entirely safe from analogous accusations. The Frequentist approach also demands the invocation of probability distribution functions as a fundamental part of inference. Often, the analyst will invoke the *CLT* in order to assume that the sampling distribution of the sample mean is Normal; such a step greatly simplifies the analysis. But then, the *CLT* does not make any comment about the Normality or otherwise of the sample mean of the data at hand, rather, it says that if we had a larger sample, then the distribution of the mean would be likely to be closer to Normal, and if we had a smaller sample, then the distribution of the mean would be likely to be less Normal. A resilient way to proceed is by resampling estimates of the mean from the sample, and comparing the resampled distribution with the Normal distribution, for example by a quantile–quantile plot—as we have done in this chapter.

Furthermore, the Frequentist model of probability depends upon the principle of an infinite sequence of repeatable instances of experiments. This an appealing principle in cases of very simple experimental setups, such as the flip of a coin or the shuffling of a deck of cards. However, the principle is less straightforward to invoke in cases of field or forest experiments (Heraclitus might have said that one cannot take a sample from the same river twice). Therefore infinite repeatability is a vexing concept in the Frequentist setup. However, the Frequentist approach captures the arguably non-repeatable aspects in the noise parameter of the model (σ^2 in Eq. 19.1). The assumption of infinite repeatability is therefore conservative, and non-repeatability is arguably compensated in the model, and the tests that arise from it, by the fact that noisy data lead to less powerful tests.

19.5.5 Conclusion

We have outlined the problem of model validation as tackled through statistical tools that arise in the Frequentist modeling paradigm. We surveyed a collection of statistical tools that have previously been applied to model validation, and pointed out that the interpretation of some of them is complicated by their very nature: they are tools designed to split, to detect differences, rather than to lump, to detect similarities. We showed that equivalence tests can fulfil this latter role, and provided two case studies demonstrating their application.

Using equivalence tests, arbitrarily small differences cannot be detected through arbitrarily large samples, and failure to detect a difference cannot be explained away by low power. The importance, as opposed to statistical significance, enters through the specification of the region of equivalence. For these reasons we argue that equivalence tests are superior to the usual hypothesis tests when the goal is model validation, and the former should be adopted whenever possible.

We conclude with an aphorism: there is only one kind of analyst, and some of them split whereas some of them lump.

Acknowledgements This study is supported in part by the Centre of Excellence for Biosecurity Risk Analysis, School of BioSciences, University of Melbourne, Australia. Thoughtful review comments by Lori Dalton, Steve Lane, Anca Hanea, James Camac, and the two editors have greatly improved this chapter.

References

Aigner, D. J. (1972). A note on verification of computer simulation models. *Management Science*, *18*(11), 615–619.

Alewell, C., & Manderscheid, B. (1998). Use of objective criteria for the assessment of biogeochemical ecosystem models. *Ecological Modelling*, *107*, 213–224.

Bartelink, H. H. (1998). Radiation interception by forest trees: A simulation study on effects of stand density and foliage clustering on absorption and transmission. *Ecological Modelling*, *105*, 213–225.

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, *11*(4), 283–319.

Capes, H., et al. (2017). The allometric quarter-power scaling model and its applicability to Grand fir and Eucalyptus trees. *Journal of Agricultural, Biological, and Environmental Statistics,* 7, 1–23.

Casella, G., & Berger, R. L. (1990). *Statistical inference*. Belmont, CA.: Duxbury Press.

Caswell, H. (1976). The validation problem. In B. Patten (Ed.), *Systems analysis and simulation in ecology* (Vol. 4, pp. 313–325). Cambridge: Academic Press.

Cohen, K. J., & Cyert, R. M. (1961). Computer models in dynamic economics. *The Quarterly Journal of Economics*, *75*(1), 112–127.

Duursma, R., Marshall, J., Robinson, A., & Pangle, R. (2007). Description and test of a simple process-based model of forest growth for mixed-species stands. *Ecological Modelling*, *203*(3–4), 297–311.

Freese, F. (1960). Testing accuracy. *Forest Science*, *6*(2), 139–145.

Gentil, S., & Blake, G. (1981). Validation of complex ecosystem models. *Ecological Modelling*, *14*, 21–38.

Gregoire, T. G., & Reynolds, M. R, Jr. (1988). Accuracy testing and estimation alternatives. *Forest Science*, *34*(2), 302–320.

Jans-Hammermeister, D. C., & McGill, W. B. (1997). Evaluation of three simulation models used to describe plant residue decomposition in soil. *Ecological Modelling*, *104*, 1–13.

Kleijnen, J. P. C. (1995). Verification and validation of simulation models. *European Journal of Operational Research, 82,* 145–162.

Kleijnen, J. P. C., Bettonvil, B., & Van Groenendaal, W. (1998). Validation of trace-driven simulation models: A novel regression test. *Management Science, 44*(6), 812–819.

Kleijnen, J. P. C. (1974). *Statistical techniques in simulation (part 1)*. New York.: Marcel Dekker.

Landsberg, J. J., Waring, R. H., & Coops, N. C. (2003). Performance of the forest productivity model 3-PG applied to a wide range of forest types. *Forest Ecology and Management*, *172*, 199–214.

Loehle, C. (1997). A hypothesis testing framework for evaluating ecosystem model performance. *Ecological Modelling*, *97*, 153–165.

Mayer, D. G., & Butler, D. G. (1993). Statistical validation. *Ecological Modelling*, *68*, 21–32.

McBride, G. B. (1999). Equivalence tests can enhance environmental science and management. *Australian and New Zealand Journal of Statistics*, *41*(1), 19–29.

Meyners, M. (2012). Equivalence tests—A review. *Food Quality and Preference*, *26*(2), 231–245.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*, 641–646.

Ottosson, F., & Håkanson, L. (1997). Presentation and analysis of a model simulating the pH response of lake liming. *Ecological Modelling*, *105*, 89–111.

Parkhurst, D. F. (2001). Statistical significance tests: equivalence and reverse tests should reduce misinterpretation. *Bioscience*, *51*(12), 1051–1057.

Pocewicz, A. L., Gessler, P., & Robinson, A. P. (2004). The relationship between effective plant area index and landsat spectral response across elevation, solar insolation, and spatial scales in a northern Idaho forest. *Canadian Journal of Forest Research*, *34*(2), 465–480.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reynolds, M. R, Jr. (1984). Estimating the error in model predictions. *Forest Science*, *30*(2), 454–469.

Reynolds, M. R, Jr., Burkhart, H. E., & Daniels, R. F. (1981). Procedures for statistical validation of stochastic simulation models. *Forest Science*, *27*(2), 349–364.

Robinson, A. (2016). *Equivalence: Provides tests and graphics for assessing tests of equivalence*. R package version 0.7.2.

Robinson, A., Duursma, R., & Marshall, J. (2005). A regression-based equivalence test for model validation: Shifting the burden of proof. *Tree Physiology*, *25*(7), 903.

Robinson, A. P., & Ek, A. R. (2000). The consequences of hierarchy for modelling in forest ecosystems. *Canadian Journal of Forest Research*, *30*(12), 1837–1846.

Robinson, A. P., & Froese, R. E. (2004). Model validation using equivalence tests. *Ecological Modelling*, *176*(3–4), 349–358.

Rykiel, E. J. (1996). Testing ecological models—The meaning of validation. *Ecological Modelling*, *90*(3), 229–244.

Sargent, R. G. (2012). Verification and validation of simulation models. *Journal of Simulation*, *7*(1), 12–24.

Vanclay, J. K., & Skovsgaard, J. P. (1997). Evaluating forest growth models. *Ecological Modelling*, *98*(1), 1–12.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Chapman and Hall/CRC.

Wykoff, W., Crookston, N., & Stage, A. (1982). *User's guide to the stand prognosis model*. USDA Forest Service Intermountain Research Station, Ogden, UT. GTR-INT 133, 113 p.

Chapter 20
Validation Using Bayesian Methods

Xiaomo Jiang, Xueyu Cheng and Yong Yuan

Abstract Quantitative model validation is playing an increasingly important role in performance and reliability assessment of a complicated system whenever simulation is involved. This chapter discusses model validation from a Bayesian perspective, considering in particular data uncertainty. First, Bayes' theorem is defined, then the Bayesian risk rule method is introduced. Explicit expressions for the Bayesian interval hypothesis testing approach are presented in both univariate and multivariate cases. The problem of non-normal validation data is addressed by the Box–Cox transformation. A generalized procedure is presented to implement Bayesian validation methods. Classic hypothesis testing method is utilized to conduct a comparison study. The impact of the data normality assumption and of the variation of the threshold on model assessment accuracy is investigated by using both classical and Bayesian approaches. The Bayesian methodology is illustrated with a reliability model of rotor blades, a univariate stochastic damage accumulation model, and a multivariate heat conduction problem.

Keywords Bayesian statistics · Hypothesis testing · Box–Cox transformation · Bayes network · Model validation · Non-normality

20.1 Introduction

Due to the rapid advance of computer technology and to the high prohibitive costs of full-scale testing, model-based simulation has been increasingly used in the design and analysis of complex systems in science and engineering (e.g., Roache 1998; Kennedy and O'Hagan 2001; Oberkampf and Trucano 2002; Babuska and Oden 2004; Chen et al. 2004, 2008). A computational model is established either from

X. Jiang (✉) · Y. Yuan
Tongji University, Shanghai 200092, China
e-mail: jiang.xiaomo@gmail.com

X. Cheng
Clayton State University, Morrow, GA 30620, USA

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_20

measured field data, from an underlying mechanism, or comprehensively from the combination of field data and a mechanism, to approximate a real-world process. The resulting model is then used to simulate the process in the target system under various scenarios, e.g., for design optimization, safety analysis, and reliability assessment of the underlying physical system. Before the model can be applied, quantitative validation is needed to establish confidence in the model predictions. The computational model, however, always contains a broad spectrum of uncertainties or unknown factors, which stem mainly from uncertainties in the measurement, modeling, coding, and parameters. These uncertainties are usually classified into two categories: aleatoric uncertainty and epistemic uncertainty (cf. Chap. 5 by Roy in this volume). The former is irreducible, as in inherent variability, such as statistically distributed properties and manufacturing variability, while the latter is potentially reducible uncertainty due to lack of knowledge, such as model form and initial and boundary condition v approximations. Quantitative approaches are usually required to handle the uncertainties in the model validation.

In the past decades, the fundamental concepts and methodologies for model verification and validation have been widely investigated by several organizations such as the United States Department of Defense (1996), American Institute of Aeronautics and Astronautics (1998), Advanced Simulation and Computing (ASC) program of the United Sates Department of Energy (2000), and American Society of Mechanical Engineers Council (2002, 2006). Figure 20.1 shows the interaction of modeling, verification and validation, which is modified from the so-called *Sargent circle* in Schlesinger (1979). A real-world physical system, e.g., a rotor blade in the engine rotor of an aircraft, is numerically modeled, leading to a mathematical model. The mathematical model, for instance, may be represented by a set of partial differential equations, which is obtained by analyzing the real physical system conceptually. Note that, the model yields a conceptual/mathematical/numerical description of the physical system, including geometrical data, material properties, and initial and boundary conditions. Next, the mathematical model is converted into a computational model in terms of a numerical algorithm via programming or coding. The computational model represents the physical system with the initial and boundary conditions as well as its material properties. *Model verification* is used to check whether the mathematical model has been converted into the computational model correctly (cf. Chap. 10 by Rider in this volume). In the process of verification, confidence needs to be established via comparing model results with analytical solutions or benchmark problems to ensure that the code is free of errors. Furthermore, confidence needs to be established that the model is an accurate representation of the real-world physical system. *Model validation* involves comparing model predictions with data from the target system. Based on the comparison results, the decision maker can judge whether to accept or reject the model. This chapter is focused on quantitative validation of the computational model under uncertainties.

In the past, subjective judgments based on graphical plots were often used to assess how good the model is. But the quality of the computational model cannot be assessed quantitatively in a graphical comparison. In addition, many critical issues, such as data correlation between multiple variables, uncertainties in both test

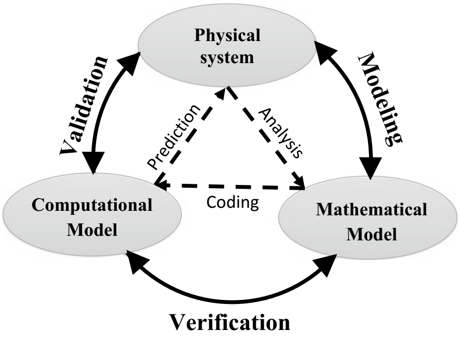


Fig. 20.1 Interaction of modeling, verification, and validation (modified from the Sargent circle originally created by Schlesinger 1979)

and model predictions, and confidence in the model, are ignored. Oberkampf and Barone (2006) have demonstrated that model validation has progressed from qualitative graphical comparisons, without considering uncertainties in either test data or model prediction, to quantitative analysis of the differences between experiments and predictions with the consideration of uncertainties. Recently, increasing attention has been paid to quantitative validation comparisons considering uncertainties in both experimental and model outputs. To develop an effective model assessment approach to ensure sufficient accuracy of model predictions, a proper validation metric is needed as quantitative measure of agreement between two sets of data under uncertainty (see e.g., Roache 1998; Oberkampf and Trucano 2002; Babuska and Oden 2004; Mahadevan and Rebba 2005; Oberkampf and Barone 2006; Rebba and Mahadevan 2006a, b, 2008; Schwer 2007; Jiang and Mahadevan 2007, 2008a, b, 2009a, b, 2010; Chenet al. 2008).

Statistical hypothesis testing is a widely used quantitative approach to validation of a computational model under uncertainty [see Oberkampf and Barone (2006) for a comprehensive state-of-the-art review]. Two types of hypothesis testing-based approaches may be pursued to develop model validation metrics: classical (also called frequentist or error statistical) and Bayesian methods. Classical hypothesis testing is a well-developed statistical method for rejecting a model on the basis of a test statistic (see e.g., Hills and Trucano 2002; Chen et al. 2004; Oberkampf and Barone 2006; see also Chap. 19 by Robinson in this volume). It leads to a decision about the model: The so-called p-value is obtained in the hypothesis test and used as a decision variable to determine whether to accept or reject the null hypothesis (i.e., to judge a valid model). In the past decade, alternatively, Bayesian methods have been developed to determine the predictive capabilities of computational models (see e.g., Kennedy and O'Hagan 2001; Chen et al. 2004, 2008; Mahadevan and Rebba 2005; Rebba and Mahadevan 2006a, b, 2008; Jiang and Mahadevan 2007, 2008a, b, 2009a, b, 2010). One major difference between the Bayesian and classical hypothesis testing approaches lies in the fact that the Bayesian approach focuses on model acceptance, whereas classical hypothesis testing focuses on model rejection. It should be mentioned that not having enough evidence to reject a model is not the same as having enough evidence to accept the model. The differences between classical and Bayesian hypothesis testing have been discussed in detail by many researchers (e.g., Berger and Delampady 1987; Hwang et al. 1992). Bayesian hypothesis testing-based methods (Zhang and Mahadevan 2003), Bayesian risk-based decision-making methodology (Jiang and Mahadevan 2007), and Bayesian structural equation modeling approach (Jiang and Mahadevan 2009b, 2010) have also been explored for model assessment. These metrics have been investigated with various model validation problems using limited amount of experimental data. Refer to Mahadevan and Rebba (2005) and Jiang and Mahadevan (2007, 2008a, b) for details of the Bayesian validation methods and their applications.

The aim of this chapter is to introduce Bayesian methods of model validation. Some basics are given in Sect. 20.2. The Bayesian decision rule is presented in Sect. 20.3. Then, Bayesian univariate interval hypothesis testing approach is discussed (Sect. 20.4). The Multivariate case is presented in terms of Bayesian hypothesis testing approaches (Sect. 20.5). Explicit expressions for the interval hypothesis testing-based Bayes factor are derived for both univariate and multivariate cases. Next, the Bayesian confidence measure is presented based on the Bayes factor. Non-normal data transformations and Bayes networks are also introduced. A generalized procedure is proposed to implement the proposed Bayesian methodology for model validation of complicated systems with either single or multiple response quantities. Finally, in Sect. 20.10, the Bayesian methods are investigated with three examples, namely, a reliability model of rotor blades in aircraft engines, a univariate stochastic damage accumulation model, and a multivariate heat conduction problem.

20.2 Fundamentals

In the decision-based approach to model validation, let d_0 denote the decision to accept the null hypothesis H_0: $y = y_0$, where y_0 and y are the predicted and actual values of a physical quantity of interest, respectively. Let likewise d_1 denote the decision to accept the alternative hypothesis H_1: $y \neq y_0$. A utility function $u(d_i, y)$ is defined in order to choose d_i based on a decision rule. Given the observed data Y, the decision d_0 is made if and only if $E[\underline{u}(d_0, y) - u(d_1, y)|Y] > 0$. In the classical statistics approach, given the experimental observations, a hypothesis test is conducted in terms of the conditional probabilities of Type I error (rejecting a correct model) and Type II error (accepting a wrong model). An expected loss function is defined based on the conditional error probabilities. Either a squared loss function or an absolute error metric is usually chosen as the difference in the loss function (Schervish 1995). The task of a decision is to minimize the expected loss. Balci and Sargent (1981) presented a classical hypothesis testing-based cost–risk decision analysis to validate a simulation model of a real system, considering the model user's risk, model builder's risk, acceptable validity range, budget, sample sizes, and cost of data collection.

In the Bayesian approach, the task of deciding between H_0 and H_1 is conceptually more straightforward. One merely needs to calculate the posterior probabilities α_0

= Pr($H_0|Y$) and α_1 = Pr($H_1|Y$), or their likelihoods given experimental data. An expected loss function is then defined as a function of α_0 and α_1. Its conceptual advantage is that α_0 and α_1 are the actual probabilities of the hypotheses in light of the observed data and prior knowledge.

There are two types of Bayesian hypothesis testing methods: point-based and interval-based approaches. The former tests whether the difference between model prediction and experimental observation is equal to zero, while the latter tests whether the difference is within an allowable limit. The study by Rebba and Mahadevan (2008) has demonstrated that the interval-based hypothesis testing method provides more consistent validation results than a point hypothesis testing method. As the amount of data increases, the interval-based method converges to the correct inference. Recently, Jiang and Mahadevan (2008b, 2009a, b) have derived an explicit expression for calculating the Bayes factor based on the interval-based hypothesis testing, with the purpose of facilitating the overall reliability assessment of model validation. Refer to Jiang and Mahadevan (2008b, 2009a) for details about the interval-based Bayesian validation metric as well as its applications.

The Bayesian approach to validation makes heavily use of Bayes' theorem, also known as Bayes' rule. The latter is a combination of traditional probabilities and statistics. Let A and B be the two events. The conditional probability of event A given that we know event B has occurred, $P(A|B)$, can be mathematically expressed as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{20.1}$$

where $P(B)$ is the probability of event B, which must be positive, i.e., $P(B) > 0$, and $P(A \cap B)$ is the probability that both A and B occurred. Since $P(A \cap B) = P(B \cap A)$, Eq. (20.1) can be rewritten as follows:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \tag{20.2}$$

By dividing the right-hand equality by $P(B)$, one obtains Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{20.3}$$

Bayes' theorem, as expressed in Eq. (20.3), links the conditional probability $P(B|A)$ to the other conditional one $P(A|B)$. It involves two basic concepts specific to Bayesian methods: the prior probability $P(A)$ and the posterior probability $P(A|B)$. The prior is the unconditioned probability which is usually used to incorporate the previous experience or knowledge, while the posterior is the conditional one.

Like traditional statistical validation metrics, for instance, the p-value from hypothesis testing, the Bayes factor can be considered to be a metric to evaluate the model accuracy. The Bayes factor is often called the likelihood ratio of data given the binary hypothesis testing (H_0: model accepted vs. H_1: model rejected). It may be derived explicitly from hypothesis testing or implicitly from a Bayes network via numerical simulation. In the following, the Bayesian decision rule is presented to determine a validation threshold considering decision costs. The Bayes factor metric is then derived from univariate and multivariate interval hypotheses to address the limited number of data and uncertainties in the model validation.

20.3 Bayesian Decision Rule

Assuming that event A is a hypothesis and that B is new data or evidence, Bayes' rule is applied to update the belief about the hypothesis A in the light of new evidence B. Specifically from Eq. (20.3), the posterior belief $P(A|B)$ can be calculated by multiplying the prior belief $P(A)$ with the conditional probability (or likelihood) $P(B|A)$ that B will occur if A is true (see Chap. 7 by Beisbart in this volume for a philosophical introduction to Bayesianism).

Within the context of binary hypothesis testing in model validation, we need to consider two hypotheses H_0 and H_1. The classical interval-based hypotheses are represented as null hypothesis $H_0 : |d| \leq \varepsilon$ versus alternative hypothesis $H_1 : |d| > \varepsilon$, where d is the difference of a univariate variable between the model prediction and validation data, and ε is a predefined value. The decision maker or model user has to decide which threshold is acceptable. The symbol |.| denotes the absolute value. Here we are testing whether the difference d is, in fact, close to zero. The prior probabilities of two hypotheses are denoted by

$$\pi_0 = \Pr(H_0) \text{ and } \pi_1 = \Pr(H_1) \tag{20.4}$$

Note that $\pi_1 = 1 - \pi_0$ for the binary hypothesis testing problem. Each time model validation is conducted given the experimental data, one of the four possible scenarios, $[H_i|H_j]$ $(i = 0, 1; j = 0, 1)$, may happen, where $[H_i|H_j]$ is the event of inferring H_i when H_j is true. In analogy to the classical testing approach, the type I and II error probabilities (α and β, respectively) are calculated as

$$\alpha = \Pr(H_1|H_0) \text{ and } \beta = \Pr(H_0|H_1) \tag{20.5}$$

Let Z represent the entire experimental data set, and Z_0 and Z_1 represent two mutually exclusive subsets of Z such that $Z_0 \cup Z_1 = Z$, and $Z_0 \cap Z_1 = \varnothing$, where $\cup$ and $\cap$ represent union and intersection, respectively, and $\varnothing$ represents an empty set or null space. Thus, Z_0 and Z_1 represent two decision regions corresponding to two hypotheses H_0 and H_1. Every possible observation Y belongs to either decision region. The problem to be solved here is to assign each experiment to Z_0 or Z_1 such that a minimum risk in model validation is obtained. Figure 20.2 shows a schematic illustration of the decision-based validation method.

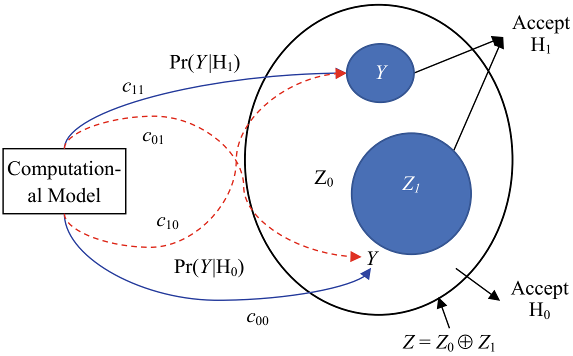


Fig. 20.2 Schematic illustration of Bayesian decision-based model validation

Assume that the observation Y has a probability density function under each hypothesis, i.e., $Y|H_0 \sim f(y|H_0)$ and $Y|H_1 \sim f(y|H_1)$. Thus, $\Pr[H_i|H_j] = \int_{Z_i} f(y|H_j)dy$ $(i = 0, 1; j = 0, 1)$ is obtained. The loss function (or risk) for model validation is defined to be the expected cost of validation, which is obtained by averaging the decision cost (the preferences of a decision maker) over two probabilities: the prior probability of the hypothesis and the probability of a particular action to be taken (Nowak and Scott 2004; Jiang and Mahadevan 2007):

$$R = \sum\nolimits_{j=0}^{1} \sum\nolimits_{i=0}^{1} \left(c_{ij}\pi_j \int_{Z_i} f(y|H_j)dy \right) \tag{20.6}$$

where c_{ij} = the cost of deciding H_i when H_j is true (decision consequence).

Based on the assumption that the total risk (cost) resulting from a correct decision is always less than the total risk resulting from a wrong decision, Jiang and Mahadevan (2007) derive the Bayes decision rule to accept the model through minimizing the total Bayes risk (Eq. 20.6) as follows:

$$\Lambda(y) = \frac{f(\text{data}|H_0)}{f(\text{data}|H_1)} = \frac{f(y|H_0)}{f(y|H_1)} > \frac{\pi_1(c_{01} - c_{11})}{\pi_0(c_{10} - c_{00})} = \eta \tag{20.7}$$

where $\Lambda(y)$ is the likelihood ratio, referred to as Bayes factor, and η is the acceptable threshold which is dependent on the prior densities of the two hypotheses and the costs of deciding H_i when H_j is true $(i = 0, 1; j = 0, 1)$. Equation (20.7) is the Bayesian decision rule developed by Jiang and Mahadevan (2007) for univariate model validation. When particular cost information c_i (e.g., $c_{00} = c_{11} = 0$ and $c_{01} = c_{10} = 1$) and prior densities π_i (e.g., $\pi_0 = \pi_1 = 0.5$) are assumed, the threshold η =1 is obtained, as the Bayes factor approach proposed by Zhang and Mahadevan (2003) for model validation. For the sake of simplicity, this particular cost information is applied in the examples for univariate and multivariate hypothesis testing cases below.

In practical applications of the Bayes risk approach to model validation, it becomes critical to efficiently compute the probability density (or likelihood) function of experimental data under each hypothesis. If data is available only on one or more intermediate quantities, a Bayes network (Jensen and Jensen 2001) approach and a Markov chain Monte Carlo (MCMC) simulation technique have been suggested by Mahadevan and Rebba (2005) to estimate the probability density of the response quantity of interest. In the following Sects. 20.4 and 20.5, the likelihoods of experimental data under two hypotheses are derived mathematically based on Bayes' theorem for both univariate and multivariate model validation, respectively.

20.4 Bayesian Univariate Hypothesis Testing

In classical interval-based hypothesis testing, the difference d must follow a normal distribution such that the one-sample t-test can be performed to determine the p-value. The decision rule based on the resulting p-value is used for the test, i.e., if the p-value is smaller than some threshold, then H_0 is rejected. Note that, not having enough evidence to reject a model is not the same as having enough evidence to accept the model, which will be demonstrated later in the illustrative examples.

Bayesian interval-based hypothesis testing is presented here to derive the Bayes factor metric explicitly. We assume that the difference d has a probability density function (PDF, or likelihood function) under each hypothesis, i.e., $d|H_0 \sim f(d|H_0)$ and $d|H_1 \sim f(d|H_1)$. Usually, we assume that (1) the difference data vector $\mathbf{d} = \{e_1, e_2, \ldots, e_n\}$ follows a normal distribution $N(\mu, \sigma^2)$ with known standard deviation σ (estimated from data), and (2) a prior probability density function for the mean μ under both null and alternative hypotheses, denoted by $f(\mu)$, is taken to be normal too: $N(\rho, \tau^2)$. The selection of the proper value of ρ and τ is still a matter of argument between Bayesians and frequentists (Migon and Gamerman 1999). If no information on $f(\mu)$ is available, the parameters $\rho = 0$ and $\tau^2 = \sigma^2$ are suggested in Migon and Gamerman (1999). This selection assumes that the amount of information in the prior is equal to that in the observation, which is consistent with the Fisher information-based method (Kass and Raftery 1995). Given a set of validation data, the likelihood ratio, referred to as the Bayes factor (B_i) in the interval hypothesis testing, is calculated using Bayes' theorem as (Jiang and Mahadevan 2009a)

$$B_i = \frac{f(\text{Data}|H_0)}{f(\text{Data}|H_1)} = \frac{\int_{-\varepsilon}^{\varepsilon} f(\boldsymbol{d}|\mu)f(\mu)d\mu}{\int_{-\infty}^{-\varepsilon} f(\boldsymbol{d}|\mu)f(\mu)d\mu + \int_{\varepsilon}^{\infty} f(\boldsymbol{d}|\mu)f(\mu)d\mu} \tag{20.8}$$

Note that by using Bayes' theorem in Eq. (20.3), $f(\mu|\mathbf{d}) \propto f(\mathbf{d}|\mu)f(\mu)$, Eq. (20.8) can be easily transferred to the area ratio of the posterior density of μ under two hypotheses, expressed as follows (see Fig. 20.3, where $\varepsilon_1 = -\varepsilon$ and $\varepsilon_2 = \varepsilon$):

$$B_i = \frac{\int_{-\varepsilon}^{\varepsilon} f(\mu|\mathbf{d})d\mu}{\int_{-\infty}^{-\varepsilon} f(\mu|\mathbf{d})d\mu + \int_{\varepsilon}^{\infty} f(\mu|\mathbf{d})d\mu} = \frac{K}{1-K} \tag{20.9}$$

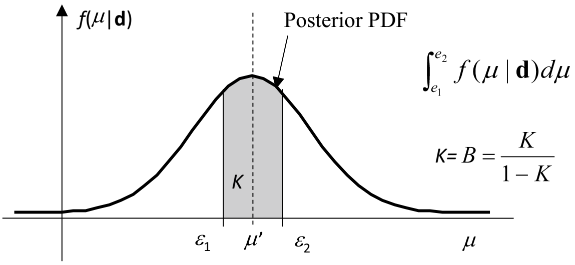


Fig. 20.3 Geometric meaning of interval-based Bayes factor method in univariate case

where $K = \int_{-\varepsilon}^{\varepsilon} f(\mu|\mathbf{d})d\mu$ represents the area of the posterior density of μ under the null hypothesis and $1 - K$ represents the area of the posterior density of μ under the alternative hypothesis (Fig. 20.3). The value of K is calculated by (Jiang and Mahadevan 2009a)

$$K = \Phi\left(\frac{\lambda_2 - \mu'}{\sigma'}\right) - \Phi\left(\frac{\lambda_1 - \mu'}{\sigma'}\right) \tag{20.10}$$

in which Φ(.) is the standard normal distribution with $\lambda_1 = -\varepsilon\sqrt{n\tau^2 + \sigma^2}$, $\lambda_2 = \varepsilon\sqrt{n\tau^2 + \sigma^2}$, $\mu' = \frac{n\bar{e}\tau^2+\rho\sigma^2}{\sqrt{n\tau^2+\sigma^2}}$ and $\sigma'^2 = \sigma^2\tau^2$. Note that the quantity K in Eq. (20.10) is dependent on the value of ε. The decision maker or model user has to decide what ε is acceptable. When $\varepsilon \to \infty$, B_i in Eq. (20.9) will increase indefinitely, thus the data will always support the model. When $\varepsilon \to 0$, B_i in Eq. (20.9) will approach zero (i.e., the nominator in Eq. (20.9) or the shadowed area in Fig. 20.3 approaches zero), thus the data will always reject the model. Refer to Jiang and Mahadevan (2009a) for the details about the derivative of the interval-based Bayes factor.

20.5 Multivariate Bayesian Hypothesis Testing

The Bayesian interval-based hypothesis testing method has been developed for ***multivariate*** model validation (Jiang and Mahadevan 2008b). Like the univariate case, explicit expressions were derived to calculate the Bayes factor based on interval hypothesis testing for the multivariate case, with the assumption of multivariate normal distribution for difference data. The multivariate model validation problem becomes testing the two hypotheses H_0: $\boldsymbol{\mu} = \mathbf{E}_0$ versus H_1: $\boldsymbol{\mu} \neq \mathbf{E}_0$ with $\boldsymbol{\mu}|H_1 \sim N(\rho, \boldsymbol{\Lambda})$, where $\boldsymbol{\mu}$ is the multivariate mean variable, $\mathbf{E}_0$ is a zeros vector, and ρ and $\boldsymbol{\Lambda}$ are the assumed mean vector and covariance matrix for the multivariate normal distribution, respectively. In this section, the multivariate interval-based hypothesis testing method is presented to facilitate the overall validation assessment of computational models with higher accuracy.

Similar to the univariate case, the Bayes factor for the multivariate case, $B_{i,\mathbf{M}}$, is expressed as follows:

$$B_{i,\mathbf{M}} = \frac{f(\text{Data}|H_0)}{f(\text{Data}|H_1)} = \frac{\int_{\boldsymbol{\varepsilon}_1}^{\boldsymbol{\varepsilon}_2} f(\boldsymbol{\delta}|\boldsymbol{D})d\boldsymbol{\delta}}{\int_{-\infty}^{\boldsymbol{\varepsilon}_1} f(\boldsymbol{\delta}|\boldsymbol{D})d\boldsymbol{\delta} + \int_{\boldsymbol{\varepsilon}_2}^{\infty} f(\boldsymbol{\delta}|\boldsymbol{D})d\boldsymbol{\delta}} = \frac{K_{\mathbf{M}}}{1 - K_{\mathbf{M}}} \tag{20.11}$$

where the multivariable integral of $K_{\mathbf{M}} = \int_{\boldsymbol{\varepsilon}_1}^{\boldsymbol{\varepsilon}_2} f(\boldsymbol{\delta}|\mathbf{D})d\boldsymbol{\delta}$ represents the volume of the posterior density of $\boldsymbol{\delta}$ under the null hypothesis, and again ε_1 and ε_2 are the lower and upper threshold that the decision maker needs to set, respectively. The value of $1 - K_{\mathbf{M}}$ represents the area of the posterior density of $\boldsymbol{\delta}$ under the alternative hypothesis. Instead of the italic subscript in Eq. (20.8) for the Bayes factor, the bold subscript is used in Eq. (20.11) to represent the multivariate case. Refer to Jiang and Mahadevan (2008b) for the details about the numerical integration in Eq. (20.11).

In Eq. (20.11), the parameter $K_{\mathbf{M}}$ can be obtained using the standard multivariate normal distribution as follows

$$K_{\mathbf{M}} = \Phi\left(\boldsymbol{\theta}_2', \mathbf{Z}_0, \boldsymbol{\Pi}\right) - \Phi\left(\boldsymbol{\theta}_1', \mathbf{Z}_0, \boldsymbol{\Pi}\right) \tag{20.12}$$

in which the parameters $\boldsymbol{\theta}_1' = (\boldsymbol{\varepsilon}_1 + \mathbf{e}_0)\sqrt{n|\boldsymbol{\Lambda}| + |\boldsymbol{\Sigma}|}$, $\boldsymbol{\theta}_2' = (\boldsymbol{\varepsilon}_2 + \mathbf{e}_0)\sqrt{n|\boldsymbol{\Lambda}| + |\boldsymbol{\Sigma}|}$, $\mathbf{Z}_0 = \frac{n\bar{\mathbf{D}}|\boldsymbol{\Lambda}|+\boldsymbol{\rho}|\boldsymbol{\Sigma}|}{\sqrt{n|\boldsymbol{\Lambda}|+|\boldsymbol{\Sigma}|}}$, and $\boldsymbol{\Pi} = \boldsymbol{\Sigma}|\boldsymbol{\Lambda}|$, and $\Phi(.)$ presents a multivariate normal cumulative distribution function, which is computed using the numerical algorithm proposed by Genz (1992). The symbol $\boldsymbol{\Sigma}$ is a $m \times m$ covariance matrix of all variables. The symbol |.| denotes the determinant of a matrix. Let $\mathbf{D}_i = [\, D_{i1}\;\, D_{i2}\; \cdots\; D_{im}\,]^{\mathrm{T}}$ ($i = 1, 2, \ldots, n$) represent the ith difference of m component variables, each having n data points, $\bar{\mathbf{D}} = [\, \bar{D}_1\;\, \bar{D}_2\; \cdots\; \bar{D}_m\,]^{\mathrm{T}}$ is the m component average values with $\bar{D}_i = \frac{1}{n}\sum_{j=1}^{n} D_{ij}$ ($i = 1, 2, \ldots, m$).

Figure 20.4 shows the geometric meaning of the interval-based Bayes factor in the bivariate case with the standard normal distribution for the two variables. The cylinder in Fig. 20.4 represents the volume of the posterior PDF under the null hypothesis, defined by the density function within the given interval $\left[\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\right]$ (i.e., $K_{\mathbf{M}}$). As the variation of the posterior density of $\boldsymbol{\mu}$ decreases, its volume within the interval increases. Obviously, if validation data are to support the model (i.e., $\varepsilon_1 \leq |D| \leq \varepsilon_2$), the volume of the posterior PDF under the null hypothesis will increase (a larger volume of the cylinder in Fig. 20.4); otherwise, the area will decrease (a smaller volume of the cylinder in Fig. 20.4).

20.6 A Bayesian Measure of Evidence

Note that, the quantity K in Eq. (20.9) or K_M in Eq. (20.11) is dependent on the initial interval threshold (e.g., ε in Eq. 20.9). The decision maker or model user has to decide which threshold is acceptable. When the threshold approaches ∞, the Bayes factor value will increase indefinitely, thus the data will always support the model. When the threshold approaches 0, the obtained Bayes factor will approach zero (i.e., the nominator in Eqs. (20.9) or (20.11) or the shadowed area in Figs. 20.3 or

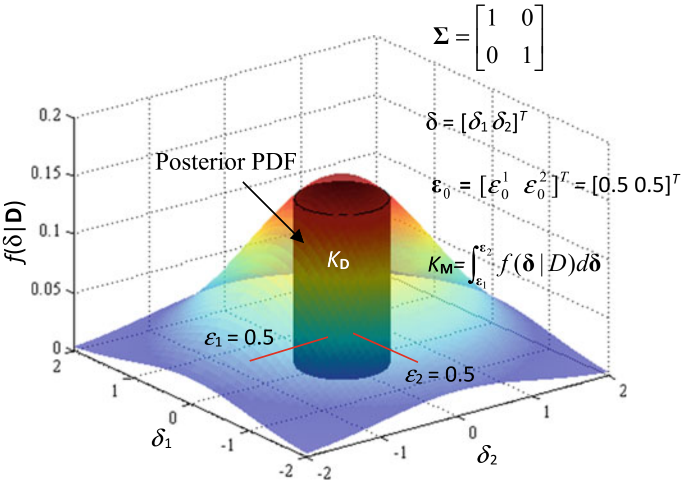


Fig. 20.4 Geometric meaning of interval-based Bayes factor in bivariate case

20.4 approaches zero), thus the data will always reject the model. Since the obtained Bayes factor is nonnegative, the Bayes factor is often converted into the logarithm scale for the convenience of comparison among a larger range of values [e.g., $b_i = \ln(B_i)$ for Eq. (20.9)]. Kass and Raftery (1995) suggest interpreting the logarithm of the Bayes factor between 0 and 1 as weak evidence in favor of H_0, between 3 and 5 as convincing evidence, and greater than 5 as very strong evidence. Negative b_i of the same magnitude is said to favor H_1 by the same amount. This thumb guideline is applicable for both the univariate and multivariate cases.

The Bayesian measure of evidence that the computational model is valid may be quantified by the posterior probability of the null hypothesis $\Pr(H_0|\mathbf{d})$. Using the Bayes theorem described previously, the relative posterior probabilities of two models are obtained as

$$\frac{\Pr(H_0|\,\mathbf{d})}{\Pr(H_1|\,\mathbf{d})} = \left[\frac{\Pr(\mathbf{d}\,|H_0)}{\Pr(\mathbf{d}|H_1)}\right]\left[\frac{\Pr(H_0)}{\Pr(H_1)}\right] \tag{20.13}$$

where the term in the first set of square brackets on the right-hand side is referred to as "Bayes factor" (Jeffreys 1961), as it is defined in Eq. (20.7). The prior probabilities of the two hypotheses are denoted by $\pi_0 = \Pr(H_0)$ and $\pi_1 = \Pr(H_1)$. Note that $\pi_1 = 1 - \pi_0$ for the binary hypothesis testing problem. Assume that $\pi_0 = \pi_1 = 0.5$ in the absence of prior knowledge of each hypothesis before testing. Thus, Eq. (20.13) becomes:

$$\Pr(H_0|\mathbf{d})\big/\Pr(H_1|\mathbf{d}) = B_i \tag{20.14}$$

where $\Pr(H_1|\mathbf{d})$ represents the posterior probability of the alternative hypothesis (i.e., model is rejected). In this situation, the Bayes factor is equivalent to the ratio of the posterior probabilities of two hypotheses. For a binary hypothesis testing we have $\Pr(H_1|\mathbf{d}) = 1 - \Pr(H_0|\mathbf{d})$. We assume that the prior probabilities of two hypotheses to be $\Pr(H_0) = \pi_0$ and $\Pr(H_1)= 1 - \pi_0$. Without loss of generality, from Bayes' theorem in Eq. (20.13), the *confidence* in the model based on the validation data can thus be quantified as

$$\kappa = \Pr(H_0|\text{Data}) = B_i\pi_0/(B_i\pi_0 + 1 - \pi_0) \tag{20.15}$$

Obviously, from Eq. (20.15), $B_i \to 0$ indicates 0% confidence in accepting the model, and $B_i \to \infty$ indicates 100% confidence. Note that, expert opinion about the model accuracy may be incorporated in the confidence quantification in Eq. (20.15) in terms of prior π_0. If no prior knowledge of each hypothesis (model accuracy) is available, the unbiased assumption $\pi_0 = 0.5$ is used to quantify the model confidence. For the multivariate case, B_i in Eq. (20.15) will be replaced by B_M to quantify the model confidence.

20.7 Bayes Network

Bayesian validation may be realized by the construction of a Bayes network (Jensen and Jensen 2001). A Bayes network (BN) is a directed acyclic (one-way) graphical representation with nodes to represent the random variables and arcs to show the conditional dependencies among the nodes. It is typically specified as a collection of conditional distributions. Data in every node can be used to update the statistics of all other nodes. This property makes the Bayes network a powerful tool to calculate the Bayes factor or validation metric in some complicated situation. The posterior distribution of a validation metric is the output of the Bayes network.

Figure 20.5 shows a typical Bayes network to update the validation decision metric d using specific inspection data. An ellipse (for example, for the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$) represents a random variable (which may be vector-valued), a bold circle represents a deterministic variable (i.e., y), and a rectangle (for example, the inspection data $\mathbf{Y}$) represents the observed data. A solid line arrow represents a conditional probability link, and a dashed line arrow represents the link of a variable to its observed data if available. The probability densities of the model parameter variables ($\boldsymbol{\theta}$ and $\boldsymbol{\phi}$) are updated using the observed data $\mathbf{Y}$. The updated statistics are then used to produce model predictions and to estimate the updated statistics of the decision variable d about the model quality in the validation domain. In addition, model predictions are related to model parameters ($\boldsymbol{\theta}$ and $\boldsymbol{\phi}$) and input variables ($\mathbf{X}$). The Bayes network thus links the relationship coefficients to the validation domain to facilitate two objectives: (1) uncertainty quantification and propagation and (2) decision inference in validation domain.

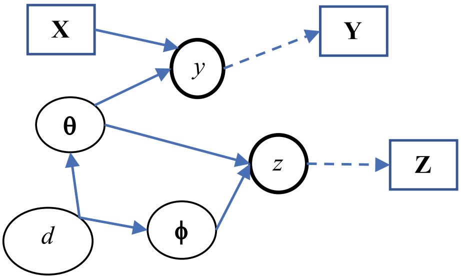


Fig. 20.5 Concept of Bayes network

20.8 Non-normal Data Transformation

So far, the data was required to follow normal distribution. In particular, Bayesian hypothesis testing-based methods for model assessment are based on the assumption that the error follows a normal distribution. This assumption is often violated in practice, which may lead to erroneous decisions on the model validity. In the case of non-normality, an appropriate transformation may be applied to convert the data, and then, a Bayesian metric needs to be derived from the transformed data to quantify the confidence on the model.

Rebba and Mahadevan (2006b) have discussed various methods for non-normality data transformation. These approaches include Rosenblatt transformation (Rosenblatt 1952), Nataf transformation (Nataf 1962), Power and modulus transformations (Box and Cox 1964), and Pericchi's Bayesian method (Pericchi 1981). Each transformation technique may be suitable for a particular application and used according to the researcher's preference. Generally speaking, Rosenblatt transformation requires an actual closed-form conditional distribution of the data, which is almost impossible in many practical cases, particularly for the multivariate case. The Nataf transformation requires that the data is not jointly normally distributed, which may not be accurate in many multivariate cases. Pericchi's Bayesian transformation method requires careful selection of priors for the data, which makes the transformation complicated to implement. In contrast, the power transformation method proposed by Box and Cos (1964) is mathematically tractable, simple to implement and free of strict requirements or significant assumptions. In addition, this method does not require exact closed-form distribution for the response quantity of interest. Therefore, the power transformation is employed to convert the non-normal data (Jiang et al. 2013a).

The power transform is a data preprocessing technique used to create a rank-preserving and normal distribution transformation of raw data by using power functions. For the univariate original data with n points, $\mathbf{d} = (d_1, d_2, \ldots, d_n)$, the Box–Cox transformation is expressed as follows

$$T_{\mathbf{d}} = \begin{cases} [(c + \mathbf{d})^{\lambda} - 1]/\lambda & \lambda \neq 0 \\ \ln(c + \mathbf{d}) & \lambda = 0 \end{cases} \tag{20.16}$$

where $T_{\mathbf{d}}$ represents the transformed data in the univariate case, c is a constant used to ensure that all (c + $\mathbf{d}$) values are positive, and λ is the transformation parameter obtained from maximum likelihood estimation of the data. The Box–Cox transformation can be applied to positive values only. If there are negative values in the data, the data set needs to be shifted by a constant (i.e., c in Eq. 20.16) to ensure that all data is positive. For $\lambda = 0$, the natural log of the data [i.e., ln(.)] is taken in Eq. (20.16). The geometric mean of the data may be used to scale the transformed data. In addition, the likelihood estimate of λ can be obtained easily by maximizing the following function:

$$f(\lambda) = -\frac{n}{2} \times \log\left(s^2\right) + (-1) \times \sum\nolimits_{i=1}^{n} \left[\log(d_i)\right] \tag{20.17}$$

where s^2 is the variance of the transformed data $T_{\mathbf{d}}$.

The transformation shown in Eq. (20.16) may be applied to each variable in multivariate cases in order to marginally transform the non-normal multivariate data into a nearly Gaussian distribution. In order to produce nearly jointly normal multivariate results, a vector of parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ may be defined to transform each of the random variables, where p is the number of variables. The entire vector can be obtained in a single estimation (Andrews et al. 1971) by finding the maximum value of the following function:

$$f(\boldsymbol{\lambda}) = -\frac{n}{2} \times \log(|S|) + \sum\nolimits_{j=1}^{p} \{(j-1) \times \sum\nolimits_{i=1}^{n} \left[\log\left(d_{ij}\right)\right]\} \tag{20.18}$$

where S is the covariance matrix of the transformed random variables $\mathbf{T_D} = \{T_{d1}, T_{d2}, \ldots T_{dp}\}$. Note that in both univariate and multivariate cases, the parameters can be estimated using any standard optimization routine such as steepest descent method, Newton–Raphson method, and genetic algorithm. The transformed data is then used for classical and Bayesian hypothesis testing in both abovementioned univariate and multivariate cases (Jiang et al. 2013a).

20.9 Bayesian Model Validation Process

Figure 20.6 shows the generalized procedure of implementing the Bayesian methodology for model validation of complicated systems with either single or multiple response quantities, which consists of seven main steps (highlighted in Fig. 20.6):

(1) Conduct a graphical comparison between validation data and model prediction. In practice, the widely used method is graphical validation through comparing graphs of prediction and observation.

This is a straightforward qualitative approach where the quantitative uncertainties in both model predictions and experimental results are not considered adequately. In this qualitative approach, a scatter plot of measured versus predicted values on a

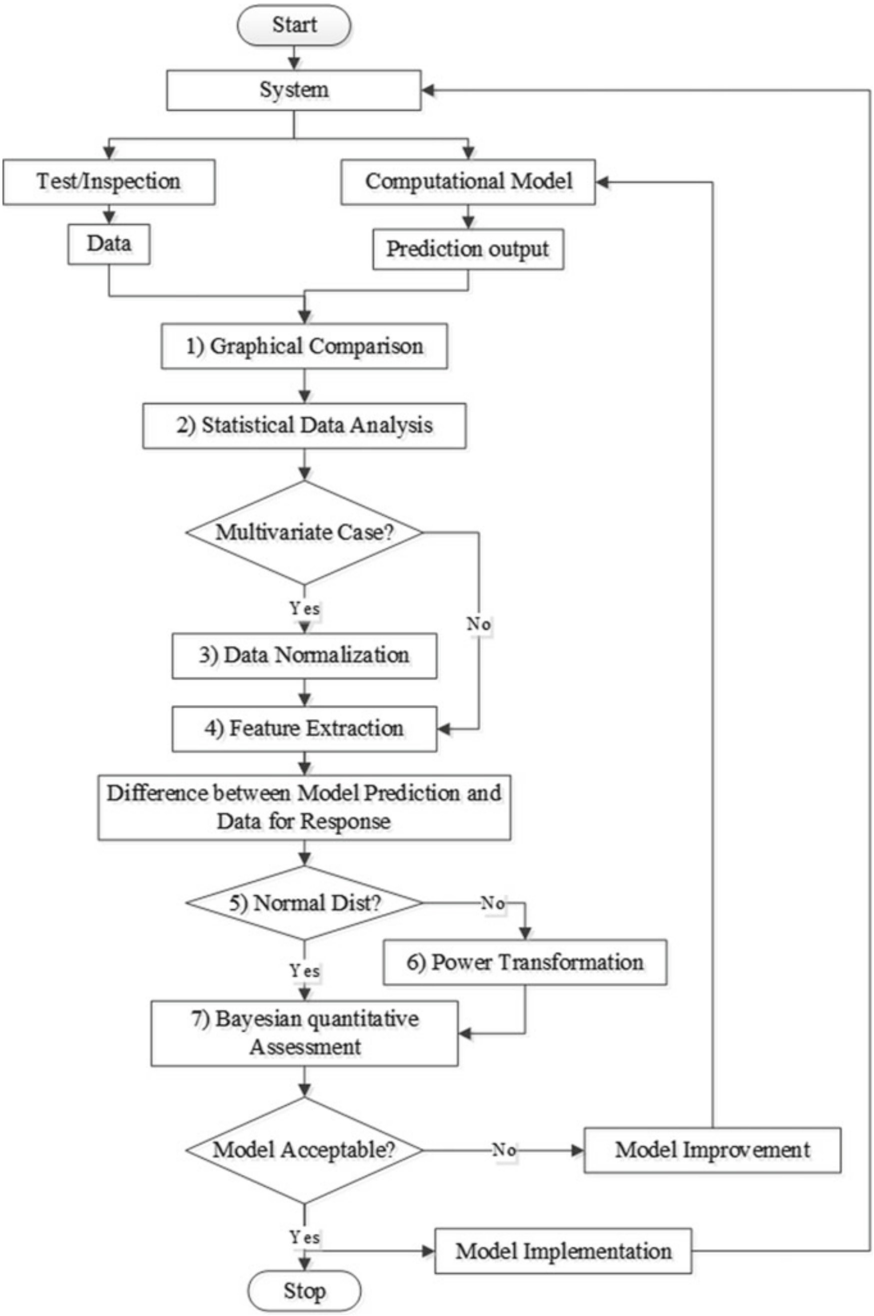


Fig. 20.6 Generalized procedure for Bayesian model validation

1:1 scale is often used. A straight line with a slope of 1 for the plotted data implies a good computer model. Most scatter points should fall within the acceptable boundary for a failure mode. For example, the model may be acceptable if more than 85% of the predicted value should fall within +/−S (S = standard deviation of the actual measurements).

(2) Conduct statistical data analysis to quantify the uncertainty for each variable, whenever repeated data from multiple experiments is available. In case that repeated data is not available, for instance, the survey data from opinion pools, a prior distribution may be assumed for the variable based on the domain expertise.
(3) Standardize each set of data to a dimensionless vector for the multivariate case whenever needed. This step enables multiple variables with different quantities to be compared simultaneously to avoid the duplicate contribution of the same variable to the model validation result.
(4) Extract features such as mean, maximum, frequency, and energy from the validation data to represent the properties of the underlying systems.
(5) Perform normality hypothesis test to verify whether the difference data is normally distributed, i.e., H_0 (null hypothesis): difference data comes from normal distribution versus H_1 (alternative hypothesis): difference data does not come from normal distribution. The Anderson–Darling goodness-of-fit test (Stephens 1974) can be utilized to perform the normality hypothesis test. Compared with other methods such as the chi-square test, this approach is more sensitive to deviations in the tails of the distribution (NIST/SEMATECH 2005). In this approach, an Anderson–Darling (AD) statistic is obtained through calculating the area between the fitted normal distribution curve and the step functions based on the plotted data points. A smaller AD value indicates that the distribution fits the data better. Furthermore, an associated *p*-value is calculated based on the AD value. If the *p*-value is less than or equal to the predetermined *a*-level (commonly 0.05), reject H_0, i.e., the data does not follow a normal distribution; otherwise, accept H_0.
(6) Apply a power transformation to convert non-normal data into Gaussian ones for hypothesis testing when needed. Refer to Jiang et al. (2013a) for more details about the power transformation.
(7) Build statistical hypotheses on the difference between test data and model prediction to assess whether the model is acceptable or not, considering uncertainties in both data sets. For either univariate or multivariate case, the Bayes factor and the confidence measure (Eq. 20.15) are calculated in the Bayesian approach for the model assessment, using the equations described previously.

The obtained quantitative information (e.g., confidence level) is then provided as an indicator to assess the model validity and predictive capacity. If the model is validated with an acceptable confidence level, it may be either applied for design optimization or asset management based on the model type (physics-based or data-driven models) or implementation purpose. For example, a physics-based computer model may be developed for design optimization to improve hardware operational performances with reduced weights, while reliability models are widely implemented for condition-based maintenance and asset management of high-value components in an engineering system. On the other hand, if the model is unacceptable, it can then be either rejected or proposed for improvement, and the validation process is repeated until satisfactory model accuracy is achieved.

20.10 Numerical Application

The Bayesian validation methodology and procedure presented in Fig. 20.6 are demonstrated with three examples. The first example is to validate the reliability model for high cycle fatigue of rotor blades in aircraft engines (Jiang and Mahadevan 2007). The second example is a univariate damage accumulation model, which was established by using real-world inspection data from subway tunnels (Jiang et al. 2013b). The third example is a multivariate heat conduction problem developed by Sandia National Laboratories (Dowding et al. 2008) as a model validation challenging problem.

20.10.1 Example 1: Bayesian Decision Rule

20.10.1.1 Problem Description

The dynamic loading and material properties of the rotor blades in aircraft engines are usually random variables. The failure probability of a single blade under high cycle fatigue can be estimated by a limit state-based reliability prediction model. The blade is assumed to have failed when the actual maximum displacement under dynamic loading exceeds the design or allowable maximum displacement. Generally, the blade is modeled as a single-degree-of-freedom oscillator and its dynamics is described by a differential equation consisting of mass, spring, dashpot and with displacement x as follows (Annis 2002)

$$F \sin \omega t = m\ddot{x} + c\dot{x} + kx \tag{20.19}$$

where F = magnitude of the external harmonic load, ω = applied load frequency, m = mass of the oscillating body, c = damping constant, and k = stiffness of the spring.

The displacement is computed as

$$x = \frac{F}{k} \frac{1}{\sqrt{[1 - (\omega/\omega_n)^2]^2 + [2(c/c_c)(\omega/\omega_n)]^2}} \tag{20.20}$$

Table 20.1 Statistics of variables (Example 1)

Variable	Type	Mean	Std Dev
ω_{nom}	Normal	2194 rpm	105 rpm
N	Normal	8800 rpm	50 rpm
b_{nom}	Normal	0.1	0.005
ξ	Lognormal	0.002	0.0005
λ_l	Normal	1	0.25
λ_m	Normal	1	0.05
D_a	Normal	15 mm	0.75 mm
α_n	Deterministic	5.7427	–
p_d	Deterministic	25	–
d_n	Deterministic	15 mm	–

where ω_n = natural frequency and c_c = critical damping factor. The performance function for the failure of the blade is thus a function of the natural frequency, damping, load factor and engine speed (all random variables):

$$g = 1 - \left[\frac{p_d}{100} \cdot \lambda_l \lambda_m \cdot \frac{\alpha}{\alpha_n} \cdot \frac{d_n}{D_a} \right] \tag{20.21}$$

where λ_l and λ_m are the load and modal shape factors, respectively, α and α_n are the design and nominal amplification factors, respectively, d_n is the allowable design displacement or the nominal allowable displacement, D_a is the allowable displacement, and p_d is the percentage of nominal allowable displacement. The limit state is denoted by the condition $g = 0$. Table 20.1 gives the statistics of variables used in this example. Refer to Jiang and Mahadevan (2007) for the calculation of the design amplification factor α and the natural frequency ω_n, as well as the validation data. The Bayesian rule approach is applied to validate the reliability of the model of the rotor blade.

20.10.1.2 Bayes Network

In order to calculate the Bayes factor, the series of quantities involved in computing the performance function g was modeled by Mahadevan and Rebba (2005) as different nodes in a Bayes network. With the availability of every experimental validation outcome of each node and the statistics of each node, the statistical distribution function associated with all the nodes in the network, including g, could be updated. Then the likelihood ratio of g is calculated (Jiang and Mahadevan 2007). In this study, assume that the measured values for two quantities ω_n (2220 rpm and 3316 rpm) and β (0.6, 0.87, and 0.9) are available. Their corresponding predicted values are 2200 rpm and 0.9. Different combinations of the validation data and the

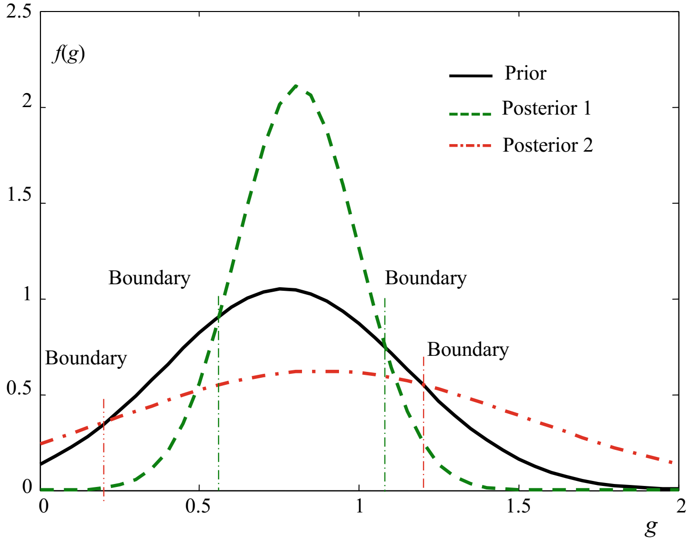


Fig. 20.7 Prior and posterior distributions of g with different validation data (Example 1)

corresponding likelihood ratio values for the overall reliability model prediction can be obtained. Figure 20.7 shows the prior and posterior densities of g resulting from Bayesian updating with validation data on single and multiple nodes. The decision regions can be easily identified through the decision boundaries with the decision threshold $\eta = 1$.

In order to calculate the Bayesian risk, a Bayesian network (Fig. 20.8) is designed in this study to update the likelihood of g [i.e., $f_0(g) = L(Y|y)$], given the measured values of ω_n and/or β. In this figure, an ellipse (for example, α) represents a random variable and a rectangle (for example, the experimental value β_{exp}) represents a constant value. The double line arrow represents a logical relationship link between two variables (computational formula) and a single line arrow represents a direct probabilistic relationship link. After any data node is added to the network, the posterior probability densities of all the nodes are computed, including g_{exp} (the experimental value of g). The statistics of the parameters shown in Table 20.1, the computational models presented in Eqs. (20.19) and (20.21), and 10,000 iterations of simulation are used in the Bayesian updating. The likelihood values for the overall reliability model prediction in different combinations of the validation data are obtained and subsequently used to perform the risk analyses for model validation in the case of $\eta = 2$.

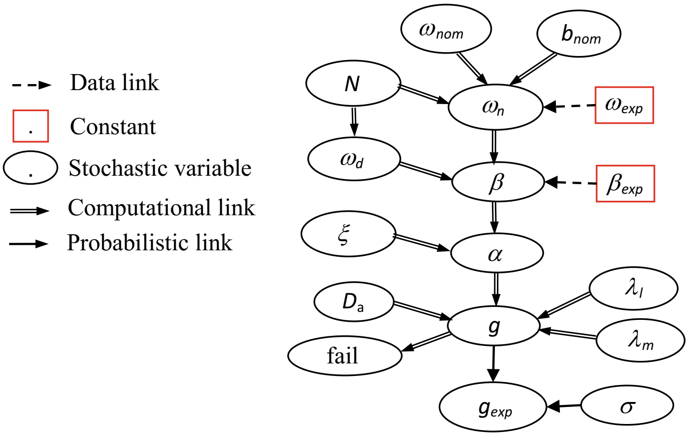


Fig. 20.8 Bayes network for updating the PDF of g (Example 1)

20.10.1.3 Case $\eta = 2$

Assume that the costs of deciding in favor of H_i when H_j is true are $\mathrm{C}_{00} = 1$, $\mathrm{C}_{11} = 1.2$, $\mathrm{C}_{01} = 2$, and $\mathrm{C}_{10} = 1.4$ (unit), and the prior information about the hypotheses still are $\pi_0 = \pi_1 = 0.5$. Thus, the threshold $\eta = 2$ is obtained from Eq. (20.7) and only the second and third experiments support the model. The 256 possible risk values are computed again using Eq. (20.21) and plotted in Fig. 20.9. The minimum Bayes risk is $R_{\min} = 8.168$ units when two experimental outputs whose likelihood ratios are larger than η $(= 2)$ are assigned to the decision region Z_0, while the other six experimental outputs whose likelihood ratios are less than η are assigned to the decision region Z_1. The maximum Bayes risk is $R_{\max} = 9.794$ units when all eight experimental outputs are wrongly assigned to the decision regions (i.e., two experiments that support the model are wrongly assigned to Z_1 and six experiments that do not support the model are assigned to Z_0).

It should be pointed out that the decision threshold η depends on both the cost information C_{ij} and the prior of each hypothesis π_i. Only the cost information is changed in this example for the purpose of illustration. However, it is easy to incorporate the engineers' or decision makers' preferences about both the cost and prior information in the threshold.

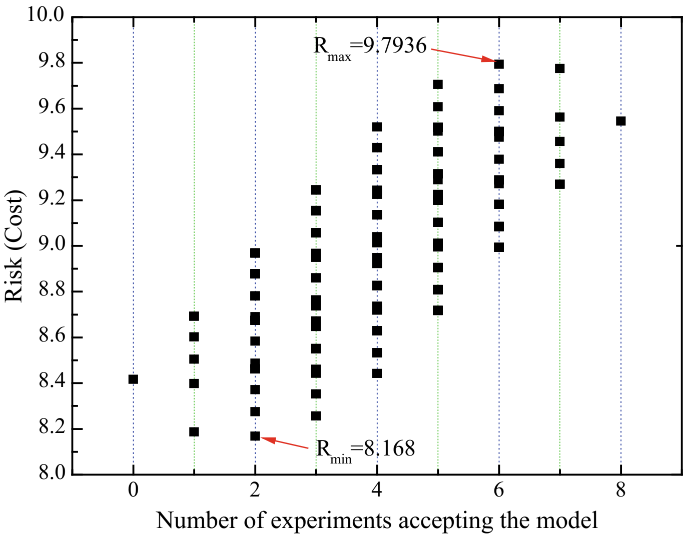


Fig. 20.9 Bayesian risk-based decision analysis for validation of rotor blade high cycle fatigue reliability model (Example 1)

20.10.2 Example 2: Univariate Model Validation

20.10.2.1 Model Description

The damage accumulation model was established from seepage data inspected on six real-life subway lines (Jiang et al. 2013b). Each subway line consists of various numbers of tubes between stations and gives rise to two sets of inspection data according to the upward and downward operating directions. In addition, there are three different types of tunnels: single-track, double-track, and cross-looping. In this example, the model output is the seepage failure ratio per kilometer, denoted by z, which is defined by the number of observed seepages divided by each segment length. The seepage damage predictive model is defined as a function of the tunnel type, operating direction, and operating duration in days, denoted by x_1, x_2, and t, respectively. Thus, the problem becomes modeling the failure quantity relationship $z = f(x_1, x_2, t)$. The tunnel type and operating directions are dealt with as deterministic variables, while the operating duration is treated as a random variable. Obviously, the tunnel duration calculation is not precise because not each segment in a given tunnel is operated in every moment of the day. Refer to Jiang et al. (2013b) for more information about this problem.

It is well recognized that many influencing factors, such as geographical condition, construction material and type, operating duration or age, as well as the surrounding condition of the subway tunnel, impact the life of a subway tunnel. This example is simplified for demonstration purpose due to the lack of adequate data about the possible influencing factors mentioned above.

A Weibull proportional hazard model (or Weibull PHM) was established to model the distribution of seepage damage as a function of tunnel type, operating direction, and operating duration in days, expressed as

$$F(l) = 1 - \exp\left[-\left(l/\eta\right)^{\beta}\right] = 1 - \exp\left[-\left(l/g_1(\mathbf{X},\boldsymbol{\theta})\right)^{g_2(\mathbf{X},\boldsymbol{\varphi})}\right] \quad (22)$$

where the scale and shape parameters, η and β, respectively, are expressed as an exponential function of vital factors (X's) and coefficients. Refer to Jiang et al. (2013b) for details about data preprocessing, model selection, parameter estimation, and model prediction. In this paper, additional 17 data points collected from 2 subway tunnel segments, denoted as Segment 1 and Segment 2, are used to assess the model validity by using univariate Bayesian hypothesis testing approach.

20.10.2.2 Bayesian Hypothesis Testing

The mean and standard deviation of original difference values (using non-normal data) for 17 validation data points, $\mu_{\mathrm{raw}} = 6.999$ and $\sigma_{\mathrm{raw}} = 29.05$, respectively, are obtained, assuming $\rho = 0$ and $\tau = \sigma_{\mathrm{raw}}$ and $\tau = \sigma_{\mathrm{trans}}$ for raw data and transformed data by using the Box–Cox transformation, respectively. Assume that the uncertainty in the validation data $\varepsilon = \sigma_{\mathrm{val}}/4$ (= 7.25) is taken for the interval hypothesis testing. From Eq. (20.8), we obtain the Bayes factor metrics in logarithm scale $b_{\mathrm{i\_raw}} = -1.7069$ and $\mathrm{B}_{\mathrm{i\_trans}} = 1.9027$, for the raw and transformed data, respectively. Note that the subscripts i_raw and i_trans represent the two cases using raw and transformed data, respectively. From Eq. (20.15), the probabilities of accepting the model are $\kappa_{\mathrm{i\_raw}} = 15.4\%$ and $\kappa_{\mathrm{i\_trans}} = 87.0\%$ for the two cases, with the assumption of $\pi_0 = 0.5$. Clearly, the model is rejected (i.e., $\mathrm{b} < 0$) when the non-normal data is used for model validation via the Bayesian interval hypothesis testing, but accepted (i.e., $\mathrm{b} > 0$) with high confidence when the transformed data is used. As a result, without the normality transformation, the Bayesian approach may produce inaccurate results for decision making when the raw, non-normal data is used in the model validation.

To investigate the effect of ε (threshold or tolerance) on model validity, various ε values are used in both classical and Bayesian hypothesis testing approaches. Figure 20.6 shows the variation of validation metrics (p-value in the classical hypothesis testing denoted by standard line or confidence in the Bayesian approach denoted by bold line) versus ε in the hypothesis testing by using raw non-normal data (dashed line) and transformed data (solid line). It is observed in Fig. 20.10 that the p-value > 0.05 (i.e., no evidence to reject H_0) is always met if a value of $\varepsilon > 0.7$ is used in classical hypothesis testing for both raw data and transformed data. Obviously, the model should not be judged to be rejected based on the classical hypothesis results in this example.

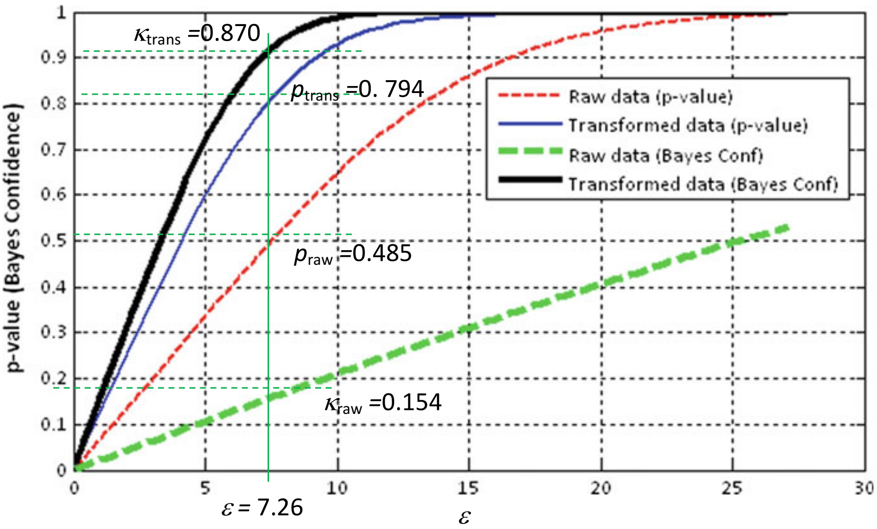


Fig. 20.10 Effect of various ε values on validation results (Example 2)

It is also observed from Fig. 20.10 that the Bayesian validation metric appears to be more sensitive to the normality of data than the classical approach. The validation metric obtained from both the classical (p-value) and Bayesian (accepting confidence) approaches provides consistent decision support if normal data is used. However, a larger value of ε is needed to accept the model if the non-normal data is inappropriately used in the Bayesian approach. As such, the Bayesian confidence/validation metric may provide erroneous conclusion on model validity if non-normal data is used in the validation.

20.10.3 Example 3: Multivariate Model Validation

20.10.3.1 Problem Description

A transient heat conduction example specified by Sandia National Laboratories in the United States (Dowding et al. 2008) as a validation challenge problem is used to demonstrate the effectiveness of multivariate Bayesian model validation. The one-dimensional heat conduction through a slab is described by a set of governing differential equations (Dowding et al. 2008). The corresponding analytical solution is approximated by a truncated infinite series. The model output is the temperature at a given spatial location and the instance of time. The inputs to the model include the initial temperature, heat flux, slab thickness, thermal conductivity, and heat capacity.

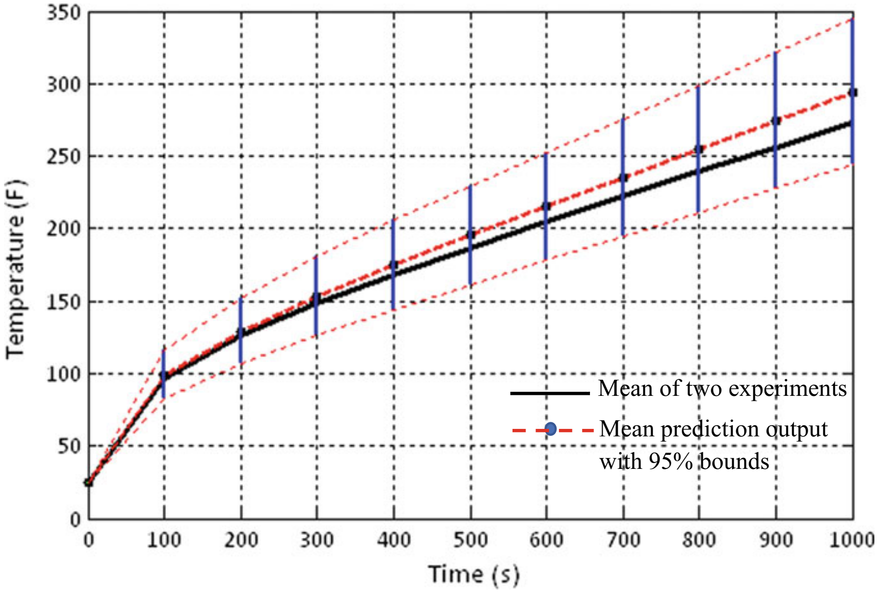


Fig. 20.11 Model prediction versus experiment results at a given configuration (Example 3)

Refer to Dowding et al (2008) and Jiang and Mahadevan (2008a) for details of this heat conduction problem.

For illustrative purposes, 21 experimental observations at a specific configuration at the time increment of 50 s over the period of 0–1000 s are used in this example. The measurements were taken at discrete, regular time intervals to provide multivariate data. Given a fixed spatial point, the predicted temperature is generated using the analytic expression with the random input parameters and various instances of time. The uncertainty in the input parameters is propagated to the response output through the approximate model repeatedly to obtain the output statistics. Thus, this example serves as a case study for multivariate model validation with observations at different spatial and temporal points for a single response quantity. Note that, the number of variables is equal to the number of observation time intervals, i.e., $m = 10$, for a given configuration, resulting in a 10-variable model validation problem (the initial temperature is treated as a constant).

Figure 20.11 shows the curves of the mean actual observation and mean value of 4000 predicted results. It is observed that the mean actual observation (dashed line) falls in the region of simulated prediction output with 95% bounds. The visual comparison appears to demonstrate that the mean values of model prediction are not too distinct from those of experimental measurements.

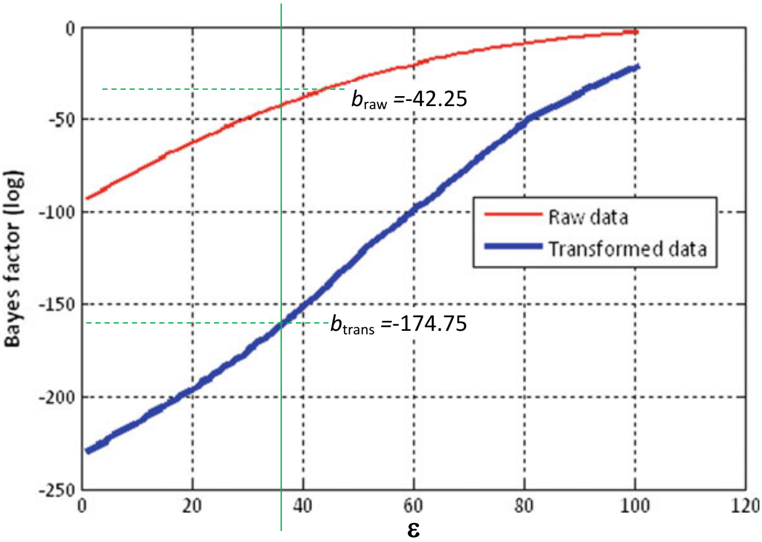


Fig. 20.12 Effect of various **ε** values on validation results (Example 3)

20.10.3.2 Bayesian Hypothesis Testing

We assume $\boldsymbol{\rho} = \mathbf{0}$ and $\boldsymbol{\varepsilon} = 0.5 \times \sqrt{diag(\Sigma)}$ as well as $\Lambda = \Sigma_{\text{raw}}$ and $\Lambda = \Sigma_{\text{trans}}$ for raw data (non-normal) and transformed data, respectively. From Eq. (20.9) we obtain the Bayes factor metrics $B_{\text{i\_raw}} = -42.25$ and $B_{\text{i\_trans}} = -174.75$, for the raw and transformed multivariate data, respectively. From Eq. (20.15), the probabilities of accepting the model are $\kappa_{\text{i\_raw}} = 4.49 \times 10^{-17}$ and $\kappa_{\text{i\_trans}} = 1.74 \times 10^{-75}$ for the two cases, with the assumption of $\pi_0 = 0.5$. Clearly, the model is rejected (i.e., $b < 0$ and $\kappa \to 0$) in case either the non-normal or the transformed data is used for model validation via the Bayesian interval hypothesis testing. It is also noted that, with the normality transformation, the Bayesian approach provides higher confidence on model rejection.

To investigate the effect of the value of **ε** on model validity, again various **ε** values are used in the Bayesian hypothesis testing approach. Figure 20.12 shows the variation of the validation metrics versus ε in the hypothesis testing by using raw non-normal data (dashed line) and transformed data (solid line). Note that, the logarithm of the Bayes factor is used for the convenience of comparison. There are two observations obtained from Fig. 20.12. First, regardless of the **ε** value selected, the condition that the logarithm of the Bayes factor be less than zero is always met for both raw and transformed data, implying that the model should be rejected based on the Bayesian hypothesis results. Second, with the Box–Cox transformation applied to the non-normal data, the logarithm of Bayes factor obtained from the Bayesian hypothesis testing approach provides stronger confidence on model rejection (smaller value).

20.11 Concluding Remarks

This chapter has presented Bayesian methods to quantitatively evaluate the validity and predictive capacity of computer models and simulations, considering data uncertainty. Explicit expressions for the interval hypothesis testing-based Bayes factor were presented for univariate and multivariate cases. The Bayesian confidence measure was defined in order to quantify the confidence on the predictive capacity of the computer model. A generalized procedure is presented to implement the Bayesian methodology for model validation of complicated systems with either single or multiple non-normality response quantities. The impact of the data normality assumption and of the decision threshold parameter in quantitative model assessment was illustrated by using Bayesian approaches. The Bayesian methodology was illustrated with a reliability model, a univariate stochastic damage accumulation model, and a multivariate heat conduction problem. Bayesian methods provide a powerful tool for quantitative model validation in order to further calibrate a computer model and simulation for practical implementation such as condition-based maintenance and asset management of complicated systems.

In the future, more research is needed on the development of Bayesian validation methods for more complicated multi-input-multi-output problems. In addition to the normality transformation of the non-Gaussian difference data, a Bayesian hypothesis testing method may be mathematically derived from a non-Gaussian distribution of data, by using for instance, a kernel density function instead. Furthermore, Bayesian validation methods should be seamlessly integrated with calibration of model parameters to produce more accurate prediction results, considering both data and modeling uncertainties.

References

AIAA. (1998). *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*, American Institute of Aeronautics and Astronautics, AIAA-G-077-1998, Reston, Virginia, USA.

Andrews, D. F., Gnanadesikan, R., & Warner, J. L. (1971). Transformations of multivariate data. *Biometrika, 27*(4), 825–840.

Annis, C. (2002). Modeling high cycle fatigue with Markov chain Monte Carlo-a new look at an old idea. In: *Proceedings of 43rd AIAA/ASME/ASCE/AHS/ASC Structure, Structural Dynamics and Materials Conference* (AIAA-2002-1380) (pp. 1351–1361). Denver CO 2002.

ASME. (2002). *Verification & Validation (V&V) methodology and quantitative reliability at confidence: Basis for an investment strategy*, PTC 60, UCRL-ID-150874, Livermore, CA: Lawrence Livermore National Laboratory.

ASME. (2006). *Guide for verification and validation in computational solid mechanics.* ASME V&V 10-2006. New York, NY: American Society of Mechanical Engineers.

Babuska, I., & Oden, J. T. (2004). Verification and validation in computational engineering and science: Basic concepts. *Computer Methods in Applied Mechanics and Engineering, 193*(36–38), 4057–4066.

Balci, O., & Sargent, R. G. (1981). A methodology for cost-risk analysis in the statistical validation of simulation models. *Communications of Association for Computing Machinery (ACM), 24*(11), 190–197.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science, 2*(3), 317–352.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, 26*(2), 211–252.

Chen, W., Baghdasaryan, L., Buranathiti, T., & Cao, J. (2004). Model validation via uncertainty propagation and data transformations. *AIAA Journal, 42*(7), 1406–1415.

Chen, W., Xiong, Y., Tsui, K. -L., & Wang, S. (2008). A design-driven validation approach using Bayesian prediction models. *Journal of Mechanical Design*, *130*(2), 021101-1-12.

DOD. (1996). *Verification, validation, and accreditation (VV&A) recommended practices guide*. Alexandria, VA: Department of Defense.

DOE. (2000). *Accelerated strategic computing initiative (ASCI) program plan*. DOE/DP-99-000010592, Washington, DC: Department of Energy.

Dowding, K. J., Pilch, M., & Hills, R. G. (2008). Formulation of the thermal problem. *Computer Methods in Applied Mechanics and Engineering (special issue), 197*(29–32), 2385–2389.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics, 1*(2), 141–149.

Hills, R.G., & Trucano, T. G. (2002). *Statistical validation of engineering and scientific models: A maximum likelihood based metric*. Technical Report Sand. No 2001-1783, Albuquerque, NM: Sandia National Laboratories.

Hwang, J. T., Casella, G., Robert, C., Wells, M. T., & Farrell, R. H. (1992). Estimation of accuracy in testing. *The Annals of Statistics, 20*(1), 490–509.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). London, UK: Oxford University Press.

Jensen, F. V., & Jensen, F. B. (2001). *Bayesian networks and decision graphs*. New York: Springer.

Jiang, X., & Mahadevan, S. (2007). Bayesian risk-based decision method for model validation under uncertainty. *Reliability Engineering and System Safety, 92*(6), 707–718.

Jiang, X., & Mahadevan, S. (2008a). Bayesian wavelet method for multivariate model assessment of dynamic systems. *Journal of Sound and Vibration, 312*(4–5), 694–712.

Jiang, X., & Mahadevan, S. (2008b). Bayesian validation assessment of multivariate computational models. *Journal of Applied Statistics, 35*(1), 49–65.

Jiang, X., & Mahadevan, S. (2009a). Bayesian inference method for model validation and confidence extrapolation. *Journal of Applied Statistics, 36*(6), 659–677.

Jiang, X., & Mahadevan, S. (2009b). Bayesian structural equation modelling method for hierarchical model validation. *Reliability Engineering and System Safety, 94*(4), 796–809.

Jiang, X., & Mahadevan, S. (2010). Bayesian nonlinear SEM approach for hierarchical validation of dynamical systems. *Mechanical Systems and Signal Processing, 24*(4), 957–975.

Jiang, X., Yuan, Y., Mahadevan, S., & Liu, X. (2013a). An Investigation of Bayesian inference approach to model validation with non-normal data. *Journal of Statistical Computation and Simulation, 83*(10), 1829–1851.

Jiang, X., Yuan, Y., & Liu, X. (2013b). Bayesian inference method for stochastic damage accumulation modeling. *Reliability Engineering and System Safety, 111*(3), 126–138.

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer experiments. *Journal of the Royal Statistical Society Series B (Statistical Methodology), 63*(3), 425–464.

Mahadevan, S., & Rebba, R. (2005). Validation of reliability computational models using Bayes networks. *Reliability Engineering and System Safety, 87*(2), 223–232.

Migon H. S., & Gamerman, D. (1999). *Statistical inference-an integrated approach*. London: Arnold, a Member of the Holder Headline Group.

Nataf, A. (1962). Détermination des distributions de probalités dont les marges sont données. *Comptes Rendus de l'Académie des Sciences, 225,* 42–43.

NIST/SEMATECH (2005). *e-Handbook of statistical methods*, National Institute of Standards and Technology. Retrieved from http://www.itl.nist.gov/div898/handbook.

Nowak, R., & Scott, C. (2004). The Bayes risk criterion in hypothesis testing. *Connexions*, Retrieved from http://cnx.rice.edu/content/m11533/1.6/.

Oberkampf, W. L., & Barone, M. F. (2006). Measures of agreement between computation and experiment: Validation metrics. *Journal of Computational Physics, 217*(1), 5–36.

Oberkampf, W. L., & Trucano, T. G. (2002). Verification and validation in computational fluid dynamics. *Progress in Aerospace Sciences, 38*(3), 209–272.

Pericchi, L. R. (1981). A Bayesian approach to transformations to normality. *Biometrika, 68*(1), 35–43.

Rebba, R., & Mahadevan, S. (2006a). Model predictive capability assessment under uncertainty. *AIAA Journal, 44*(10), 2376–2384.

Rebba, R., & Mahadevan, S. (2006b). Validation of models with multivariate outputs. *Reliability Engineering and System Safety, 91*(8), 861–871.

Rebba, R., & Mahadevan, S. (2008). Computational methods for model reliability assessment. *Reliability Engineering and System Safety, 93*(8), 1197–1207.

Roache, P. J. (1998). *Verification and validation in computational science and engineering in: Science and engineering*. Albuquerque, NM: Hermosa Publishers.

Rosenblatt, M. (1952). Remarks on multivariate transformation. *Annals of Mathematical Statistics, 23*(3), 470–472.

Schlesinger, S. (1979). Terminology for model credibility. *Simulation, 32*(3), 103–104.

Schervish, M. J. (1995). *Theory of statistics*. New York: Springer.

Schwer, L. E. (2007). Validation metrics for response histories: Perspectives and case studies. *Engineering with Computers, 23*(4), 295–309.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association, 69*(347), 730–737.

Zhang, R., & Mahadevan, S. (2003). Bayesian methodology for reliability model acceptance. *Reliability Engineering and System Safety, 80*(1), 95–103.

Chapter 21
Imprecise Probabilities

Seamus Bradley

Abstract This chapter explores the topic of imprecise probabilities (IP) as it relates to model validation. IP is a family of formal methods that aim to provide a better representation of severe uncertainty than is possible with standard probabilistic methods. Among the methods discussed here are using sets of probabilities to represent uncertainty, and using functions that do not satisfy the additvity property. We discuss the basics of IP, some examples of IP in computer simulation contexts, possible interpretations of the IP framework and some conceptual problems for the approach. We conclude with a discussion of IP in the context of model validation.

Keywords Imprecise probabilities · Lower previsions · Credal sets · Formal epistemology · Computer simulation

21.1 Introduction

Model validation is an important aspect of quality control when modelling some phenomenon about which we are uncertain. So, accommodating and representing uncertainty is of central importance to model validation. Probability theory provides the standard suite of tools for dealing with uncertainty, but this theory has its limits. For example, models will often contain parameters whose true value we don't actually know. Now, we can't run a simulation without providing a value for this parameter, so for each simulation we run, we must pick *some* value. If we sample this value from a distribution, and run several simulations—sampling from this distribution each time—we can, to some extent, accommodate uncertainty in the parameter value. In doing so, however, we are assuming that a certain sort of distribution is the "right" distribution to be sampling from. If the parameter fluctuates randomly and

S. Bradley (✉)
TiLPS, Tilburg University, Tilburg, Netherlands
e-mail: s.c.bradley@leeds.ac.uk

S. Bradley
University of Leeds, Leeds, UK

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_21

we have data on the distribution of fluctuations, perhaps a particular distribution can be justified. If not, it is typical to pick some "non-committal" distribution that will not skew the results of the simulation. For example, if bounds can be put on the range of values the parameter can take, a uniform distribution is often selected. Now, a uniform distribution seems innocuous, non-committal, but the distribution's being uniform for some parameter means that distributions for related parameters are not uniform. Ferson and Ginzburg (1996) give the example of two independent uniformly distributed parameters that give rise to a non-uniformly distributed product. Or consider two inversely related parameters (like "ice fall rate in clouds" and "ice residence time in clouds"): if one is uniformly distributed, the other is not. This is, in essence, the problem that Joseph Bertrand pointed out at the end of the nineteenth century that is today known as "Bertrand's paradox."

The practice of sampling unknown parameters from distributions chosen for convenience rather than for empirically grounded reasons is a necessary aspect of standard modelling practice. Imprecise Probabilities is an approach that attempts to mitigate some of the problematic consequences of such a methodology. This chapter will outline the basic idea of IP, give some examples of IP in modelling contexts, discuss how we might interpret the IP framework and point to some potential problems for IP. We will conclude with a discussion of IP in the context of validation.

21.2 Basics

The core idea of Imprecise Probabilities (IP) is to represent uncertainty using a *set* of probability measures rather than a single such measure (although there are a great many related formal models that we'll discuss later in this section).[1] The basics of uncertainty quantification is introduced in Chap. 5 by Roy and Chap. 22 by Dougherty et al., and the probabilistic/Bayesian approach to uncertainty are discussed in Chap. 7 by Beisbart and Chap. 20 by Jiang et al. in this volume, so let's jump straight to the basic idea of IP. We use **pr** to signify a probability function. The basic idea of IP is that we represent uncertainty, not by a single such function, but by a set of them—$\mathscr{P}$ – defined over the same state space. If X is an event over which the **pr**s are defined, then we can let $\mathscr{P}(X) = \{\mathbf{pr}(X), \mathbf{pr} \in \mathscr{P}\}$. That is, we can take $\mathscr{P}(-)$ to be a set-valued function that returns the set of values assigned to X by members of $\mathscr{P}$. We can then apply the rest of the Bayesian machinery "pointwise." So conditionalising $\mathscr{P}$ involves conditionalising on each member of $\mathscr{P}$ and taking the resultant set of

[1] Although one can find precedents going back to Keynes or even Boole, IP really started in the middle of the twentieth century with work by people like Koopman (1940), Good (1952, 1962), Smith (1961) and Dempster (1967). Work in philosophy on IP really starts with Levi (1974, 1980, 1986). Important formal and philosophical work on IP was carried out by Seidenfeld (1983, 1988); Seidenfeld et al. (1989) (Seidenfeld was Levi's graduate student). Walley (1991) was a hugely influential book which, until recently, was still the go-to monograph for many formal details of the theory. The state of the art in terms of formal theory can be found in Augustin et al. (2014) and Troffaes and de Cooman (2014). Bradley (2014) provides a philosophical overview.

conditional probabilities. Expected values also become sets of values, determined pointwise for each $\mathbf{pr} \in \mathscr{P}$.

Recall that a probability function $\mathbf{pr}$ is a real-valued function on an algebra of events that has the following properties:

Bounded For all X, $0 = \mathbf{pr}(\bot) \leq \mathbf{pr}(X) \leq \mathbf{pr}(\top) = 1$ (where $\bot$ and $\top$ are the bottom and top elements of the algebra, respectively)
Superadditive If $X \wedge Y = \bot$ then $\mathbf{pr}(X \vee Y) \geq \mathbf{pr}(X) + \mathbf{pr}(Y)$
Subadditive If $X \wedge Y = \bot$ then $\mathbf{pr}(X \vee Y) \leq \mathbf{pr}(X) + \mathbf{pr}(Y)$

A function that satisfies the first and second of these properties is called a "lower probability" and a function that satisfies the first and third is called an "upper probability."

The idea of a set of functions and the idea of a lower probability are intimately related. Take a set of probabilities, $\mathscr{P}$ and define $\underline{\mathscr{P}}(X) = \inf \mathscr{P}(X)$, the lowest value assigned to X by some member of $\mathscr{P}$. This function $\underline{\mathscr{P}}$ is a lower probability.[2] Likewise, $\overline{\mathscr{P}}(X) = \sup \mathscr{P}(X)$ is an upper probability. Moreover, $\underline{\mathscr{P}}(\neg X) = 1 - \overline{\mathscr{P}}(X)$.

And going the other way, take a lower probability $\mathbf{lpr}$, and define the associated credal set of $\mathbf{lpr}$ as the set of probability functions such that $\mathbf{pr}(X) \geq \mathbf{lpr}(X)$ for all X, the set of probability functions that "pointwise dominate" it. If $\mathbf{lpr}$ is a lower probability as defined above, such a set is non-empty; let $M(\mathbf{lpr})$ be the associated credal set of $\mathbf{lpr}$. Since $M(\mathbf{lpr})$ is a set of probabilities, we can take the "lower envelope" as we did above: $\underline{M(\mathbf{lpr})}(X)$. The lower envelope theorem entails that $\mathbf{lpr}$ is a lower probability if and only if $\underline{M(\mathbf{lpr})}(X) = \mathbf{lpr}(X)$ for all X (see Sect. 3.3 of Walley 1991 or Sect. 2.2.2 of Augustin et al. 2014). Note that distinct credal sets might result in the same lower probability.[3]

As well as credal sets and lower probabilities, there is a huge range of other related formal methods for representing uncertainty (see, for example, Halpern 2003; Augustin et al. 2014; Klir and Smith 2001). For example, Dempster–Shafer theory (sometimes called Evidence Theory) uses a *belief function* which is a lower probability with the further property of being infinite monotone (a sort of strengthening of superadditivity). DS theory comes equipped with a slightly different interpretation (see Sect. 21.4) and an alternative kind of updating/aggregation rule.[4]

This brief discussion merely scratches the surface of the rich and interesting theory of IP. Many aspects of the statistical method have been replicated inside the IP framework including statistical inference, graphical models (e.g. Bayes nets) and

[2] Consider some $\mathbf{pr} \in \mathscr{P}$ for which $\mathbf{pr}(X \vee Y) = \underline{\mathscr{P}}(X \vee Y)$. $\inf \mathscr{P}(X) + \inf \mathscr{P}(Y) \leq \mathbf{pr}(X) + \mathbf{pr}(Y)$, since $\mathbf{pr} \in \mathscr{P}$, so $\underline{\mathscr{P}}$ is superadditive. Boundedness is trivial, and much the same reasoning works if the set $\mathscr{P}$ doesn't attain its bounds (just think in terms of the closure of the set).

[3] There is a one-to-one correspondence between lower probabilities and a subset of the set of credal sets, namely those with some nice topological properties. We don't need to discuss this here, but see the above-listed references for details.

[4] See Oberkampf and Helton (2004) for a discussion of DS theory in an engineering context.

stochastic models (e.g. Markov chains).[5] See, for example Augustin et al. (2014), Troffaes and de Cooman (2014).

21.3 Examples

In this section, we'll explore two examples of IP-like ideas that show up when attempting to model physical systems on computers.

21.3.1 Unknown Parameters

Oberkampf and Roy (2010) discuss an example of how IP arises in scientific computing.[6] They start from the position of wanting to keep apart *epistemic uncertainty* – uncertainty arising from things unknown to the experimenter—and *aleatory uncertainty*—randomness or natural variability.[7] Now, whether some kind of uncertainty counts as one or the other of these kinds is somewhat a matter of perspective, the distinction is important. Aleatory uncertainty about a parameter can be accommodated by having a probability distribution over that parameter in the model. Epistemic uncertainty, on the other hand, is captured by having a set of such distributions, i.e. having a credal set.

The idea is that if a parameter is subject to aleatory uncertainty, you can sample values for that parameter (using the given distribution), run the simulation using those sampled values and then take the distribution of outcome values as telling you something about uncertainty in the outcomes. However, with epistemic uncertainty, you must pick specific values of the parameter to run through the model (and you must pick them using *some* distribution). You can't take the distribution of outputs as telling you about the uncertainty in the outcomes: you can only take the *range* of outcome values as telling you what ranges of values of outcome values are possible given your uncertainty about the parameter. Or perhaps, a more careful way to phrase the same thing: the distribution of output values might be in part due to the choice of input distribution for the unknown paramaters. If that distribution were chosen merely for convenience, then we had better not read too much into the output distribution. As Oberkampf and Helton (2004) say:

> If extreme system responses correspond to extreme values of these parameters (i.e. values near the ends of the uniform distribution), then their probabilistic combination could predict

[5] For introductions to these aspects of IP, see Augustin et al. (2014) Chapters 7, 9 and 11, respectively.

[6] I am drawing mainly from Sect.13.4, but similar ideas appear in a number of other places in the book.

[7] This is one dimension of the many ways one can categorise different kinds of uncertainty. See Chap. 5 by Roy in this volume or Morgan and Henrion (1990), Chap. 4.

> a very low probability for such extreme responses. Given that the parameters are only known to occur within the intervals, however, this conclusion is grossly inappropriate (p. 10–3).

As we discussed earlier: the product of independent uniformly distributed variables will not be uniform: it will give more weight to those values in the "middle" of the interval of possible values. If we have no reason to think the variables really are uniformly distributed, then it seems unwise to discount these possible extreme responses as the standard approach implicitly seems to.

This is sometimes known as "probability bounds analysis" or "p-boxes". The following two quotations give you the flavour of this approach.

> In a probability bounds analysis, the uncertainty about the probability distribution for each input variable is expressed in terms of interval bounds on the cumulative distribution function. These bounds form a p-box for each input variable Ferson and Hajagos (2004, p. 136).

> Basically, interval analysis should be used to propagate ignorance, and probability theory should be used to propagate variability Ferson and Ginzburg (1996, p. 133).

A similar approach is advocated by Stainforth et al. (2007b), where they suggest that we interpret the range of values produced by ensemble members to be a "non-discountable envelope" of values of that variable: a range of values that we cannot dismiss as impossible.

21.3.2 The Challenge Problems

In 2002, a workshop was organised around the idea of a set of "challenge problems" that were intended to serve as a kind of standard suite of tests for a theory of uncertainty (Helton and Oberkampf 2004). Oberkampf et al. (2004) presents the challenge problems, and many of the papers in that special issue of the journal respond to them. The problems are designed to highlight issues of "representation, aggregation, and propagation of uncertainty through mathematical models" (Oberkampf et al. 2004, p. 15). The challenge is to come up with some way to predict the behaviour of a system given a model of the system and some evidence as regards some unknown parameters of that system. Each problem has two unknown parameters, and some sort of mathematical model whose output depends on those unknowns. The information about the unknown parameters might be given in a number of different ways. A simple example is in problem 1, we are told that parameter a is somewhere in the interval $[a_1, a_2]$. A more complex example is given by problem 3c where you are told that you have n independent sources of information regarding parameter b, each witness j tells you that b lies in an interval $[b_1^j, b_2^j]$. The model whose outputs depends on the parameters can also be more or less complex. For example, for some of the models, it is simply a function of the parameters. In other cases, the parameters are meant to represent physical constants of some simple physical system.

Several papers have presented broadly IP solutions to this problem set. For example, de Cooman and Troffaes (2004) use the theory of lower previsons to address the problems, while Ferson and Hajagos (2004) use p-boxes.[8]

21.3.3 Nonprobabilistic Odds

Frigg et al. (2014) offer a cautionary tale that suggests that treating distributions of model output as capturing decision-relevant probabilities is dangerous when the target system appears to behave chaotically. The starting point is the scepticism about ensemble forecast probability distributions expressed by Stainforth et al. (2007a):

> The frequency distributions across the ensemble of models may be valuable information for model development, but there is no reason to expect these distributions to relate to the probability of real-world behaviour. One might (or might not) argue for such a relation if the models were empirically adequate, but given nonlinear models with large systematic errors under current conditions, no connection has been even remotely established for relating the distribution of model states under altered conditions to decision-relevant probability distributions (p. 2154).

Frigg et al. (2014) develop a simple example that illustrates this point. They start with a simple mathematical system that an agent wants to use to model a variable of interest. The target system's dynamics are similar to but not identical to the system used for prediction (the one timestep error is always less than one in a thousand). Unfortunately, both the target system and the model exhibit chaotic behaviour, which means that these errors compound and grow. If we are predicting at about eight timesteps out, the errors can grow to such an extent that the distribution of model outputs for an ensemble of nearby initial conditions can be located entirely on the left-hand side of the unit interval while the ensemble of outputs for the target system is entirely on the right-hand side. What this means is that the model appears to be telling you that it's overwhelmingly likely that the variable will be less than 0.5, while the truth is that it's overwhelmingly likely to be greater than 0.5. Obviously betting using these ensemble probabilities would be disastrous. What Frigg et al. (2014) show is that, in fact, it's very often disastrous to bet using ensemble probabilities in a case like this where the dynamics are nonlinear and there's a chance of model error.

They suggest that instead of taking the ensemble probabilities at face value, they should be manipulated to produce "nonprobabilistic odds" which don't yield ruinous betting strategies. How exactly this process should be effected is still up for debate, but what is clear is that the nonprobabilistic odds thus produced will be inversely proportional to upper probabilities, in the same way that probabilistic odds are inversely related to standard probabilities.

[8] See also Fetz and Oberguggenberger (2004) and Helton et al. (2004) for further examples of IP approaches to the challenge problems. See Ferson et al. (2004) for an overview of the range of responses to the challenge problems.

21.4 Interpretations

What does it mean to say that our uncertainty is captured by a set of probability measures (or an upper probability, or a p-box, or…)? In this section, we shall discuss some ways of interpreting such claims. I will discuss several such ideas, but I do not mean to suggest that this survey is exhaustive, nor that the ideas presented here are mutually exclusive: it is certainly possible to be motivated by more than one of these interpretations of the formalism.[9]

21.4.1 One-Sided Betting

Consider betting on the value of a random variable.[10] Let $\mathscr{X}$ be the set of values that random variable X can take. A bet on the value of X can be described by a function from $\mathscr{X}$ to the real numbers. Call such functions *gambles*. How much would you be willing to pay for a gamble g? That is, when do you find the gamble $g - \mu$ desirable? (Where μ is the constant gamble that corresponds to the amount you pay to take g). It seems like the highest price you are willing to pay to take g reflects your valuation of g. Let $\mathbf{lpr}(g) = \sup\{\mu \in \mathbb{R} : g - \mu \text{ is desirable}\}$. Now consider the minimum price you would accept to sell the gamble g. That is, the minimum price at which you find $\mu - g$ desirable. Given some reasonable coherence constraints on your set of desirable gambles,[11] if you require that this should be equal to **lpr**, then **lpr** is a *linear prevision*. And indeed, if we consider gambles over a set of indicator functions for some set of states, then **lpr** gives a probability function. The exploration of linear previsions as a foundation for probability theory goes back to Bruno de Finetti. This is one version of what is known as the "Dutch book theorem," since the coherence constraints on desirability essentially prevent you accepting a collection of bets that guarantee you a sure loss (see Chap. 7 by Beisbart in this volume).

If you drop the requirement that your maximum buying price and your minimum selling price should be the same—if you move away from "two-sided" betting to "one-sided" betting—then **lpr** behaves somewhat like a "lower expectation" operator (called a lower prevision) and its restriction to gambles on indicator functions is a lower probability as defined above. The theory of lower previsions was first systematically set out in Walley (1991), Troffaes and de Cooman (2014) provides an admirably clear self-contained treatment of the theory, as well as significant refinements. Note that, the bets discussed in Frigg et al. (2014) are one-sided bets in this sense.

[9] For more on the interpretation of IP, see Bradley (2014).

[10] We earlier described probability theory in terms of events rather than random variables, but the difference is mostly cosmetic. Real-valued random variables are functions from events to real numbers, events are "indicator functions" in the space of random variables.

[11] For example, if you find f desirable, and you find f' desirable, you should find $f + f'$ desirable; or if a gamble's payout is always non-negative, then it is desirable.

If we consider real-world instances of bookmakers or financial traders, there is typically a difference between their buying and selling prices for their commodities (bets, financial products, whatever). Now, part of this spread is explained by the desire to make a profit, but there is evidence that the "bid-ask spread" can also be responsive to the amount of uncertainty about the future performance of the instrument (Smith and van Boening 2008).

So, if we interpret lower previsions and lower probabilities as reflecting the agent's limiting willingness to bet—as is standard in Bayesian approaches—this gives us a natural interpretation of the formalism that is broadly in line with the standard precise probabilist picture.

21.4.2 Indeterminate Belief

One way to interpret credal sets is to take them to reflect an *indeterminacy* in rational belief. If $\mathscr{P}(X)$ is a set of values, this means that it is indeterminate—vague—what rational belief you ought to adopt in X given the evidence that determined that $\mathscr{P}$. This approach takes inspiration from the *supervaluationist* theory of logic, which uses a *set* of truth valuation functions to characterise the satisfaction of a vague predicate. If Wayne is a borderline case of the predicate "bald," then "Wayne is bald" is true according to some members of the set of valuations—"true on some precisifications"—and "Wayne is bald" is false on others (Williamson 1994). Rinard (2013, 2015) has argued for a supervaluationist understanding of credal sets: if it is indeterminate whether the agent's credence in X is stronger than 0.5, some members of the credal set have $\mathbf{pr}(X) > 0.5$ while others $\mathbf{pr}'(X) < 0.5$. If every member of the credal set agrees on something (e.g. that $\mathbf{pr}(X) > 0.1$), then it is determinately true that the agent believes that $\mathbf{pr}(X) > 0.1$. This idea is sometimes characterised by the metaphor of the "credal committee" (Joyce 2010): each probability in the credal set is a committee member and the committee as a whole must decide what to do. When there is unanimity in the committee then things are easy, when there is conflict—disagreement—then things are tricky.[12]

If you want to treat your credal sets or your lower probabilities not as subjective credences but as something akin to objective chances, then you might still be able to take a view of this form: the imprecision in your probabilities is due to objective indeterminacy in the world. This is an underexplored possibility, but see Bradley (2016).

[12] The idea of IP as reflecting unresolved conflict—either between persons or within a person—is one that Isaac Levi discussed in great detail Levi (1980, 1986).

21.4.3 Robustness Analysis

Let's say you run your model with a particular set of parameters, but you are not confident that the parameters you chose are the actual ones. If there's a danger that the result you obtain depends in a big way on the specific value of the parameter chosen, then perhaps, it's best to explore how robust your result is when changing those parameter values. The range of output values, the range captured by the set of probability functions, reflects a robust range of possible outcomes. This "robust bayesian analysis" has a rich history. See, for example Ruggeri et al. (2005). This idea is very closely connected to the discussion of "probability bounds analysis" and "non-discountable envelopes" discussed in Sect. 21.3.1.

If one is taking a Bayesian approach to validation (cf. Chap. 7 by Beisbart and Chap. 20 by Jiang et al. in this volume), then one has to respond to the "problem of the priors": the criticism that Bayesian methods rely on epistemically unmotivated prior probability. One response to such a criticism would be to move to an "imprecise Bayesian" perspective which is, essentially, to apply the robustness analysis approach to the prior. The set of priors allows one to be confident that one's conclusions are not artefacts of the particular prior one chose.

21.4.4 Evidence Theory

Instead of interpreting **lpr** as a degree of confidence, or a limiting willingness to bet, one might want to interpret $\mathbf{lpr}(X)$ as "the degree to which the evidence supports X". This is the interpretation typically associated with the "Dempster–Shafer function" approach to evidence. A "Dempster–Shafer belief function" is a lower probability that has the additional property of being "infinite monotone." The actual formal description of this property is a little messy, and not particularly illuminating in the current context, but see Halpern (2003, Chap. 2.4) for the basics of Dempster–Shafer theory, and see Augustin et al. (2014, Chaps. 4 and 5) and Troffaes and de Cooman (2014, Chaps. 6 and 7) for belief functions and their relation to lower probabilities. Dempster–Shafer theory also has a distinct theory of evidence combination which is beyond the scope of this chapter (but see Halpern (2003, Chap. 3.4)).[13]

The motivating idea behind this degree of support idea is that your evidence can support X to degree p without thereby supporting $\neg X$ to degree $1 - p$ (as would be required if degree of support were probabilistic). Hawthorne (2005) argues that Bayesians need to keep degree of belief and degree of support distinct (and that both concepts are useful).

[13] See also Oberkampf and Helton (2004) for an example of DS theory in an engineering context.

21.5 Problems

IP suffers from a number of issues. Here, we shall outline some of them. As we'll see, not all of them are really that worrying in the context of model validation.

21.5.1 Updating

Recall that earlier, we said that we conditionalise a set of probabilities pointwise. That is, $\mathscr{P}(X|Y) = \{\mathbf{pr}(X|Y), \mathbf{pr} \in \mathscr{P}, \mathbf{pr}(Y) > 0\}$.[14] There are two problems with conditionalising this way that is worth mentioning briefly.

First, dilation. An early, important discussion of dilation is Seidenfeld and Wasserman (1993).[15] Consider a set of probabilities $\mathscr{P}$ constrained as follows for all $\mathbf{pr} \in \mathscr{P}$:

- $\mathbf{pr}(H) = \frac{1}{2}$
- $\mathbf{pr}(H|X) = \mathbf{pr}(H)$

Now consider the proposition Y which is equivalent to $(H \wedge X) \vee (\neg H \wedge \neg X)$. It's easy to show that for all $\mathbf{pr} \in \mathscr{P}$, $\mathbf{pr}(Y) = \frac{1}{2}$. Let $\mathscr{P}$ contain all probability functions over these propositions other than those ruled out by the above constraints. So $\mathscr{P}(X) = [0, 1]$. Note that, it follows from the definition of Y and some basic probability theory that for all $\mathbf{pr} \in \mathscr{P}$, $\mathbf{pr}(H|Y) = \mathbf{pr}(X)$. Therefore $\mathscr{P}(H|Y) = [0, 1]$. Note that the same reasoning entails that $\mathscr{P}(H|\neg Y) = [0, 1]$. This, in essence, is the phenomenon of dilation. What is this considered a problem? To see this, let's consider an example that gives some meaning to the variables.[16]

I have two coins, one fair and one mystery coin of unknown bias. Let H be the event that the fair coin lands heads, and let X be the event that the mystery coin lands heads. (Verify that the above discussed probabilistic constraints seem reasonable given this interpretation of the propositions). Now I toss both coins and announce that the two coins landed the same way up (either both heads or both tails), call this proposition Y. What is your posterior in the fair coin having landed heads? $\mathscr{P}(H|Y) = [0, 1]$. Your belief in the fair coin's having landed heads has *dilated*: the interval of probability values has spread out from $\{\frac{1}{2}\}$ to $[0, 1]$. And this happens regardless of whether you learn Y or $\neg Y$. This seems puzzling. Learning the fact that the two coins landed the same way up doesn't seem like it should cause me to change my belief in whether the fair coin landed heads. It seems like you have learned something irrelevant to H and it has caused you to become more uncertain about H. That seems like a strange way to arrange your credences.

[14] The restriction to non-zero probability in the conditioning event is for convenience: if we had defined credal sets in terms of Popper functions or similar we could do without such a restriction.

[15] A recent characterisation of dilation is found in Pedersen and Wheeler (2014).

[16] This description of the puzzle follows Joyce (2010).

Dilation, despite initial appearances, isn't as problematic as some (e.g. White (2010)) take it to be. For example, in different ways, Joyce (2010), Bradley and Steele (2014b), Pedersen and Wheeler (2014) and Hart and Titelbaum (2015) all argue that dilation is actually the correct response to the evidence as specified. Gong and Meng (2017) argue that dilation is a symptom of a mis-specified statistical inference problem, not a problem for IP *per se*.

We can think of the problem as follows. The constraints we placed on our model place no constraint at all on what values $\mathbf{pr}(H|Y)$ might take. It is somewhat intuitive that H and Y say nothing about each other. If we take this "silence" to be modelled by probabilistic independence—$\mathbf{pr}(H|Y) = \mathbf{pr}(H) = \frac{1}{2}$ – then our $\mathscr{P}$ becomes the singleton with $\mathbf{pr}(X) = \mathbf{pr}(H|Y) = \frac{1}{2}$. But, as Pedersen and Wheeler (2014) point out, independence for sets of probabilities can be much more subtle (Cozman 2012). As Bradley (2014) explains, probabilistic independence is not the appropriate characterisation of the "silence" of Y with respect to H. As it stands, it is compatible with the problem set up that Y would be very informative about H, if only we knew something about the bias of the mystery coin. That is, if we knew that the mystery coin was biased towards heads, learning that the coins landed the same way up would be evidence in favour of heads on the fair coin. It is not that H and Y are unrelated: it's just that the nature of their relationship is unknown.

Let's turn now to another puzzle related to updating sets of probabilities: the problem of *belief inertia*. This problem, though not under that name, goes back to Sect.13.2 of Levi (1980), and is also discussed by Walley (1991); Vallinder (2018) provides a nice discussion of the current state of the art. Consider the mystery coin again. Recall that proposition X is "mystery coin lands heads up", and $\mathscr{P}(X) = [0, 1]$. Now consider learning that in ten flips of the mystery coin, 8 were heads. Call this proposition Z. This seems like some evidence that could potentially move your credences about. But note that $\mathscr{P}(X|Z) = (0, 1)$. Why? Because, even if we assume that all the priors in $\mathscr{P}$ are "well behaved" beta distributions over the unknown bias of the mystery coin, there are some distributions in $\mathscr{P}$ that put so much weight on the probability for landing heads being really really low that even evidence Z doesn't move them very far away from 0. In the case that $\mathscr{P}(X) = [0, 1]$ the "credal committee" contains members that are so stubborn that they are moved an arbitrarily small distance by the evidence. And likewise for the top end of the unit interval. Starting with a vacuous prior like this seems to make learning impossible.

The imprecise probabilities that are likely to arise in a validation setting are not likely to be vacuous, so perhaps this is less of a concern in the current context.

21.5.2 Decision-Making

Ultimately, we often want to use the results of our simulations for decision support: we want to take our simulation of the behaviour of a nuclear reactor to inform safety standards for new reactors, for example. This boils down to the question: how do we translate our uncertain predictions into policy advice? We want to take into account the uncertainty in our simulations and perhaps err on the side of caution by focusing on, for example the worst case among the plausible scenarios consistent with our evidence. So how do we make decisions with sets of probabilities?

If you had a single probability function, you can act so as to maximise expected utility. What is the analogue decision rule for imprecise probabilities? There are a number of possibilities, each with drawbacks. Should you act to maximise minimum expectation over probabilities in your set? Pick an option that maximises expected value with respect to some $\mathbf{pr} \in \mathscr{P}$?[17] Find some way to average over the set of expectation and maximise that?[18] Elga (2010) argues that no imprecise decision theory is even minimally adequate, although the consensus now seems to be that Elga overstated his case (Bradley and Steele 2014a; Chandler 2014; Sahlin and Weirich 2014). In any event, it is still true that providing IP with an adequate decision theory is an unresolved issue. In a sense, it is not surprising that decision-making with IP is difficult: the whole point is that we are being careful to represent the full extent of our lack of knowledge, and we shouldn't expect decision- making to be easy in such a case. Indeed, it would appear to be a surprise if decision-making were as easy as it is in a case where we know the objective probabilities of the events, or have some reason to believe our subjective estimates are on the right track.

21.6 Validation and IP

So where does all of the above leave the practitioner? What should someone who works with computer simulations take away from this discussion of imprecise probabilities? We'll discuss the issues of interpretation and the problems in the sections below, but first, I want to say something about where to situate IP in general. Some, particularly in philosophy, seem to see IP as a rival to the standard Bayesian view of subjective probability. I think this is a mistake. IP is a suite of tools, a range of methods that extend and improve on the standard probabilistic tools. They are provided in order to overcome some problems that the standard theory has with severe uncertainty, careful propagation of uncertainty and giving appropriate weight to serious dangers. Questions remain about when it is appropriate to deploy the admittedly more complex machinery of IP, and when it is best to stick with the simpler tools of standard probability, but the above discussion of the "challenge problems" highlights that many practictioners do see value in the use of IP.

[17] This option, called "E-admissibility" by Levi (1974)—and discussed in depth in Levi (1986)—is a popular one among some IP theorists.

[18] See Bradley (2015) for some discussion of the options.

21.6.1 Interpretations

We discussed four ways of interpreting the mathematical framework of IP: betting, indeterminate belief, robustness analysis and evidence theory. Which of these will be appropriate depends on your goals and how you are using the tools. Probabilities show up in validation, and there is a question about how to interpret them.[19] Your general interpretative view on probabilities is going to inform what you think about IP. If your inclination is to treat probabilities in a subjective Bayesian way, then the betting approach seems a natural fit for interpreting imprecise probabilities. If, however, your inclination is to treat probabilities as objective chances or objective evidential probabilities, then perhaps the "indeterminate belief/chance" route is a better fit.

If the imprecise probabilities arise from responding to the aleatory/epistemic uncertainty distinction in an "unknown parameters" context (see Sect. 21.3.1) then it makes sense to see IP as a kind of "robustness analysis".

The case of the "evidence theory" view of IP—$\mathbf{lpr}(X)$ represents the degree to which the evidence supports X—is an interesting one. This interpretation is strongly associated with Dempster–Shafer belief functions which are a special case of lower probabilities. It seems that much of the literature on this topic doesn't make a distinction between the "robustness analysis" view and the "evidence theory" view.

21.6.2 Problems

We looked at two kinds of problems for IP: problems related to updating and problems related to decision- making.

It seems that the problems for updating aren't all that problematic in the validation applications of IP. First, note that conditionalisation plays a relatively minor role outside of the Bayesian approach. And even where conditionalisation does play a role, the goal is to propagate the uncertainty. So if that yields large intervals of output variables then that is a good thing: the uncertainty has been adequately propagated. Take dilation: dilation occurs when you have two variables whose interaction is unknown. By that I mean, when it is unknown whether they are positively or negatively correlated.[20] In a validation context, if you were in that situation, you would want to know the range of possible system responses if the parameters were positively correlated or if they were negatively correlated: you would *want* to see that range of responses represented in your model output. Likewise for belief inertia. If it is consistent with your evidence that X might not be affected very much by conditioning on Z, then your model output should accommodate that possibility. In short, the first two problems discussed are problematic when we interpret $\mathscr{P}(A|B)$ as rational credence in A having learned B, but in a context where the probability models are representing

[19] See Hájek (2011) for an introduction to interpretations of probability.

[20] For a more careful and rigorous characterisation of dilation, see Pedersen and Wheeler (2014).

possible relationships between variables in a model, the phenomena of dilation and inertia do not seem so problematic.

So what about decision-making? If we are required to make a decision on the basis of some simulation-based prediction that involves several underconstrained parameters, it would be a mistake to make decision-making too easy. Take a simple example of deciding on how high to build your flood defences. Let's imagine that you use a climate model to predict whether sea level rises will be small, moderate or large. We run a bunch of climate models, varying some unknown parameters, and come up with a range of possible future scenarios. A precise probabilistic approach might say that on average sea level rise will be moderate. Now, it might be tempting to build flood defences that can cope with moderate rise but not with a large rise. It might also be tempting to not bother with investing in defences that can be extended in the case of a large rise: after all, the models say that won't happen. Of course, such a decision maker has made a mistake in paying too much attention to the headline "moderate rise" and not enough attention to the small print "range of scenarios". By taking seriously the task of propagating the uncertainty and presenting the full range of scenarios consistent with the physical constraints, the IP approach highlights the ranges of cases that the decision maker cannot discount in her deliberations. That makes it harder to make a decision, but it does so in a way that will improve the decisions made. Of course, one can go too far: the range of climate scenarios consistent with our models range from ice age to everybody dies of heatstroke. It's hard to make any sort of decision that will lead to good outcomes across the board there. But propagating that uncertainty, presenting the "non-discountable envelope" of scenarios, prevents the decision maker from being misled by the headline model average.

21.7 Conclusion

Imprecise probabilities can provide an expressively rich and sophisticated theory of uncertainty that builds on and extends orthodox probability theory. The formal foundations of IP are fairly solid although there are still some conceptual sticking points that need work. Such a theory has the potential to find many useful applications in the field of computer simulation validation.

References

Augustin, T., Coolen, F. P., de Cooman, G., & Troffaes, M. C. (Eds.) (2014). *Introduction to imprecise probabilities*. John Wiley and Sons.

Bradley, S. (2014). Imprecise probabilities. In E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy*.

Bradley, S. (2015). How to choose among choice functions. In *ISIPTA 2015 proceedings* (pp. 57–66).

Bradley, S. (2016). Vague chance. *Ergo*, *3*(20)

Bradley, S., & Steele, K. (2014a). Should subjective probabilities be sharp? *Episteme*, *11*, 277–289.

Bradley, S., & Steele, K. (2014b). Uncertainty, learning and the “problem” of dilation. *Erkenntnis*, *79*, 1287–1303.

Chandler, J. (2014). Subjective probabilities need not be sharp. *Erkenntnis*, *79*, 1273–1286.

Cozman, F. (2012). Sets of probability distributions, independence and convexity. *Synthese*, *186*, 577–600.

de Cooman, G., & Troffaes, M. (2004). Coherent lower previsions in systems modelling: Products and aggregation rules. *Reliability Engineering and System Safety*, *85*, 113–134.

Dempster, A. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, *38*, 325–339.

Elga, A. (2010). Subjective probabilities should be sharp. *Philosophers' Imprint*, 10.

Ferson, S., & Ginzburg, L. R. (1996). Different methods are needed to propagate ignorance and variability. *Reliability Engineering and System Safety*, *54*, 133–144.

Ferson, S., & Hajagos, J. G. (2004). Arithmetic with uncertain numbers: Rigorous and (often) best possible answers. *Reliability Engineering and System Safety*, *85*, 135–152.

Ferson, S., Joslyn, C. A., Helton, J. C., Oberkampf, W. L., & Sentz, K. (2004). Summary from the epistemic uncertainty workshop: Consensus amid diversity. *Reliability Engineering and System Safety*, *85*, 355–369.

Fetz, T., & Oberguggenberger, M. (2004). Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. *Reliability Engineering and System Safety*, *85*, 73–87.

Frigg, R., Bradley, S., Du, H., & Smith, L. A. (2014). Laplace's demon and the adventures of his apprentices. *Philosophy of Science*, *81*, 31–59.

Gong, R., & Meng, X. -L. (2017). Judicious judgment meets unsettling update: Dilation, sure loss and simpson's paradox. arXiv:1712.08946v1.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society Series B (Methodological)*, *14*, 107–114.

Good, I. J. (1962). Subjective probability as the measure of a non-measurable set. In *Logic, methodology and philosophy of science: Proceedings of the 1960 international congress* (pp. 319–329).

Hájek, A. (2011). Interpretations of probability. In E. N. Zalta (ed.), *The stanford encyclopedia of philosophy*. Stanford. http://plato.stanford.edu/archives/fall2007/entries/probability-interpret/.

Halpern, J. Y. (2003). *Reasoning about uncertainty*. MIT Press.

Hart, C., & Titelbaum, M. G. (2015). Intuitive dilation? *Thought*, *4*, 252–262.

Hawthorne, J. (2005). Degree-of-belief and degree-of-support: Why Bayesians need both notions. *Mind*, *114*, 277–320.

Helton, J. C., Johnson, J., & Oberkampf, W. L. (2004). An exploration of alternative approaches to the representation of uncertainty in model predictions. *Reliability Engineering and System Safety*, *85*, 39–71.

Helton, J. C., & Oberkampf, W. L. (2004). Alternative representations of epistemic uncertainty. *Reliability Engineering and System Safety*, *85*, 1–10.

Joyce, J. M. (2010). A defense of imprecise credence in inference and decision. *Philosophical Perspectives*, *24*, 281–323.

Klir, G. J., & Smith, R. M. (2001). On measuring uncertainty and uncertainty-based information: Recent developments. *Annals of Mathematics and Artificial Intelligence*, *32*, 5–33.

Koopman, B. O. (1940). The bases of probability. *Bulletin of the American Mathematical Society*, *46*, 763–774.

Levi, I. (1974). On indeterminate probabilities. *Journal of Philosophy*, *71*, 391–418.

Levi, I. (1980). *The enterprise of knowledge*. The MIT Press.

Levi, I. (1986). *Hard Choices: Decision making under unresolved conflict*. Cambridge University Press.

Morgan, M. G. & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitiative risk and policy analysis*. Cambridge University Press.

Oberkampf, W. L., & Helton, J. C. (2004). Evidence theory for engineering applications. In *Engineering design reliability handbook* (pp. 197–226). CRC Press.

Oberkampf, W. L., Helton, J. C., Joslyn, C. A., Wojtkiewicz, S. F., & Ferson, S. (2004). Challenge problems: Uncertainty in system response given uncertain parameters. *Reliability Engineering and System Safety*, *85*, 11–19.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge University Press.

Pedersen, A. P., & Wheeler, G. (2014). Demystifying dilation. *Erkenntnis*, *79*, 1305–1342.

Rinard, S. (2013). Against radical credal imprecision. *Thought*, *2*, 157–165.

Rinard, S. (2015). A decision theory for imprecise probabilities. *Philosophers' Imprint*, *15*, 1–16.

Ruggeri, F., Insua, D. R., & Martín, J. (2005). Robust bayesian analysis. In *Handbook of statistics* (Vol. 25, pp. 623–667). Elsevier.

Sahlin, N.-E., & Weirich, P. (2014). Unsharp sharpness. *Theoria*, *80*, 100–103.

Seidenfeld, T. (1983). Decisions with indeterminate probabilities. *The Brain and Behavioural Sciences*, *6*, 259–261.

Seidenfeld, T. (1988). Decision theory without "independence" or without "ordering". what's the difference? *Economics and philosophy* (pp. 267–290).

Seidenfeld, T., Kadane, J. B., & Schervish, M. J. (1989). On the shared preferences of two Bayesian decision makers. *The Journal of Philosophy*, *86*, 225–244.

Seidenfeld, T., & Wasserman, L. (1993). Dilation for sets of probabilities. *Annals of Statistics*, *21*, 1139–1154.

Smith, C. A. (1961). Consistency in statistical inference and decision. *Journal of the Royal Statistical Society Series B (Methodological)*, *23*, 1–37.

Smith, V. L. & van Boening, M. (2008). Exogenous uncertainty increases the bid-ask spread in the continuous double auction. In *Handbook of experimental economics results* (vol. 1). Elsevier.

Stainforth, D. A., Allen, M. R., Tredger, E., & Smith, L. A. (2007a). Confidence uncertainty and decision-support relevance in climate models. *Philosophical Transactions of the Royal Society*, *365*, 2145–2161.

Stainforth, D. A., Downing, T., Washington, R., Lopez, A., & New, M. (2007b). Issues in the interpretation of climate model ensembles to inform decisions. *Philosophical Transactions of the Royal Society*, *365*, 2163–2177.

Troffaes, M., & de Cooman, G. (2014). *Lower previsions*. Wiley.

Vallinder, A. (2018). Imprecise bayesianism and global belief inertia. *The British Journal for the Philosophy of Science*, *69*, 1205–1230.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities, volume 42 of Monographs on Statistics and Applied Probability*. Chapman and Hall.

White, R. (2010). Evidential symmetry and mushy credence. In T. Szabo Gendler, & J. Hawthorne (Eds.), *Oxford studies in epistemology* (pp. 161–186). Oxford University Press.

Williamson, T. (1994). *Vagueness*. Routledge.

Chapter 22
Objective Uncertainty Quantification

Edward R. Dougherty, Lori A. Dalton and Roozbeh Dehghannasiri

Abstract When designing an operator to alter the behavior of a physical system, the standard engineering paradigm is to begin with a scientific model describing the system, mathematically characterize a class of operators, define a performance cost relative to the operational objective, and pick an operator that minimizes the performance cost. Validation ipso facto plays a role because the scientific model needs to be validated. With complex systems, or those for which experiments are costly, there may be insufficient data for system identification, with validation being outside the realm of possibility. Given the resulting model uncertainty, the best one can do is to design a "robust" operator that is optimal relative to both the objective and the uncertainty. This robust optimization paradigm entails optimal experimental design: should one not be satisfied with the performance, choose the experiment that maximally reduces the uncertainty as it pertains to the objective. In this chapter, we address these problems and present examples in the context of gene regulatory network intervention.

Keywords Bayesian experimental design · Robust design · Mean absolute cost of uncertainty · Uncertainty quantification

22.1 Introduction

In the classical deterministic scenario, a simple model consists of a few variables and a few parameters that can be estimated from a handful of experiments. The model

E. R. Dougherty (✉) · R. Dehghannasiri
Texas A&M University, College Station, TX 77843, USA
e-mail: edward@ece.tamu.edu

R. Dehghannasiri
e-mail: roozbehdn@tamu.edu

L. A. Dalton
The Ohio State University, Columbus, OH 43210, USA
e-mail: dalton@ece.osu.edu

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_22

can be tested with a few numerical predictions, and it is contingently accepted if predictions are deemed to be concordant with observations. As model complexity grows to hundreds and thousands of variables and parameters, as is typical of many computer models and simulations, the classical procedure becomes problematic. The difficulty is exacerbated by stochasticity because prediction then includes testing the accuracy of probability distributions in the model. Systems with thousands of variables are virtually unvalidatable.

For a deterministic model, initial conditions can be set and, in principle, the state at some future time determined. If the initial conditions of a test experiment are aligned with those of the model and the experiment run to some future time, then agreement between the final model and experimental states can be checked. Large-scale deterministic systems have high-dimensional state vectors, so that test experiments are more demanding; nevertheless, the ultimate comparison is still between model and experimental state vectors. It is prudent to run tests using a variety of initial conditions so that a large portion of the state space is tested.

With a stochastic model, the situation is more challenging. Given an initial state, the final state will not be determined exactly; rather, there will a probability distribution of possible final states. Hence, comparison must be between the state distribution, which is generally multivariate, and a state histogram generated by many experimental runs, the number of required runs growing exponentially with the number of variables. Distributional statistical tests are required. For instance, with hypothesis testing one decides between two hypotheses: the distributions match or they do not match. A decision to accept the theory depends on the acceptance threshold. The theory and test are inter-subjective, but the decision to accept or reject depends on subjective considerations, as with a hypothesis test, where the acceptance region depends on a chosen level of significance. The overall procedure can be onerous (or impossible) depending on the number of experimental runs required, especially with complex systems, where distributions are high-dimensional.

For complex systems, not only is it practically impossible to validate a model, it is not even possible to obtain a specific model. Parameter estimation is required for model construction, but when the number of parameters is too large for the amount of data, accurate parameter estimation becomes impossible. It may be possible to assume ranges for the parameters, but then there is model uncertainty: all models with parameters lying in the given ranges constitute an uncertainty class of possible models. In this situation, the classical concept of validation makes no sense. The desire for knowledge concerning complex systems and the impossibility of testing corresponding models, or even formulating such models, forms a salient epistemological crisis of the twenty-first century (Dougherty 2016).

From an engineering perspective, one has to ask a basic question: How do we mathematically formulate action in nature if we cannot posit a single description of the part of nature that we wish to act upon? Or: How do we determine optimal operations to alter natural systems when we cannot precisely specify the mathematical relations among the variables of the system? The answer is that we must determine the best operation, not only relative to our objective (diagnosing the system, controlling the system, predicting future system behavior, etc.), but also relative to our uncertainty regarding the system. To do this, the uncertainty must be quantified and merged with our operational objective to form a criterion to judge operator performance across the uncertainty class.

Historically, there have been many attempts to characterize and differentiate between types of uncertainty. For instance, aleatory uncertainty is irreducible and inherent in a model, while epistemic uncertainty is reducible given additional knowledge or data (Bae et al. 2004) (see also Chap. 3 by Oberkampf and Chap. 5 by Roy in this volume). This chapter takes a Bayesian approach, modeling uncertainty in our scientific knowledge itself (cf. Chap. 7 by Beisbart and Chap. 20 by Jiang et al. in this volume). We assume a prior distribution over an uncertainty class of plausible parametric models (states of nature). These can be any scientific model, including stochastic differential equations, Gaussian or non-Gaussian stochastic processes, feature-label distributions (as in classification), or Markov chains (as with regulatory networks), to name a few. We assume that aleatory uncertainty is accounted for in the scientific model. For example, given a constrained number of measurement variables, gene regulation is inherently stochastic; thus, we integrate this inherent randomness into our gene regulation models through the introduction of a perturbation probability. However, our main focus is not on modeling aleatory uncertainty but on quantifying epistemic uncertainty in our scientific knowledge.

Rather than quantifying uncertainty in a global sense, for instance using the variance or entropy of model parameters, we focus on uncertainty quantification from the perspective of an operational objective. How do we characterize uncertainty when the objective is to design an operator to optimally achieve a certain engineering goal, such as filtering an image, classifying an observed object, or controlling a physical system? Consider a class of candidate operators, and a cost function for each operator and state-of-nature pair. The cost may be, for example, the error rate of a given operator. Given a prior distribution on the states of nature, we define the *intrinsically Bayesian robust* (IBR) operator to be an operator corresponding to the minimum expected cost, averaged over the prior. We then quantify uncertainty using the *mean objective cost of uncertainty* (MOCU) (Yoon et al. 2013), which is the increase in cost between the IBR operator and the optimal operator for a given state of nature, averaged over the prior. Unlike variance or entropy, MOCU quantifies uncertainty in terms of our ability to make inferences. In gene networks, the engineering goal is structural intervention, which amounts to blocking interaction between a pair of genes, thus severing an edge in the regulatory network. For a given intervention and network pair, cost is quantified by the long-run probability that a certain combination of undesirable gene activity levels occurs simultaneously. Given a set of candidate structural interventions, and uncertainty in the underlying regulatory functions in the actual network, an IBR intervention corresponds to an intervention with minimal expected cost, and the MOCU quantifies uncertainty in our ability to intervene without knowing the actual state of nature precisely.

An excellent application of MOCU is in the sequential information collection problem, where the objective is to devise a strategy of selecting real or computer experiments from a set of available options. A key component of any experimental design framework is uncertainty. As experiments are observed, uncertainty in some underlying mechanism or state of nature should be reduced, thereby aiding future decisions. A basic strategy for experimental design is thus to quantify uncertainty given various potential experimental observations, and to choose the experiment corresponding to maximal expected uncertainty reduction in the future. One possible option is to choose the experiment resulting in maximal reduction in variance or entropy, but this approach can spend many experiments reducing uncertainty in variables that are irrelevant to the engineering objective at hand. A better strategy would be to minimize the expected MOCU in the future, where the expectation is over an assumed distribution over the experimental outcomes. Then, the future IBR operator (after the experiment) will have minimal expected cost, averaged with respect to the state of nature and the future experiment. In other words, we directly maximize our ability to make accurate inferences in the future, and the optimal inference operator automatically falls out of the uncertainty analysis.

We begin with a review of our example, gene regulatory networks, in Sect. 22.2. This is followed by a review of optimal operators in Sect. 22.3, and an example of optimal intervention in regulatory networks in Sect. 22.4. We review IBR operators in Sect. 22.5, which have optimal expected cost in the presence of uncertainty, and present examples of IBR intervention in regulatory networks in Sect. 22.6. In Sect. 22.7, we review MOCU and discuss how this framework can be used to characterize optimal experimental design. We discuss examples of optimal experimental design for interventions in regulatory networks in the presence of uncertainty in Sect. 22.8. In Sect. 22.9, we discuss general issues pertaining to Bayesian inference and experimental design. We conclude in Sect. 22.10.

22.2 Gene Regulatory Networks

We will illustrate the theory by considering intervention in regulatory networks, in particular, cell regulation. The regulatory system in a cell is mainly based on its genetic structure. The basic paradigm has two parts. *Transcription* is the process by which genetic information in a gene is copied into messenger RNA (mRNA). When this process is occurring the gene is said to be *expressing* (or activated). Expression is governed by signaling proteins attaching themselves (binding) to the gene's promoter region. In essence, each gene is controlled by the states of a set of genes, so that activation depends on the expression levels of its regulating genes. *Translation*, which occurs subsequent to transcription, refers to the production of protein, based on the code carried by the mRNA, that can either be involved in maintaining the cell structure or function as a signal (*transcription factor*) to instigate or prohibit further gene expression.

A gene regulatory network (GRN) is a mathematical model comprised of a set of entities called "genes" and a regulatory structure that governs their behavior over time. GRNs can be finely detailed, as with differential-equation models, or coarse-grained, with discrete expression levels transitioning over discrete time. We consider a Boolean network, in which a gene can have logical values 1 or 0, corresponding to expressing or not expressing, respectively, and regulation is specified by logical operations among genes, so that functional relationships between genes can be specified by a truth table (Kaufmann 1993). Although the original formulation (and the one we consider) is two-valued, 0 or 1, the Boolean-network concept applies to any number of discrete gene values and has been extended to probabilistic Boolean networks in which the logic is affected by regulation outside the network (Shmulevich et al. 2002).

An n-node *Boolean network* (BN) is a pair $(\mathbf{V}, \mathbf{F})$, where $\mathbf{V} = \{X_1, X_2, \ldots, X_n\}$ is a set of binary-valued nodes and $\mathbf{F} = \{f_1, f_2, \ldots, f_n\}$ is a set of Boolean functions such that $f_i : \{0, 1\}^{k_i} \to \{0, 1\}$ is the Boolean function that determines the value of X_i, 0 or 1. In the context of gene regulatory networks, the nodes correspond to “off” and “on” states of the genes and the Boolean functions represent gene regulation. The vector $X(t) = (X_1(t), \ldots, X_n(t))$ gives the state of the network at time t. The value of X_i at the next time point,

$$X_i(t+1) = f_i\left(X_{i_1}(t), X_{i_2}(t), \ldots, X_{i_{k_i}}(t)\right), \tag{22.1}$$

is determined by the values of k_i predictor nodes at time t. Given an initial state, a Boolean network will eventually reach a set of states, called an *attractor cycle*, through which it will cycle endlessly. Each initial state corresponds to a unique attractor cycle and the set of states leading to a specific attractor cycle is known as the *basin of attraction* of the attractor cycle.

To illustrate validation and uncertainty in the context of Boolean networks, consider the tumor suppressor gene p53, which in mammalian genomes is a transcription factor for hundreds of downstream genes that modulate cell cycle progression, repair damaged DNA, and induce senescence and apoptosis (cell self-destruction). Figure 22.1, adapted from Batchelor et al. (2009), shows some major pathways involving p53 that are activated in the presence of DNA double-strand breaks. An arrow indicates an activation signal and a blunt end indicates suppression. This kind of pathway model, common in biology, is under-determined, in the sense that many networks might generate the pathways. We consider two such Boolean networks having states [ATM, p53, Wip1, Mdm2]. An external input signal, denoted dna_dsb, takes on the value 1 or 0, depending on whether there is or is not DNA damage. This leads to two 4-gene Boolean networks determined by the following logical rules:

$$\begin{aligned}
\mathrm{ATM}_{\mathrm{next}} &= \overline{\mathrm{Wip1}} \wedge \mathrm{dna\_dsb} \\
\mathrm{p53}_{\mathrm{next}} &= \overline{\mathrm{Mdm2}} \wedge \mathrm{ATM} \wedge \overline{\mathrm{Wip1}} \\
\mathrm{Wip1}_{\mathrm{next}} &= \mathrm{p53} \\
\mathrm{Mdm2}_{\mathrm{next}} &= (\overline{\mathrm{ATM}} \wedge (\mathrm{p53} \vee \mathrm{Wip1})) \vee (\mathrm{p53} \wedge \mathrm{Wip1})
\end{aligned}$$

The symbols $\wedge$, $\vee$, and $\overline{}$ represent logical “and”, “or”, and “not”, respectively.

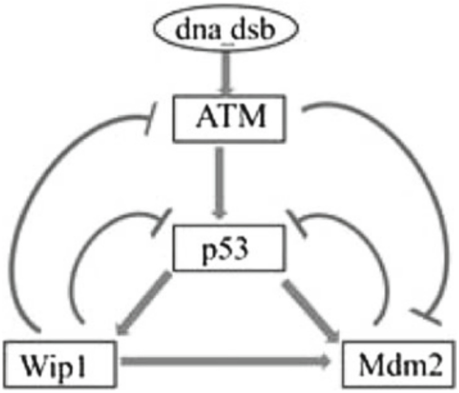


Fig. 22.1 Major pathways involving p53 that are activated in the presence of DNA double-strand breaks. [Dougherty, E. R., *The Evolution of Scientific Knowledge: From Certainty to Uncertainty*, SPIE Press, Bellingham, 2016; © SPIE and being used with permission.]

The state transition diagrams for these networks are shown in Fig. 22.2: (a) dna_dsb = 0; (b) dna_dab = 1. Absent damage, from any initial state, the network evolves into the single attractor state 0000; with damage, the network evolves into a 5-state attractor cycle in which p53 (state number 2) oscillates between expressing and not expressing. There are several ways to validate this kind of genomic network based on long-run (attractor) behavior (Dougherty 2007).

Now, suppose that the regulatory function for ATM is unknown in the sense that we know that it is of the form $\mathrm{ATM}_{\mathrm{next}} = x \wedge$ dna_dsb. The variable x can be either p53, Wip1, Mdm2, or any of their negations. Thus, there is an uncertainty class consisting of six possible networks. Validation is impossible since we don't know which of the six is to be validated. One might argue that if we attempt to validate all six and only the correct one, x being the negation of Wip1, is validated, not only have we found the correct network but also validated it. The problem is that this assumes sufficient data for validation, which would be more than sufficient to decide on the hypothetical network to begin with, which is precisely what we cannot do.

Putting uncertainty aside, suppose there is a mutation and the network of Fig. 22.2b is altered so that state 0000 is an attractor. Then the network stays in 0000 when there is DNA damage. This is a bad mutation because p53 remains off when there is DNA damage so that the corrective downstream effects are not actuated. To treat this condition, we desire a drug that will structurally intervene to alter the mutated regulatory logic and best treat the condition. It might not be possible to completely correct the logic, but among a set of possible drugs, which one provides the optimal correction, where the notion of "optimal" must be medically defined? The present chapter discusses optimal operations (such as structural intervention in a regulatory network) when there is model uncertainty.

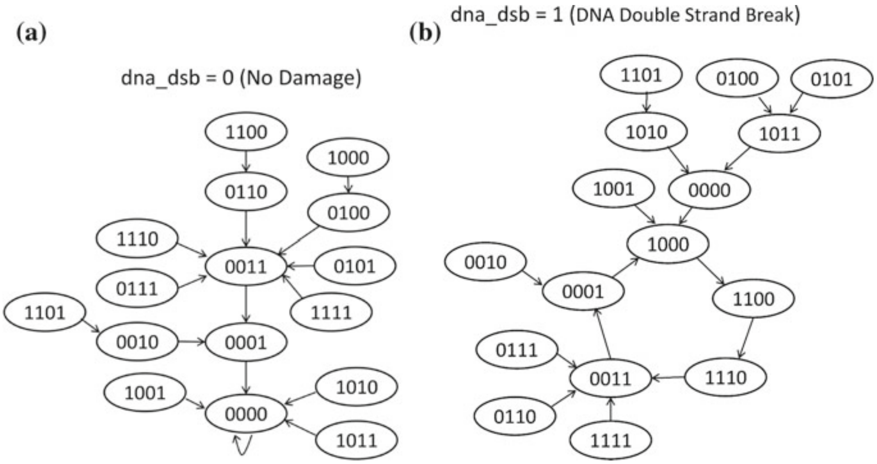


Fig. 22.2 Boolean network state transition diagrams for states [ATM, p53, Wip1, Mdm2]: **a** dna_dsb = 0, **b** dna_dab = 1. [Dougherty, E. R., *The Evolution of Scientific Knowledge: From Certainty to Uncertainty*, SPIE Press, Bellingham, 2016; © SPIE and being used with permission.]

22.3 Optimal Operators

Modern engineering begins with a scientific model, but in addition to the model, there is an objective. The situation changes from simply modeling behavior to action. In medicine, engineering is popularly called *translational science*, which accurately describes modern engineering. A scientific model, whose purpose is to provide a conceptualization of some portion of the physical world, is transformed into a model characterizing human action in the physical world. Scientific knowledge is translated into practical knowledge by expanding a scientific system to include external inputs that can be adjusted to affect the behavior of the system and outputs that monitor the effect of the external inputs and feed back information on how to adjust the inputs (Dougherty and Bittner 2011). For example, in biomedical science models are created with the intention of using them for diagnosis, prognosis, and therapy.

If one is going to transform a physical process, then the conceptualization of that physical transformation takes the form of a mathematical operator on some mathematical system, which itself is a scientific model for the state of a portion of nature absent the transformation. The product of science is a validated model, whereas the product of translational science is an operator that transforms some aspect of nature in a quantifiably useful manner. For translation, the scientific model is an intermediate construct used to facilitate control of nature; its descriptive power is of concern only to the degree that it affects the operator designed from it.

There are two basic operator problems concerning systems. One is *analysis*: given a system, characterize the properties of the transformed system resulting from the operator in terms of the properties of the original system. The second, and the one that concerns us here, is *synthesis*: given a system, design an operator to transform the system in some desirable manner. Synthesis forms the basis of modern engineering (translational science). Synthesis begins with the relevant scientific knowledge constituted in a mathematical theory. This is used to derive an optimal operator for accomplishing a desired transformation under the constraints imposed by the circumstances. A criterion, called a *cost function* (objective function) is defined to judge the goodness of the response—the lower the cost, the better the operator. The objective is to minimize the cost function.

For translational science, synthesis generally involves four steps:

1. Construct the mathematical model (graphical network, ordinary differential equations, stochastic differential equations, etc.).
2. Define a class of operators.
3. Define the optimization problem via a cost function.
4. Solve the optimization problem.

The optimization problem takes the following form: among all operators ψ in the operator class $\mathscr{F}$, find an operator ψ_{opt} that minimizes the cost $C(\psi)$ of applying operator ψ to the model.

22.4 Optimal Intervention in Regulatory Networks

Turn back to our example of regulatory networks. Different types of internal stochasticity can be placed into a Boolean network. Here we restrict ourselves to a single, simple type. In a Boolean network with perturbation (BNp), each node may randomly flip its value at a given time with a perturbation probability $p > 0$, independently from other nodes. Hence, for a BNp, $X(t+1) = \mathbf{F}(X(t))$ with probability $(1-p)^n$, when there is no perturbation, but $X(t+1)$ may take a different value, with probability $1-(1-p)^n$, when there exists one or more random perturbations. On account of perturbation, the network can jump out of an attractor cycle into a different basin of attraction and then transition to a new attractor cycle. Thus, from any state, there is a positive probability of reaching any other state at the next time point. In a BNp, the sequence of states over time can be regarded as a Markov chain.

A discrete-time, finite-state, homogeneous *Markov chain* is a vector stochastic process $X(t)$ completely defined by its initial state $X(0)$ and its one-step *transition probability matrix* (TPM) $\mathbf{P}$ whose i, j component is the transitional probability $P(X(t+1) = j|X(t) = i)$. If there exists a probability distribution $\{\varphi_j\}$ such that for all states i, j,

$$\lim_{t\to\infty} P(X(t+1) = j|X(0) = i) = \varphi_j, \tag{22.2}$$

meaning that in the long-run (as $t \to \infty$) the probability of transitioning to state j equals φ_j no matter the initial state i, then $\{\varphi_j\}$ is known as the *steady-state* (long-run) distribution. Equivalently, the probability of being in state j in the long-run is φ_j, independent of the initial state.

For a BNp, transitions are made according to a fixed TPM **P** and Markov chain theory can be applied for analyzing network dynamics. The general formula of a TPM using Boolean functions and perturbation probability are derived in Faryabi et al. (2009). The Markov chain possesses a steady-state distribution.

The issue for intervention in Markovian networks is to choose from a family of interventions the one that best alters the steady-state distribution of the network. The Markov-chain states are partitioned into sets D and U of desirable and undesirable states, respectively, possessing total steady-state probability mass $\varphi_D = \sum_{i \in D} \varphi_i$ and $\varphi_U = 1 - \varphi_D$, respectively. The operator class $\mathscr{F}$ consists of a family of transformations ψ on the TPM **P** and the cost function C is the total steady-state probability mass of the undesirable states. An optimal intervention minimizes

$$C(\psi) = \varphi_U^\psi = \sum_{i \in U} \varphi_i^\psi, \tag{22.3}$$

where φ_U^ψ and φ_i^ψ are the undesirable steady-state mass and steady-state probability of state i, respectively, following the intervention ψ.

As an illustration, consider the cell cycle, which controls cell duplication and division. The model contains ten genes: CycD, Rb, p27, E2F, CycE, CycA, Cdc20, Cdh1, UbcH10, and CycB, with genes numbered in this order. The cell cycle in mammals is controlled via extra-cellular stimuli. Positive stimuli activate Cyclin D (CycD) in the cell, thereby leading to cell division. CycD inactivates the Rb protein, which is a tumor suppressor. The regulatory model, shown in Fig. 22.3, has blunt arrows representing suppressive regulations and normal arrows representing activating regulations. This regulatory model can also be summarized by a regulatory matrix $\mathbf{R} = (R_{ij})$, where R_{ij} represents the regulatory relation from gene j to gene i: $R_{ij} = 1$ if the relation between genes j and i is activating, $R_{ij} = -1$ if the relation between genes j and i is suppressive, and $R_{ij} = 0$ if there is no relation between genes j and i. To construct a BNp from the regulatory model, we assume the majority vote rule, where at each time point a gene takes the value 1 if the majority of its regulator genes are activating and the value 0 if the majority of its regulatory genes are suppressive; otherwise, it remains unchanged. Under this rule, $X_i(t+1) = f_i(X(t))$ with probability $1 - p$, where

$$f_i(X(t)) = \begin{cases} 1 & \text{if } \sum_j R_{ij} X_j(t) > 0 \\ 0 & \text{if } \sum_j R_{ij} X_j(t) < 0 \\ X_i(t) & \text{if } \sum_j R_{ij} X_j(t) = 0 \end{cases}. \tag{22.4}$$

When gene p27 and either CycE or CycA are active, the cell cycle stops, because Rb can be expressed even in the presence of cyclins. States in which the cell cycle continues even in the absence of stimuli are associated with cancerous phenotypes. For this reason, states with down-regulated CycD, Rb, and p27 (corresponding to $x_1 = x_2 = x_3 = 0$) are undesirable. A structural intervention removes an arrow from the regulatory graph because it blocks a regulation between two genes. The interven-

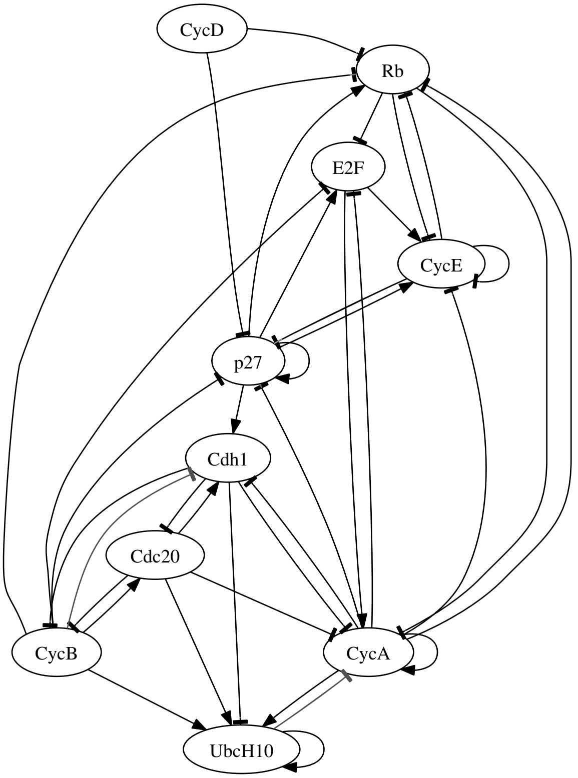


Fig. 22.3 Regulatory model for the mammalian cell cycle. [Dougherty, E. R., *The Evolution of Scientific Knowledge: From Certainty to Uncertainty*, SPIE Press, Bellingham, 2016; © SPIE and being used with permission]

tion alters the TPM and hence the steady-state distribution. The structural intervention that maximally lowers undesirable steady-state probability blocks the regulatory action from gene CycE to p27 and reduces total undesirable steady-state probability from 0.3401 to 0.2639 (Qian and Dougherty 2008). The steady-state distributions for the original network and the treated network are shown in Figs. 22.4 and 22.5.

22.5 Intrinsically Bayesian Robust Operators

The cost function upon which optimization is based depends on the scientific model being known with certainty. In many problems, this assumption is warranted. At worst, there is insignificant deviation between the behavior of the variables in the model and the empirical behavior they represent. However, this is not always the case, especially with complex systems. Specifically, while some parameters of the

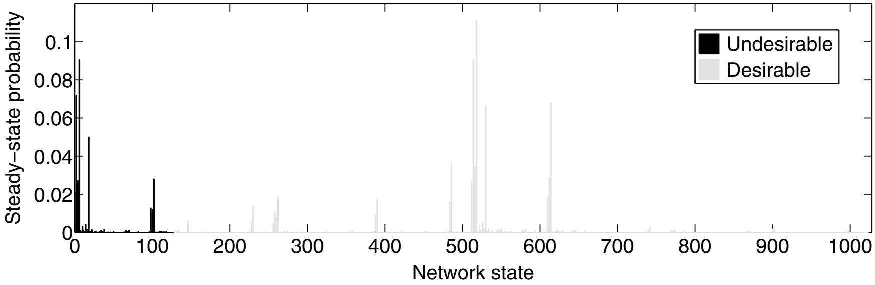


Fig. 22.4 Steady-state distribution of the original network. [Dougherty, E. R., *The Evolution of Scientific Knowledge: From Certainty to Uncertainty*, SPIE Press, Bellingham, 2016; © SPIE and being used with permission]

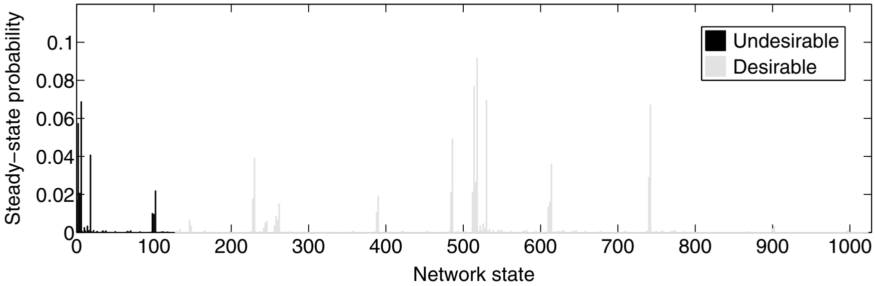


Fig. 22.5 Steady-state distribution of the treated network. [Dougherty, E. R., *The Evolution of Scientific Knowledge: From Certainty to Uncertainty*, SPIE Press, Bellingham, 2016; © SPIE and being used with permission]

model can be assumed to be known with certainty, for others this assumption is unwarranted. The result is an *uncertainty class* of models determined by a parameter vector θ consisting of the unknown parameters. If Θ is the set of possible values of θ, then the uncertainty class is referred to as Θ because the uncertainty class of models is in one-to-one correspondence with Θ. It is crucial to recognize that Θ defines a class of mathematical processes that characterize our uncertain scientific knowledge.

To formulate optimization when there is model uncertainty, let $\mathscr{F}$ be a family of operators whose performance on model $\theta \in \Theta$ is measured by the cost function C_θ. For each operator $\psi \in \mathscr{F}$, there is a cost $C_\theta(\psi)$ of applying ψ on model $\theta \in \Theta$. An *intrinsically Bayesian robust* (IBR) operator $\psi_{\mathrm{IBR}}^{\Theta} \in \mathscr{F}$ minimizes the expected value over Θ, among all operators in $\mathscr{F}$, of the cost $C_\theta(\psi)$, the expected value being with respect to a prior probability distribution $\pi(\theta)$ over Θ (Yoon et al. 2013; Dalton and Dougherty 2014). An IBR operator is *robust* because on average it performs well over the whole uncertainty class. Since each parameter vector θ corresponds to a model, a probability distribution on the space of possible models quantifies our belief that some models are more likely to be the actual model than are others. It quantifies our prior knowledge. If there is no prior knowledge beyond the uncertainty class itself, then the prior distribution is taken to be uniform, meaning that all models are assumed to be equally likely.

In many instances, the solution for the IBR operator takes a form analogous to the optimal operator when there is no uncertainty. Specifically, when an optimal operator is expressed via characteristics of the random processes constituting the underlying scientific model, where by a characteristic we mean some entity derived from the processes, such as a covariance matrix, then it is often the case that the IBR operator is expressed in the same manner except that the random-process characteristics are replaced by *effective characteristics* that summarize the information across the uncertainty class. For instance, for wide-sense stationary random processes, the classical Wiener filter is expressed in terms of power spectra of the processes. When the random processes are uncertain, the IBR Wiener filter is expressed in terms of effective power spectra (Dalton and Dougherty 2014). In classification, the optimal classifier is expressed in terms of the class-conditional densities. When the feature-label distribution is uncertain, the IBR classifier is expressed in terms of the effective class-conditional densities (Dalton and Dougherty 2013). Other examples include morphological filtering (Dalton and Dougherty 2014) and Kalman filtering (Dehghannasiri et al. 2017).

When it is possible to construct effective characteristics, the four-step schema for optimal operator design is extended by the following five steps (Dougherty 2016):

5. Identify the uncertainty class.
6. Construct a prior distribution.
7. State the IBR optimization problem.
8. Construct the appropriate effective characteristics.
9. Prove that the IBR optimization problem is solved by replacing the model characteristics by the effective characteristics.

We shall not pursue an example using effective characteristics. First, we want to avoid a lot of domain-specific mathematics, and second, we want to give an example that is appreciable to a wide audience. The mathematics is avoided when the uncertainty class and operator class are finite, so that finding an IBR filter reduces to a search among all possible operators to find the best. There is a price because the computational burden is typically too great, especially when we proceed to experimental design. Hence, in practice one must use some sort of complexity reduction, thereby obtaining an approximate solution. For instance, one might approximate the underlying scientific model to reduce the number of parameters. A general approach is to reduce the operator optimization from being over all operators in the operator class $\mathscr{F}$ by only considering operators that are optimal for some model in the uncertainty class (Grigoryan and Dougherty 1999). The resulting operator is known as a *model-constrained Bayesian robust* (MCBR) operator. These have been considered for various classes of operators.

22.6 IBR Intervention in Regulatory Networks

Consider uncertainty in the mammalian cell cycle network resulting from there being k pairs of genes for which it is known that there is a regulatory relationship but the type of relationship, activating or suppressing, is unknown. The uncertainty class consists of 2^k networks, each $\theta \in \Theta$ corresponding to a specific assignment of regulation types to the k uncertain edges. Since we have no knowledge beyond the existence of regulatory relations, the uncertainty class is governed by a uniform prior distribution $\pi(\theta) = 2^{-k}$. A structural intervention blocks the regulatory action between a pair of genes in the network and the cost function is the total undesirable steady-state probability. Based on the given mammalian cell cycle network, simulations have been run in Yoon et al. (2013) that incrementally increase the number of edges with unknown regulation from $k = 1$ to $k = 10$. In each case, 50 uncertain networks are created by randomly selecting uncertain edges while keeping the regulatory information for the remaining edges.

For the set of models with from 1 to 5 uncertain edges, 54.0% of the time the IBR structural intervention, which minimizes the expected undesirable steady-state mass, is the optimal intervention for the true network, which blocks the regulation from CycE to p27. As noted previously, when applied to the true model, this reduces the total undesirable steady-state probability to 0.2639. 41.6% of the time the IBR intervention blocks the regulation from CycE to Rb, and reduces the total undesirable steady-state probability to 0.2643. Four other interventions are chosen a total of 4.4% of the time. When the simulation is run with 6 to 10 uncertain edges, blocking CycE to p27 or blocking CycE to Rb accounts for 88.8% of the IBR interventions, as opposed to 95.6% of the IBR interventions for 1 to 5 uncertain edges. This change reflects the greater uncertainty.

22.7 Objective Cost of Uncertainty

While optimal over the uncertainty class, an IBR operator is not likely optimal relative to the true model. The loss of performance is the cost of uncertainty. To quantify this cost, for $\theta \in \Theta$, let ψ_θ be an optimal operator for θ. Then $C_\theta(\psi_\theta) \leq C_\theta(\psi_{\text{IBR}}^{\Theta})$. For any $\theta \in \Theta$ and operator family $\mathscr{F}$, the *objective cost of uncertainty (OCU)* relative θ is defined by

$$\mathrm{U}_{\mathscr{F}}(\theta; \Theta) = C_\theta(\psi_{\text{IBR}}^{\Theta}) - C_\theta(\psi_\theta). \tag{22.5}$$

If we knew the true network, then we could insert the corresponding value of θ in the preceding expression to find the actual objective cost of uncertainty; however, we do not know it. Hence, we take the expectation of the OCU over the prior distribution, which gives the *mean objective cost of uncertainty (MOCU)* (Yoon et al. 2013):

$$\mathrm{M}_{\mathscr{F}}(\Theta) = E_\Theta[\mathrm{U}_{\mathscr{F}}(\theta; \Theta)]. \tag{22.6}$$

The MOCU provides the desired uncertainty quantification. Historically, uncertainty has often been measured by the entropy of a distribution, but entropy ignores the translational objective. There may be large entropy but with most (or all) of the uncertainty irrelevant to the objective. For instance, in controlling a network there may be much uncertainty in the overall network but a high degree of certainty regarding the mechanisms involved in the control. In this case, the entropy might be large but the MOCU small, which is what matters from a translational perspective.

The MOCU can be used to design experiments to optimally reduce the uncertainty relevant to the operational objective. Suppose there are k experiments $T_1, \ldots, T_k$, where experiment T_i exactly determines the uncertain parameter θ_i in $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Question: Which experiment should be conducted first? Let $\theta|\bar{\theta}_i = \theta|(\theta_i = \bar{\theta}_i)$ be the *conditional uncertainty vector* composed of all uncertain parameters other than θ_i, with the experiment now having determined that $\theta_i = \bar{\theta}_i$. Let $\Theta|\bar{\theta}_i = \{\theta|\bar{\theta}_i : \theta \in \Theta\}$ be the *reduced uncertainty class* given $\theta_i = \bar{\theta}_i$. Elements of $\Theta|\bar{\theta}_i$ are of the form

$$\theta|\bar{\theta}_i = (\theta_1, \ldots, \theta_{i-1}, \bar{\theta}_i, \theta_{i+1}, \ldots, \theta_k). \tag{22.7}$$

The IBR operator for $\Theta|\bar{\theta}_i$ is denoted $\psi_{\mathrm{IBR}}^{\Theta|\bar{\theta}_i}$ and is called the *reduced IBR operator* relative to $\bar{\theta}_i$.

If the outcome of experiment T_i is $\bar{\theta}_i$, then the *remaining MOCU* given $\theta_i = \bar{\theta}_i$ is defined by

$$\mathrm{M}_{\mathscr{F}}(\Theta|\bar{\theta}_i) = \mathrm{E}_{\Theta|\bar{\theta}_i}[C_{\theta|\bar{\theta}_i}(\psi_{\mathrm{IBR}}^{\Theta|\bar{\theta}_i}) - C_{\theta|\bar{\theta}_i}(\psi_{\theta|\bar{\theta}_i})], \tag{22.8}$$

where the expectation is relative to the conditional distribution $\pi(\theta|\bar{\theta}_i)$. The remaining MOCU is the MOCU for the reduced IBR operator relative to the reduced uncertainty class.

The expected remaining MOCU given parameter θ_i is called the *experimental design value*, $\mathrm{D}(\theta_i)$. It is the expectation of $\mathrm{M}_{\mathscr{F}}(\Theta|\theta_i)$ with respect to the marginal distribution $\pi(\theta_i)$ over θ_i:

$$\mathrm{D}(\theta_i) = \mathrm{E}_{\theta_i}[\mathrm{M}_{\mathscr{F}}(\Theta|\theta_i)]. \tag{22.9}$$

An experiment T_{i^*} minimizing the experimental design value is called an *optimal experiment*: $\mathrm{D}(\theta_{i^*}) \leq \mathrm{D}(\theta_i)$ for $i = 1, 2, \ldots, k$ (Dehghannasiri et al. 2015b). θ_{i^*} is called an *primary parameter*. Putting the definitions together, it is not difficult to show that i^* minimizes the *residual IBR cost* for experiment T_i,

$$\mathrm{R}(\theta_i) = \mathrm{E}_{\theta_i}[\mathrm{E}_{\Theta|\theta_i}[C_{\theta|\theta_i}(\psi_{\mathrm{IBR}}^{\Theta|\theta_i})]], \tag{22.10}$$

which is the expectation over the possible outcomes of experiment T_i of the expected cost of the reduced IBR operator over the reduced uncertainty class. The primary parameter is the parameter corresponding to i^*, where i^* minimizes both $\mathrm{D}(\theta_i)$ and $\mathrm{R}(\theta_i)$, $i = 1, 2, \ldots, k$.

Rather than performing a single experiment, one can perform a sequence of experiments to iteratively reduce the number of uncertain parameters.

22.8 Optimal Experimental Design for Regulatory Networks

We now apply optimal experimental design to regulatory networks when there are uncertain regulations and the aim is to find the optimal regulation based on the MOCU. Consider a BNp in which network regulations are governed by a regulatory matrix **R** that characterizes the regulatory relations between every pair of nodes, and where nodes are regulated according to the majority vote rule. Uncertainty is introduced by assuming that for certain node pairs, although a regulatory relation is known to exist, the type of the regulation (activating or suppressive) is unknown. The uncertain parameters are the unknown regulatory relations. An uncertain parameter θ_i equals 1 for an activating regulation and -1 for a suppressive regulation. If there are k uncertain regulations, then Θ contains 2^k networks. Experimental design selects a primary parameter to optimally improve structural intervention.

Computational complexity is a major issue. The complexity of network intervention grows exponentially with network size. It is much worse when performing experimental design because the optimal intervention must be found for every network in the uncertainty class. Model reduction for structural intervention has been studied to reduce complexity (Dehghannasiri et al. 2015a) but we do not consider the issue here. To reduce computation, we utilize the MCBR intervention rather than the IBR intervention.

Proceeding with our example, simulations have been performed with $k = 2, 3, 4, 5$, assuming a uniform prior distribution over Θ (Dehghannasiri et al. 2015b). Networks have six nodes, $X_1, \ldots, X_6$, each having three regulators. A random BNp is generated by randomly selecting three regulators for each node with uniform probability and randomly assigning 1 or -1 to the corresponding entries in the regulatory matrix **R**. The perturbation probability is set to $p = 0.001$. States with $X_1 = 1$ are undesirable. For each k, 1,000 synthetic BNps are generated and 50 different sets of k edges (regulations) are randomly selected for each network. The regulatory information of other edges is retained while that of the k selected edges is assumed to be unknown.

Unlike real networks, which can be controlled to a certain extent, many randomly generated networks may not be controllable. Hence, regardless of the intervention, the resulting steady-state distribution shift may be negligible and the difference between optimal and suboptimal experiments insignificant. Thus, to examine the practical impact of experimental design, we must take controllability into account. We use the percentage decrease of total steady-state mass in undesirable states after optimal structural intervention as a measure of controllability:

Table 22.1 The average gain of conducting the optimal experiment predicted by the proposed experimental design strategy in comparison to other suboptimal experiments

Unknown edges	Average η_1	Average η_2	Average η_3	Average η_4
$k = 2$	0.0584	N/A	N/A	N/A
$k = 3$	0.0544	0.0718	N/A	N/A
$k = 4$	0.0545	0.0750	0.0855	N/A
$k = 5$	0.0474	0.0696	0.0803	0.0863

$$\Delta = \frac{\varphi_U - \varphi_U^{\psi}}{\varphi_U} \times 100\%, \tag{22.11}$$

where controllable networks have a larger Δ.

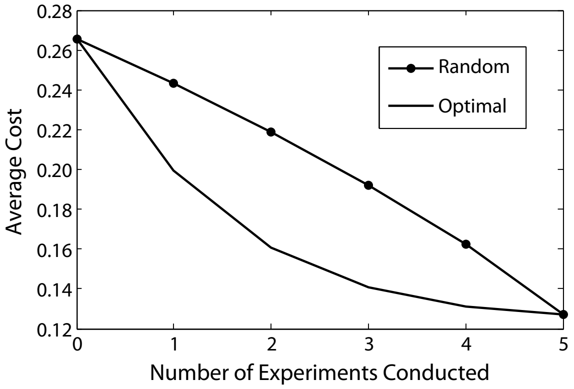
Rank the experiments $\langle 1\rangle, \langle 2\rangle, \ldots, \langle 5\rangle$ according to which provide the greatest reduction in expected remaining MOCU. By definition, experiment $\langle 1\rangle$ is optimal. For $i = 1, 2, 3, 4$, let

$$\eta_i = C_{\theta_{\text{true}}}\left(\psi_{\text{MCBR}}^{\Theta|\bar{\theta}_{\langle i+1\rangle}}\right) - C_{\theta_{\text{true}}}\left(\psi_{\text{MCBR}}^{\Theta|\bar{\theta}_{\langle 1\rangle}}\right) \tag{22.12}$$

be the cost difference between applying the MCBR intervention derived for the reduced uncertainty class that results from conducting the $(i + 1)$-ranked experiment to the true network and the cost of applying the MCBR intervention obtained from conducting the optimal experiment. Table 22.1 summarizes the average gain of performing the optimal experiment over other suboptimal experiments according to η_i. The average is taken over different sets of uncertain regulations and different networks with $\Delta \geq 40\%$.

Regarding sequential experimentation, Fig. 22.6 compares the average cost after performing k experiments iteratively chosen by optimal experimental design to the average when experiments are chosen randomly. The curves agree at the outset when no experiments have been performed, and at the end when all parameters have been determined. The key point is that the cost reduction is much greater when only one or two experiments are performed; indeed, the cost when two experiments have been chosen optimally is approximately the same as when four have been chosen randomly. This is because at each stage the experimental design procedure selects the experiment that will provide the maximal expected reduction of model uncertainty related to the operational objective. In practice, this means that one can perform a fraction of the possible experiments and eliminate most of the objective uncertainty.

Fig. 22.6 The average cost of robust intervention after performing the sequence of experiments predicted by the proposed strategy and the average cost after performing randomly selected experiments



22.9 Discussion

There are numerous Bayesian methods to handle inference in the presence of model uncertainty (Bernardo and Smith 2001; Clyde and George 2004; Madigan and Raftery 1994). One approach is Bayesian model selection, which aims to select the most probable model in an uncertainty class given observed data (Barbieri and Berger 2004; Wasserman 2000; Chen et al. 2003). Though an inferred model produced by Bayesian model selection, or any other model selection method, could be used to make predictions about a given quantity, the resulting prediction is expected to be suboptimal relative to that of the IBR operator that directly and optimally infers the same quantity.

Bayesian model averaging predicts a given quantity by finding the weighted average of predictions across an uncertainty class of models, where weights in the average are the posterior probability of the corresponding model (Hoeting et al. 1999; Raftery et al. 1997; Wasserman 2000). There are several key distinctions between Bayesian Model Averaging and IBR operators: (1) Whereas Bayesian Model Averaging implicitly assumes the minimum mean-square error cost function, IBR operators allow other cost functions. (2) While Bayesian model averaging is typically implemented with a finite number of models to control computational complexity, IBR operators leverage existing fixed-model optimal operator design methods to efficiently compute the average, often over a continuum of models. (3) The IBR framework summarizes all model uncertainty using either *effective processes* or *effective characteristics*. Effective processes are not required to be a member of the uncertainty class, or even valid stochastic processes, and effective characteristics distill model uncertainty even further to certain partial descriptors of the stochastic process. (4) Perhaps most importantly, while Bayesian model averaging and other standard Bayesian inference techniques place uncertainty on parameters of the operational model, e.g., the regression coefficients, the IBR framework quantifies uncertainty on scientific knowledge itself by placing priors directly on the underlying stochastic process.

There are also numerous Bayesian approaches to sequential experimental design (Ryan et al. 2016). However, again, most Bayesian sequential design methods quantify uncertainty on parameters of the operational model, rather than quantifying uncertainty in the science itself. An example is the knowledge gradient policy (Frazier et al. 2009), which assigns independent Gaussian rewards (sign-flipped cost) to each alternative experiment and correlated multivariate Gaussian beliefs to the mean values of these rewards. Although the knowledge gradient approach is a special case of the MOCU-based experimental design framework, there are fundamental differences: (1) under knowledge gradient the experiment space and action space must be the same, while under MOCU they may be different, and (2) under knowledge gradient uncertainty is modeled directly on the reward function and there is no direct connection between assumptions in the operational model and the underlying physical model, while under MOCU prior knowledge and uncertainty in the underlying physical system can be incorporated into the modeling framework.

Many Bayesian experimental design methods also quantify uncertainty using information theoretic entropy or other global measures that are not tailored to the objective. This is in contrast with the MOCU-based framework, which specifically aims to quantify the degree to which model uncertainty affects the engineering objective. To illustrate, consider work in Huan and Marzouk (2016), which presents a general review of Bayesian experimental design and almost immediately assumes Kullback–Leibler (KL) divergence as a design objective. On page 8, the authors state that KL divergence may be used as a "general-purpose objective that seeks to maximize learning about the uncertain environment" and "should lead to good performance for a broad range of estimation tasks." While this is no doubt true in many applications, there are perhaps few cases where experimental design is conducted in the absence of a more specific objective.

22.10 Conclusion

Our approach to uncertainty quantification has two salient aspects: (1) it is based on scientific uncertainty regarding the underlying random processes and (2) it quantifies the effect of scientific uncertainty on operator design. In particular, as opposed to Bayesian methods that place prior distributions on parameters of an operator model, such as standard Bayesian linear regression, here the prior is placed on the underlying model and therefore directly reflects our scientific uncertainty. Uncertainty in the operator follows directly from the form of the operator and uncertainty regarding the random processes involved. For instance, in IBR classification, prior distributions are on the parameters of the feature-label distribution, not on the parameters of the classifier. Not only does this place the mathematical formulation of the uncertainty where it belongs, it also facilitates the design of prior distributions directly from our scientific knowledge—for instance, deriving prior distributions for phenotype classification from known biological signaling pathways (Esfahani and Dougherty 2014; Boluki et al. 2017).

Objective-based optimal experimental design follows directly from the definition of the mean objective cost of uncertainty. Our interest in the scientific model lies solely in our ability to use it to facilitate operator performance. The curves in Fig. 22.6 clearly demonstrate the importance of determining the parameters most relevant to operator performance. In basing uncertainty quantification on MOCU, we are quantifying uncertainty relative to an operational objective. The aim is translational scientific knowledge (engineering) as opposed to general scientific knowledge, which might be the aim if one were to use entropy over the uncertainty class to quantify uncertainty. MOCU-based optimal experimental design has been applied in both biology and materials discovery (Mohsenizadeh et al. 2016; Dehghannasiri et al. 2017).

References

Bae, H.-R., Grandhi, R. V., & Canfield, R. A. (2004). An approximation approach for uncertainty quantification using evidence theory. *Reliability Engineering & System Safety*, *86*(3), 215–225.

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, *32*(3), 870–897.

Batchelor, E., Loewer, A., & Lahav, G. (2009). The ups and downs of p53: Understanding protein dynamics in single cells. *Nature Reviews Cancer*, *9*(5), 371–377.

Bernardo, J. M., & Smith, A. F. (2001). Bayesian theory. *Measurement Science and Technology*, *12*(2), 221.

Boluki, S., Esfahani, M. S., Qian, X., & Dougherty, E. R. (2017). Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors. *BMC Bioinformatics*, *18*(Suppl 14), 552.

Chen, M.-H., Ibrahim, J. G., Shao, Q.-M., & Weiss, R. E. (2003). Prior elicitation for model selection and estimation in generalized linear mixed models. *Journal of Statistical Planning and Inference*, *111*(1–2), 57–76.

Clyde, M., & George, E. I. (2004). Model uncertainty. *Statistical Science*, *19*(1), 81–94.

Dalton, L. A., & Dougherty, E. R. (2013). Optimal classifiers with minimum expected error within a Bayesian framework-Part I: Discrete and Gaussian models. *Pattern Recognition*, *46*(5), 1301–1314.

Dalton, L. A., & Dougherty, E. R. (2014). Intrinsically optimal Bayesian robust filtering. *IEEE Transactions on Signal Processing*, *62*(3), 657–670.

Dehghannasiri, R., Esfahani, M. S., & Dougherty, E. R. (2017). Intrinsically Bayesian robust Kalman filter: An innovation process approach. *IEEE Transactions on Signal Processing*, *65*(10), 2531–2546.

Dehghannasiri, R., Xue, D., Balachandran, P. V., Yousefi, M. R., Dalton, L. A., Lookman, T., et al. (2017). Optimal experimental design for materials discovery. *Computational Materials Science*, *129*, 311–322.

Dehghannasiri, R., Yoon, B.-J., & Dougherty, E. R. (2015a). Efficient experimental design for uncertainty reduction in gene regulatory networks. *BMC Bioinformatics*, *16*(Suppl 13), S2.

Dehghannasiri, R., Yoon, B.-J., & Dougherty, E. R. (2015). Optimal experimental design for gene regulatory networks in the presence of uncertainty. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *12*(4), 938–950.

Dougherty, E. R. (2007). Validation of inference procedures for gene regulatory networks. *Current Genomics*, *8*(6), 351–359.

Dougherty, E. R. (2016). *The evolution of scientific knowledge: From certainty to uncertainty*. Bellingham: SPIE Press.

Dougherty, E. R., & Bittner, M. L. (2011). *Epistemology of the cell: A systems perspective on biological knowledge*. New York: Wiley.

Esfahani, M. S., & Dougherty, E. R. (2014). Incorporation of biological pathway knowledge in the construction of priors for optimal Bayesian classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*(1), 202–218.

Faryabi, B., Vahedi, G., Datta, A., Chamberland, J.-F., & Dougherty, E. R. (2009). Recent advances in intervention in Markovian regulatory networks. *Current Genomics*, *10*(7), 463–477.

Frazier, P., Powell, W., & Dayanik, S. (2009). The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing*, *21*(4), 599–613.

Grigoryan, A. M., & Dougherty, E. R. (1999). Design and analysis of robust binary filters in the context of a prior distribution for the states of nature. *Journal of Mathematical Imaging and Vision*, *11*(3), 239–254.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401.

Huan, X., & Marzouk, Y. M. (2016). Sequential Bayesian optimal experimental design via approximate dynamic programming. arXiv preprint arXiv:1604.08320.

Kaufmann, S. (1993). *The origins of order*. New York: Oxford University Press.

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*(428), 1535–1546.

Mohsenizadeh, D., Dehghannasiri, R., & Dougherty, E. R. (2016). Optimal objective-based experimental design for uncertain dynamical gene networks with experimental error. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *15*(1), 218–230.

Qian, X., & Dougherty, E. R. (2008). Effect of function perturbation on the steady-state distribution of genetic regulatory networks: Optimal structural intervention. *IEEE Transactions on Signal Processing*, *56*(10), 4966–4976.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179–191.

Ryan, E. G., Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, *84*(1), 128–154.

Shmulevich, I., Dougherty, E. R., & Zhang, W. (2002). From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. *Proceedings of the IEEE*, *90*(11), 1778–1792.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107.

Yoon, B.-J., Qian, X., & Dougherty, E. R. (2013). Quantifying the objective cost of uncertainty in complex dynamical systems. *IEEE Transactions on Signal Processing*, *61*(9), 2256–2266.

Part VI
Methodology—The Organization and Management of Simulation Validation

Chapter 23
Standards for Evaluation of Atmospheric Models in Environmental Meteorology

K. Heinke Schlünzen

Abstract This chapter focuses on evaluation guidelines developed in the field of environmental meteorology. Definitions for verification, validation, and evaluation as used in the field of environmental meteorology are given. A generic structure of a model evaluation guideline is introduced consisting of three parts: (A) Specification of application area, (B) evaluation steps to be performed by the model developer, and (C) evaluation steps to be performed by the model user. The generic structure is detailed using two examples from environmental meteorology. For both examples, an accepted standard for model evaluation was achieved by involving the relevant stakeholders in the harmonization process. The methodology to achieve a standard and why standards are relevant in environmental meteorology is outlined.

Keywords Verification · Validation · Evaluation · Environmental meteorology · Guideline · Standard

Abbreviations

CCA-EM	Commission for Clean Air
EGa	VDI 3783 Part 7 (VDI 2017a)
EGb	VDI 3783 Part 9 (VDI 2017b)
LES	Large eddy simulation
MQI	Model quality indicator
MQO	Model quality objective
RANS	Reynolds-averaged Navier–Stokes

K. H. Schlünzen (✉)
Meteorological
Institute, Center for Earth System Research and Sustainability (CEN), Universität Hamburg, Bundesstr. 55, 20146 Hamburg, Germany
e-mail: heinke.schluenzen@uni-hamburg.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_23

23.1 Introduction

Evaluation of models and their results is of uttermost relevance in meteorology. This is true not only for weather forecast (see Chap. 29 by Theis and Baldauf in this volume), which is a daily product of meteorological services worldwide and used by different communities (e.g., industry, public). Evaluation is also a relevant topic for climate models (see Chap. 30 by Rood in this volume), since results of climate models are most relevant for planning mitigation and adaptation measures. In the field of environmental meteorology, evaluation is a long-standing issue, since also in this field model results are used to take decisions relevant for humans and the environment. Environmental meteorology investigates, broadly speaking, effects of anthropogenic changes of the atmosphere that influence the environment. The assessed environmental changes are often short distance (e.g., 1 km to a few hundred kilometers) and concern direct impacts on humans or on ecosystems. Global climate change studies would not be addressed, while local impacts of global change and locally driven local and regional climate changes are part of environmental meteorology studies. This is especially true for studies of the environmental impacts of intense local forcing, as found in urban areas resulting in heavy pollution and urban climate.

Models are a very important planning tool in environmental meteorology, for instance, for decisions on industrial settlement, in traffic planning or on urban development. All these decisions are costly in terms of both time and resources. Therefore, it is important that the models used are tailored for the tasks they are meant to do and that they deliver reliable results for these tasks. These ideas lead to the need for evaluation guidelines that should preferably become standards (Sect. 23.3). These standards shall ensure that the models applied by consultants provide reliable results. The standards are of interest to the scientific community as well, since they describe the current state of knowledge with respect to model evaluation. Therefore, they can also be used to ascertain quality of research models if, for instance, the test cases are included in a benchmark suite used to evaluate new model versions.

In the process of guideline development, it turned out that definitions for verification, validation, and evaluation were quite differently used in the environmental meteorology community. An agreement was achieved; the definitions are summarized in Sect. 23.2. A generic structure of an evaluation guideline is introduced in Sect. 23.4. Examples on how this generic structure is detailed in environmental meteorology are given in Sect. 23.5. Conclusions on limits of the current approach and future developments in the field of environmental meteorology are given in the conclusions (Sect. 23.6).

23.2 Definitions Used

In environmental meteorology, a longer discussion took place in the 1990s on the meaning of evaluation, validation, and verification that eventually lead to agreeing on definitions as given in this section. Additional definitions of words frequently used in this chapter are added but have not been intensely discussed in the environmental meteorology community.

In the following, citations of previously published papers are used. Additions to the original citation are given in square brackets, and words omitted marked with dots.

23.2.1 Specifics of an Atmospheric Model

23.2.1.1 Model and Program

A model shall represent, in a simplified but physically sound way, the real atmospheric processes and relations "… complying with fundamental physical principles and using fundamental physical equations, approximations, parameterizations and certain boundary conditions." (VDI 2017a). It is further noted in VDI (2017a) that "the systems of equations of the models … are solved with the help of numerical methods with specified boundary and initial conditions."

For the realization as computer code, mostly newest versions of the programming language Fortran are used. Fortran is well suited for numerical models and scientific computing, and the compilers produce well-performing programs for the types of equations, number of grid points (10^6 to 10^8 per variable), and number of time steps (~10^6 to 10^7) solved in atmospheric models. By the word "program" as used in this chapter, the "implementation of the model on a computer" (VDI 2017a) is understood. This includes the executable on a specific computer.

The thinking of model and program being combined is relevant for the evaluation of a model, which is not seen independently from its realization.

23.2.1.2 Scale

Characteristic size and lifetime of an atmospheric phenomenon determine the spatial and temporal scale, respectively. Examples of the scales are included in Fig. 23.2. The scales determine which approximations of the fundamental equations might be made and which phenomena a model should be able to simulate.

23.2.1.3 Averaging Approaches

The fundamental equations are averaged in space and time (filtering of equations) since current computers are not able to resolve all atmospheric phenomena to the detail. Due to this filtering, phenomena of a specific scale are no longer resolved. Two filtering approaches are used in atmospheric models. Large eddy simulation models (LES models) use a spatial filtering and temporally resolve developments of vortices with a temporal resolution consistent with the time step used for numerical reasons. Models using the Reynolds-averaged Navier–Stokes equations (RANS models) filter in space and in time. The spatial filtering is the same as in the LES model of same resolution, but the remaining timescales are larger in the RANS model. In a RANS model, subgrid-scale turbulent fluxes are parameterized (Sect. 23.2.1.4). This parameterization is set up in a way that it filters fast changes, so that the characteristic temporal resolution of a RANS model is about 10–20 min. The actual spatial and temporal resolution of an atmospheric model needs to be determined per case. This can be done by comparing spatial and temporal spectra of model results with those derived from measured data.

23.2.1.4 Parameterization

A parameterization is used in a model to consider the effects of influential processes that cannot be resolved in a model with the applied resolution or averaging approach. The effects are described as a function of the simulated variables using some constants or parameters that may also depend on the simulated variables. To give a well-known example, the subgrid-scale turbulent momentum fluxes are frequently parameterized in dependence of the spatial gradients of the wind field and the exchange coefficient. The exchange coefficient may again depend on the spatial gradients of the flow field and on the atmospheric stratification (i.e., how temperature and humidity change with height).

23.2.2 Modeling

With "modeling" not only the model and its realization are meant but also the use of the realized model is covered. This use includes the preparation of input data, the use of the model by its user, and how the output data are further refined (e.g., interpolation in space and/or time). Thus, modeling consists of two parts:

I. Theoretical basics and realization of the model in a program: This includes the basic equations, approximations, parameterizations, in some cases chemical mechanisms, boundary conditions, initialization method, numerical schemes (incl. discretization), programming language, internal error checking, and detection of user errors by code internal checks;

II. Model use: The user of a model specifies the initial data and how the model shall be initialized, which boundary values and which time steps are to be used. Also, the specification of the output times may be relevant for result use. The user is also responsible for the correctness of the input data (e.g., surface characteristics), for specifying the domain size (large enough?) and the grid resolution (uniform/nonuniform). The model user needs to determine, if the phenomenon to be simulated can be simulated with the selected model (is the model applicable?). The technical surrounding and the experience of the model user are further influencing the model results.

Since how a model is used (part II) impacts the model results, even a perfect model can deliver wrong results ("garbage in—garbage out"), if the user employs the model in a wrong way. Thus, user training is very relevant for model application and model results. Training exists, mainly at the meteorological services and universities with meteorology lectures. It reaches from a few days for model output users (Geertsema et al. 2018) to several months for model developers (e.g., lectures and postgraduate studies at universities and research institutions).

23.2.3 Guideline

A guideline recommends how to act, in this case when evaluating a model. It helps to harmonize different approaches and aims at setting sound practices. A guideline is not mandatory or legally binding and thus cannot be enforced.

23.2.4 Standard

A standard is a guideline that was accepted by a broader community (involvement of relevant stakeholders) and therefore is setting a norm. Developing a standard often involves compromises between scientifically based wishes and practical needs and practicability.

23.2.5 Verification

VDI 3783 Part 9 (VDI 2017b) defines verification as follows: "Confirmation that the program corresponds to the model." This definition is checking for the realization of the model in the program but is not including the check for the realism of model results. This, however, would hardly be possible: Based on Popper (1982), Schlünzen (1997) outlines this for atmospheric models: "To verify a model completely, it has to be proved that the model is able to simulate all atmospheric phenomena of the model application area with the correct solution. Such a proof could only be constructed if we knew all atmospheric phenomena of the application area, could simulate them all, and could successfully compare model results with measurements. Apart from the fact that it is not at all a simple task to compare model results and measurements, a proof is impractical for complex models, because we neither know all the atmospheric phenomena nor all the initial data. The impossibility of a proof becomes even more evident if one considers that all possible variations of input parameters have to give realistic results. Therefore, complex atmospheric models cannot in general be verified completely. However, model results can be verified for single-case studies, e.g., by comparing them with measurements. In addition, a verification of single aspects of a model or of a simple model might be possible." Since the model verification is practically impossible, VDI (2017b) restricted the definition to software verification.

23.2.6 Validation

VDI 3783 Part 9 (VDI 2017b) defines validation as follows: "An examination to find out to what extent (with what accuracy) a program describes within the formal scope of the model the phenomena it is meant to model." Similar to verification, it is difficult to perform this in general for a model. However, for single cases, this can be achieved and, therefore, the application of a model to a single situation and its comparison with reference data is often named validation in environmental meteorology.

23.2.7 Evaluation

VDI 3783 Part 9 (VDI 2017b) defines evaluation as follows: "Assessment of a model and of the associated program with regard to accuracy." This means that theory and model basics are not separated from the model realization in a computer code. Both have to be jointly evaluated with respect to model performance characteristics. As detailed in Baklanov et al. (2014, Sect. 6.3), "The aims of model evaluation are to assess the suitability of a model for a specific application ('fit for purpose'); benchmarking model performance against reality and other models; quantifying uncertainties; testing individual model components; and providing guidance for future model developments."

Based on a paper by Dennis et al. (2010), different types of evaluation can be distinguished: operational, diagnostic, dynamic, and probabilistic evaluation. For a full evaluation of a model, all evaluation types should be considered.

23.2.7.1 Operational Evaluation

This "involves the direct comparison of model output with routine observations … using statistical metrics such as normalized mean bias, root mean square error, etc." Baklanov et al. (2014, Sect. 6.3).

23.2.7.2 Diagnostic Evaluation

The diagnostic evaluation "…examines individual processes and input drivers that may affect model performance and requires detailed … measurements that are not, typically, routinely available." Baklanov et al. (2014, Sect. 6.3).

23.2.7.3 Dynamic Evaluation

The dynamic evaluation "… investigates the model's ability to predict changes … in response to changes in … [drivers (input or boundary values)]. …" Baklanov et al. (2014, Sect. 6.3).

23.2.7.4 Probabilistic Evaluation

The probabilistic evaluation "…explores the uncertainty of model predictions and is used to provide a credible range of predicted values rather than a single estimate. It is based on knowledge of uncertainty embedded in observations and model predictions, the latter often being approximated by an ensemble of model simulations." Baklanov et al. (2014, Sect. 6.3). Thus, here the uncertainty of model results (e.g., from ensemble simulations) as well as the uncertainty of reference data (here observations) are considered.

23.2.8 Model Quality Indicator

A model quality indicator (MQI) characterizes the quality of a model result in comparison to reference data. A summary of those used in environmental meteorology can be found in Chap. 5 of Schlünzen and Sokhi (2008) and Appendices D-F therein (see also Chap. 13 by Marks and Chap. 18 by Saam in this volume). It should be noted that many of the statistical metrics used as MQIs assume Gaussian distributions of the compared values (model results and reference data) or at least of the differences. However, often the differences are not normally distributed (Fig. 23.1).

The question arises what should be reflected by an MQI. If the interest is merely in checking how close model results and reference data agree, then the percentage

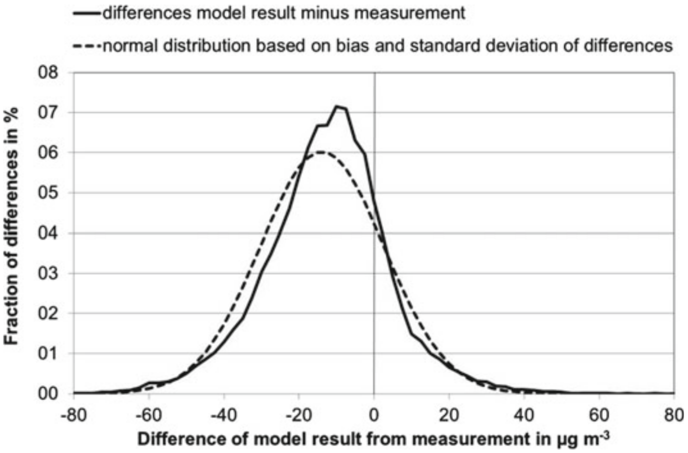


Fig. 23.1 Frequency distribution of differences between model results and reference data (straight line) taking NO_2 data as example. Using the bias (here $-13.85\ \mu g\ m^{-3}$) and the standard deviation (here $16.55\ \mu g\ m^{-3}$) of the differences, the frequency distribution is also given as normal distribution (dashed line)

of differences might be determined that are in an acceptable range. This is expressed by using the hit rate H (Cox et al. 1998; Schlünzen and Katzfey 2003):

$$H = \frac{1}{N}\sum_{j=1}^{N}\begin{cases} 1 \text{ for } \left|M_j - O_j\right| < A, \text{ or } \left|\frac{M_j - O_j}{O_j}\right| < D \\ \qquad 0 \text{ else} \end{cases} \tag{23.1}$$

In Eq. (23.1), M_j denotes an individual model result and O_j the corresponding reference value taken from a comparison data set with N values. This MQI checks for absolute and relative differences. It is assumed that the model result is correct (hit), if either the absolute difference remains below the given value A or the relative difference is below a given value D. Both values should be prescribed depending on the accuracy of the available reference data so that their uncertainty can be considered in the calculation of MQI. The allowed differences A, D can be larger, if less accuracy is needed from the model. By calculating hit rates, two advantages are achieved: the MQI is independent of the Gaussian distribution, and the reference data uncertainty can be considered in the evaluation, which is hardly possible in most other MQIs.

To give an example: Assuming an absolute difference and thus an uncertainty of the measured data of $A = 10$ and (for simplicity) a relative difference of $D = 0$ one receives from the distribution shown in Fig. 23.1 a hit rate of $H = 37.5\%$. This is higher than from the corresponding Gaussian distribution that fits the very same values of bias and standard deviation ($H = 34.9\%$). In this case, the agreement of measured and simulated values is better if one uses the correct distribution and not the Gaussian distribution. If the errors were less frequent within the assumed allowed absolute difference A, the values calculated on the basis of a Gaussian distribution would show a better agreement. This discrepancy can result for the same values of bias and standard deviation, since they do not account for a non-Gaussian error distribution. This uncertainty is inherent to all MQIs mentioned before, if the differences are not normally distributed.

One has to be aware that the deviation from a Gaussian error distribution might lead to wrong indications of the model performance, since the MQIs measure the model results' quality in correspondence to that type of distribution. In the example given, the MQIs based on the normal distribution indicate worse goodness of the model results than actually achieved. The possibility that the differences of model results and reference data are not normally distributed holds for all evaluation types mentioned in Sect. 23.2.7. Therefore, the use of MQIs independent of the normal distribution is preferable. Furthermore, data often cannot be quality checked due to time constrains. Then, they have a higher uncertainty, and a good agreement of model results and data should not be expected. In any case, the MQIs should consider uncertainty of the reference data.

23.2.9 Reference Data

Reference data are those data the model results are compared with (see Chap. 15 by Murray-Smith on data in this volume). These reference data can be derived from analytic solutions, plausibility assessments, other model results, or observed data. All of these data are not perfect, but include, besides other errors, conformity problems (e.g., model complexity and analytic solutions), uncertainties (e.g., representativeness, aliasing), and instrumental problems (e.g., measurable minimum or maximum values). All these uncertainties influence resulting values for the MQIs. The uncertainties need to be considered when selecting and calculating MQIs.

23.3 From Guidelines to Standards

23.3.1 Historical Background

Research in environmental meteorology was intensified after a heavy pollutant episode more than half a century ago, a 5-day winter smog episode in London in 1952 during which ~4000 people died (GLA 2002). Thereafter, first models (analytic solution of the dispersion equation, Gaussian plume models) were developed. These developments were less motivated by scientific curiosity, but more by the need to improve the environmental conditions for the people, in this case by understanding dispersion and reducing air pollution. Already at that time, results were compared with available field data which were taken from dedicated field experiments like those in 1957 by Hay and Pasquill. Many more field experiments and dedicated measurements as well as improved model approaches followed in later decades, and comparisons of model results with reference data were performed. Mostly, the outcome of these comparisons was visually evaluated as “Model results and measurements agree well.” in published papers, without using any quantitative model quality indicator. However, in the past 20–30 years, quantitative MQIs were more and more introduced to assess the abilities of current models. A summary of evaluation outcomes in the research field of environmental meteorology showed that, in example, 50% of the model evaluations have a bias in temperature between −1.1 K and +0.3 K, a correlation coefficient of above 0.62 for wind speed or a root mean square error of 71° for wind direction (Schlünzen et al. 2016). The evaluation approaches used in the analyzed scientific publications were (and many still are) different from test case to test case. A general agreement on how to evaluate atmospheric models applied to environmental problems is yet not achieved by the scientific community.

To ensure that concentration and deposition values neither affect human health nor ecosystems, standards for air quality have been introduced (EC 1980). In 1980, the European Communities Programme for Action on Environment set the limit level for sulfur dioxide concentrations among EU member states (EC 1980). For US, Canada, Japan, and other countries, similar limits exist. The introduction of limit values for more pollutants and the lowering of the limit values (e.g., EC 2008) increased measurement activities to monitor concentration levels. At the same time, it became clear that impacts of emissions into the atmosphere that result from new developments (e.g., industrial plants, roads, harbors, and livestock farming) need to be assessed in advance to minimize negative effects for the environment. Following EU regulations, models can be applied for this assessment, if they fulfill some quality criteria and if expected concentrations are much lower than the limit values. However, detailed hints on how to evaluate models were not given.

Several models have been developed and are used for assessment studies. Attempts to harmonize the modeling approaches started already in 1991 (HARMO conferences; Olesen 2017). The first complete evaluation protocol for atmospheric mesoscale models tailored for pollutant transport studies and using grid sizes of 500 m to 5 km was suggested by Schlünzen (1997), leading to generic protocols (Schlünzen, and Sokhi 2008; Schlünzen et al. 2018) as well as detailed evaluation concepts for high-resolution atmospheric models (Di Sabatino et al. 2011a, b; Franke et al. 2011). Some of these approaches were further developed and became national standards (e.g., VDI 2017a, b). Several guidelines have also been developed and are applied in US, Japan, and Germany for regulations relevant for the atmospheric environment (Meroney et al. 2016). Currently, a European-wide approach intends to develop a standard for the evaluation of air quality assessment models (Nordmann et al. 2017). This shall enhance result comparability throughout the different states of the European Union.

To summarize, standards for model evaluation are developed for the following five reasons:

(1) A quantitative assessment of an actual situation or of scenarios, e.g., of a future situation, is needed.

(2) Assessment cannot be done reliably by using common sense but need complex models that are not understandable and verifiable (definition see Sect. 23.2) at first glance.

(3) The model results have (economic) relevance for different stakeholders (individuals, community, and developers).

(4) There are regulations that put the model results in a legal framework with clear decisions to be taken by the management of a company or the administration of a community.

(5) Different results of the same scenario that have been achieved with different methodologies might trigger lawsuits and delay further development of a community.

Thus, evaluation standards become necessary if the model results are used in a legal and economic framework. However, to become a standard acceptance by the scientific community is not sufficient. A guideline can only become a standard if all relevant stakeholders are involved in the development and they finally agree on and accept a guideline as a standard.

23.3.2 How to Achieve a Standard

In order to illustrate how a standard can be achieved, the development of standards within the environmental meteorology air quality division of the Commission for Clean Air (shortened CCA-EM hereafter) is given. The CCA-EM standard development takes place within the Association of German Engineers (VDI) and the German Institute for Standardization (DIN). All standards are in German and English. The corresponding European-wide standardization organization is CEN (European Committee for Standardization) and worldwide it is ISO (International Organization for Standardization). Within CCA-EM currently, around 70 standards are in development or use, plus 2 and 7 developed in the framework of CEN and ISO, respectively.

The development of a standard for application in environmental meteorology needs to involve the relevant stakeholders, which are all experts. They come from executive boards (e.g., state administrators), from industry, as well as from science or are consultants involved in environmental impact assessments. These experts should be from diverse enough groups so that they represent the relevant stakeholders. They form the working group of that guideline (5–20 people) and jointly prepare a draft of a standard.

Typical development times for drafting a standard are 3–5 years, but it can be much faster (VDI 2018) or take even longer (VDI 2017a). A main reason for different development times is the gap in ascertained knowledge about the specific methodology that needs to be standardized. Another reason is the limitation in resources, since the work has no basic funding for the experts involved in the different working groups.

Once an agreed draft of the standard exists, this is sent by the CCA-EM secretariat to a wider circle of interested parties. They function as external reviewer board in a similar way as in a scientific journal, but they do mostly not express scientific disagreements but more practical problems or possible misunderstanding in the draft provided. Furthermore, they can object to specific methods or regulations suggested. The interested parties can formulate their objections and comments, and send them back within a specified deadline (about 2–3 months) to CCA-EM secretariat, which then ensures that the working group of that standard addresses all comments, improves the draft of the standard correspondingly, and provides individual answers to all reviewers. This review process can easily deliver 50 minor to major comments, even lead to a rejection of the draft of the standard. All objections and comments have to be addressed if they are not rejected for sound reasons. The draft of the standard has to be improved correspondingly. If the changes are severe and change the meaning of the standard, a second review and objection round is foreseen; otherwise, the changed standard will be published and thereby becomes a so-named VDI standard. In rare cases, the objections initiate a complete rewriting of the standard.

Once a standard is accepted and published, it does not mean this standard is applied. For this, it has to become part of a legal regulation. As an example of an environmental regulation, Appendix 3 of the German regulation TA Luft (2002) is taken. The requirement of using a Lagrangian particle model for dispersion calculations is specified there. The model has to be consistent with VDI 3945 Part 3 (VDI 2000). This ensures that environmental assessments are based on the same method at least for the dispersion simulation. Further standards are available concerning the meteorology fields (VDI 2017a, b), model use (VDI 2015), and assessment setup (VDI 2010) and might be taken up in a future update of the German air quality regulations (BMUB 2016).

23.4 Generic Structure of an Evaluation Guideline

Guidelines for model evaluation are not only helpful for standards (e.g., VDI 2017a, b), but are also of interest for the scientific community. A model developer might identify model shortcomings much faster by using standardized test cases that come with robust and tested reference data and model quality objectives (MQOs). Test cases can also be used as benchmarks for checking code errors in the process of the further continuous model development (see Chap. 18 by Saam on validation benchmarks in this volume). Model users have the advantage that the abilities and limits of a model are well documented once a model is evaluated with a standard (at least for the application area the model has been checked for). And, last not least, scientific publications based on model results of evaluated models will include fewer model results which lack model quality.

A comparable method for evaluating models also allows comparing the performance of different models. This helps to distinguish general shortcomings of many models from deficits of a single model. General shortcomings hint on deficits in our scientific understanding, while deficits in single models mostly hint on needed improvements of that one model.

A generic structure of an evaluation guideline is outlined in the following. It can be applied to all types of models; it consists of three parts:

A. Specification of application area,
B. Evaluation steps to be performed by the model developer, and
C. Evaluation steps to be performed by the model user.

The specification of the application area (Part A of generic structure; Sect. 23.4.1) is relevant since a model can hardly be verified or validated in general (Sect. 23.2). A model may, however, be evaluated for one or several specific application areas by the model developer; the limitations always need to be clearly stated (Part B of generic structure, Sect. 23.4.2). Since even validated models can deliver unrealistic results, hints need to be given to the model user as well (Part C of generic structure, Sect. 23.4.3).

23.4.1 Specification of Application Area

The target variable for the model's application (e.g., temperature, population growth) as well as the application type of the model (e.g., single-case versus statistical averages; forecast versus assessment) do not only determine what needs to be checked in the guideline but also the model's theoretical basics, the scales to be considered, and applications to be evaluated (Sect. 23.4.2). It also should clearly be stated in the application area where the limits of the evaluation guideline are, thus it needs to be specified what might be outside the application area of a model successfully evaluated following the specified guideline.

23.4.2 Evaluation Steps to be Performed by the Model Developer

This part of the evaluation is grouped into three parts, general evaluation, scientific evaluation, and test cases used as benchmarks.

23.4.2.1 General Evaluation

The general evaluation is not specific for an application area: the model should be comprehensible, meaning it should be documented, a third party should be allowed to inspect the code, peer-reviewed publications of model results and on model theory should exist. The documentation should consist of a brief description in the type of a data/fact sheet, a detailed model description including model theory, a manual that also includes installation aspects, and an evaluation report that summarizes the evaluation outcomes. If the program sources are used by externals, an evaluated model should also come with a technical reference specifying, e.g., programming conventions, variable names, etc.

It should be noted that all these above parts, with exception of the publication in peer-reviewed journals, are in addition to other scientific work and time-consuming. However, for reliable scientific results, not only publications but also the model documentation is important, even if a model is used only by the model developer. The documentation will support the continuous use and further development of the code since by documenting it, violations of the coding norms or error-prone realization can more easily be recognized by the developer. However, to achieve this documentation status for models and thus a higher code quality, this part of scientific work needs to be valued higher and thereby get a higher reputation in the scientific community.

23.4.2.2 Scientific Evaluation

For the scientific evaluation, the theoretical requirements on a model are specified based on the scientific knowledge. In consideration of the application area of the model evaluation guideline (Sect. 23.4.1), these requirements could concern the spatial and temporal resolution of the model, necessary output parameters, equations to be solved, theoretical concepts to be applied, solution methods, boundary values to be used, and the initialization method.

The criteria for the scientific evaluation should be specified by a group of experts for the respective application area. It should include the state of the science, but be open to new developments to not hinder any scientific progress in the application area. This might happen if, for instance, an evaluation guideline would be chosen by a funding agency as a pre-condition for a research project, and the guideline is very specific and detailed with respect to the characteristics of a model. In that case, the guideline might help to perpetuate outworn modeling methodologies. Thus, the scientific evaluation should only include criteria which are supported by the scientific community and that are open to new scientific developments.

23.4.2.3 Test Cases

The test cases have to be specified in detail to validate the model. These test cases should be relevant for the intended application area and should check the target variables. The test cases should be selected to cover the whole solution space of the application area. Even so it is impossible to check the validity of a model in the whole solution space (even more impossible to verify it), a thoughtful selection of test cases sampled from the solution space could sufficiently reflect possible solutions. One way is to select the test cases so that they check for possible extremes: Assuming the solution space to be a cube, the solutions at the corners and at some arbitrarily taken points within the cube could be chosen. Another way is to tailor the test cases to check for typical model shortcomings, e.g., horizontal homogeneity or stationarity, if these criteria fit the application area specified in the evaluation guideline. In practice, a mixture of test cases from both groups is chosen. Also, the test cases should be selected to perform the different evaluation types (Sect. 23.2.7) so that not only operational evaluations of the model are performed, but also reasons for the results (diagnostic evaluation), scenario reliability (dynamic evaluation), and results uncertainty are investigated (probabilistic evaluation).

All these different requirements call for many test cases. However, often the time needed to develop a test case is so large and the following application time so long that the number of test cases is quite limited in reality (5 to 20 test cases). Therefore, the test cases are mostly selected to detect typical model shortcomings. Consequently, the model user has a high responsibility for checking the model results (Sect. 23.4.3).

For each test case, a general description with the specifically checked model quality should be given. This will help to detect model shortcomings based on checking a specific model quality. The test case should be described with all details, e.g., domain size, resolution to be used, topography in the domain, initialization time, initial values, forcing values, boundary values, output needed, etc. In summary, all information needed to reliably run and evaluate a model should be given. The test case description should also include details on the reference data, and at best they should be provided with the guideline. Furthermore, details should be given how to compare model results and reference data. This includes MQIs (e.g., correlation coefficient r, hit rate H) as well as model quality objectives (MQOs). The last states how large (or small) an MQI has to be in order to comply with the test case criterion (e.g., $r > 0.9$, $H > 66\%$, $-2\ \text{K} < \textit{bias} < 2\ \text{K}$).

23.4.3 Evaluation Steps to be Performed by the Model User

As mentioned before, even a perfect model might produce wrong results especially if the model user is not trained. Therefore, an evaluation guideline should also include recommendations that have to be considered by the model user. Reasons for additional evaluation steps to be taken by the model user are manifold: The application intended by the model user might not be covered by the evaluation guideline applied by the model developer for evaluating the model. In addition, the impossibility to completely verify a model but only being able to falsify it (Popper 1982; see also Chap. 6 by Beven and Lane on falsificationism in this volume) might lead to arguable results even within the application area. Furthermore, every model solution depends on (uncertain) initial and boundary values, which might be wrongly chosen by the model user. Thus, model applications have to be verified or at least evaluated by the model user.

In a guideline, it should be specified how a user might prepare input data of sufficient quality, what needs to be considered when setting up model grid and domain, and how it can be ensured that a model solution in the focus area is sufficiently independent of selected domain characteristics as well as initial and boundary values. It should also be specified what evaluations a model user shall perform for each case the model is applied to, as routinely done by meteorological services (see Chap. 29 by Theis and Baldauf in this volume). Last not least, hints have to be given how the model simulations performed by the model user are to be documented so that they could be reproduced and repeated by any other model user. Furthermore, some test cases should be specified for training of the model user.

23.5 Examples for Standards

As already outlined in Sect. 23.3, several standards exist in environmental meteorology. The following example is taken from guidelines evaluating flow fields (VDI 2017a, b; Schlünzen et al. 2017) that are needed as input for atmospheric dispersion studies.

23.5.1 Comparing Application Areas of Two Standards

The application areas of VDI 3783 Part 7 (shortened EGa hereafter) and VDI 3783 Part 9 (shortened EGb hereafter) are outlined in Fig. 23.2 with respect to the phenomena that have to be simulated by the models (in italics in Fig. 23.2). While environmental meteorology investigates atmospheric phenomena of all scales, the two standards taken as examples here focus on microscale to mesoscale phenomena. Spatial scales are a few kilometers (EGb) or up to 100 km (EGa). Both standards are developed for atmospheric models using Reynolds-averaged Navier–Stokes equations. Thus, typical model results have an inherent time filter of 10–20 min as a result of the parametrizations employed and the boundary values used (Sect. 23.2.1.3). This is independent of the time step needed for numerical reasons or the output frequency of the model result. Typical forecast times as considered in the standards are hours (EGb) to days (EGa).

Models evaluated by using EGb are assessed toward their ability to simulate building wakes effects. If a model is evaluated by employing EGa, this model shall simulate katabatic flows, urban heat islands, orographic, and sea-breeze effects (phenomena in italics in Fig. 23.2). Effects of phenomena smaller than the scale resolved by the model need to be parameterized (e.g., turbulence, and in addition building effects in EGa). Influences of phenomena larger than the model domain and timescale have to be considered via the boundary values. In environmental meteorology assessments, a time slice approach (Schlünzen et al. 2011) is often used; necessary time series are combined from stationary solutions.

As mentioned before, standards EGa and EGb both assume to use models solving Reynolds-averaged equations. Therefore, they are (in theory) not applicable to the evaluation of large eddy simulation (LES) models. Letzel et al. (2012) compare in their paper results of an obstacle resolving LES model with wind tunnel data. Since

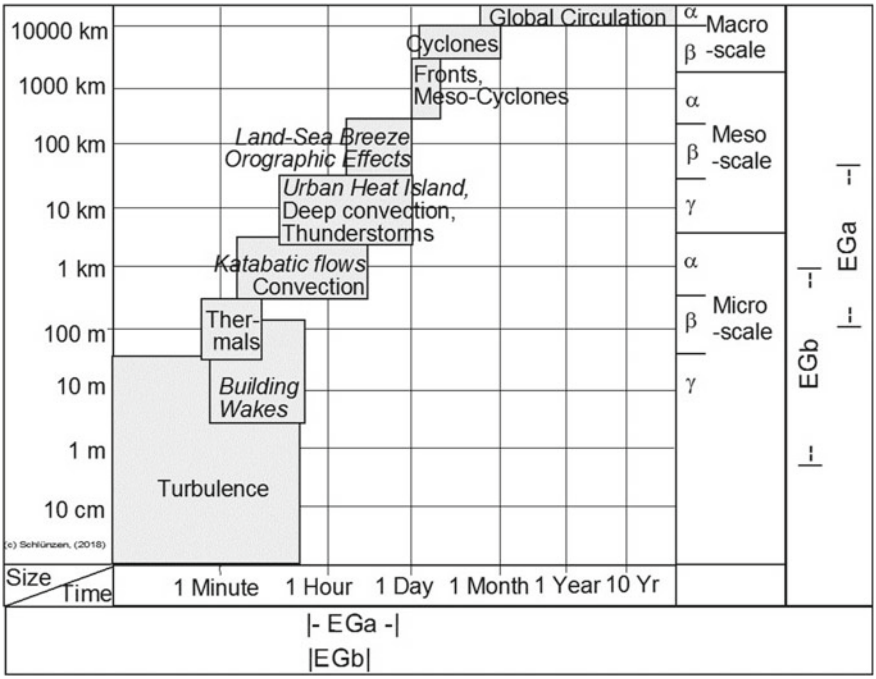


Fig. 23.2 Possible application areas (in italics) of different wind field evaluation standards with respect to atmospheric phenomena of characteristic horizontal size (extension) and characteristic time scales (lifetimes). EGa denotes VDI 3783 Part 7 (VDI 2017a), EGb VDI 3783 Part 9 (VDI 2017b). For more details, see text. Scales based on Schlünzen (1996)

they compare time-averaged values (2 h average), EGb might also be employed for assessing time-averaged LES model results. Then, the resulting time-averaged values are evaluated, but not the statistics that LES models can additionally provide.

23.5.2 Detailed Specification of an Application Area

Details below are only taken from VDI 3783 Part 7 (EGa). Even with this restriction, these examples represent only a small number of the specifications given, since this standard has 83 pages and includes details that are beyond the scope of the present more general paper. The examples selected shall provide some ideas for the development of standards in other fields of science. They are also chosen to illustrate that a standard can be detailed without hindering scientific development (see Sect. 23.4.2 on scientific evaluation).

In EGa, it is specified by the 10 experts involved in the development of the draft, in which cases the standard should be applied: The terrain slope should be larger than

1:5 (>11°). This is specified, since for smaller slopes more simplified approaches may be used. An alternative reason to apply models evaluated by EGa is the interaction of flow fields resulting from temperature gradients, e.g., katabatic flows. Also, in cases of time-dependent flow fields, models evaluated using EGa should be employed.

The application area is specified in EGa as atmospheric flow fields above structured terrain up to a height of 200 m above ground in domain sizes of a few kilometers to 100 km (Fig. 23.2). Since the target is the flow field, the evaluation focuses on the wind (components, or speed and direction). The standard is developed for evaluating Reynolds-averaged modeling approaches. Cloud formation is not considered, but a supersaturation by 10% is permitted.

Limitations of the standard are given as follows: thermodynamic values are calculated, but not explicitly evaluated (Sect. 23.5.3). If the horizontal grid size is below 100 m and the extension of buildings or high vegetation is larger than half the grid size, then the chosen grid needs to be justified—not explicitly mentioned is that additional evaluation is needed in that case. If temperature or other thermodynamic values are needed, additional evaluations have to be performed that are dedicated to these thermodynamic variables.

23.5.3 Some Detailed Evaluation Steps to be Performed by the Model Developer

23.5.3.1 General Evaluation

The general evaluation of EGa is prescribing all documentation given in Sect. 23.4.2. A minimum of two peer-reviewed publications of the model physics or its results is defined. These publications have to be in two different professional journals, but one can also be a doctoral or postdoctoral thesis.

From a scientific point of view, the number of publications is very low. The chosen minimum number is a compromise between a wish for a multitude of external model evaluations, as peer-reviewed papers can provide, and the desire for remaining open to new model developments in the field of environmental sciences. If the limit were higher, it might take too long before new model approaches could be used in practice.

23.5.3.2 Scientific Evaluation

The group of experts specified the theoretical requirements on the models to be fulfilled when using EGa: The fundamental equations that need to be solved are the conservation of mass, energy and momentum equations as well as the ideal gas law. The equations can be used in derived form (e.g., vorticity) or in basic form (for wind or momentum components), but all three wind components need to be calculated prognostic and be Reynolds-averaged. A temperature measure, e.g., potential temperature and specific humidity, should be computed with prognostic equations. Rotation of the Earth resulting in Coriolis force has to be considered. Acceptable approximations are given as well; these are mainly the anelastic approximation and the Boussinesq approximation and, for near-surface turbulent fluxes, the Monin–Obukhov similarity theory. Processes not directly solved by the model should be parameterized. Stability dependence of turbulence and for subgrid-scale convection has to be considered. The continuity of the parameterized turbulent fluxes has to be ensured. Concerning buildings and other land uses, their properties have to be considered in the model, e.g., by using a roughness length approach. The minimum requirement for solving the surface energy and humidity budgets is to apply a force-restore method (Deardorff 1978) and a budget equation, respectively. Since orography is considered, the effects of slopes on incoming radiation as well as shading effects by hills need to be addressed. Some additional model properties are mentioned, since they are frequently used, but are not a requirement to apply EGa. These include the use of a nonuniform model grid, the symmetry of the shear stress tensor, and the use of multilayer soil models for the surface energy and humidity budgets.

The scientific evaluation does not restrain new model developments in the field of environmental meteorology, since the restrictions are small: while the basic equations are set and a need for considering Coriolis force explicitly mentioned, there are very few additional restrictions. Especially, no restrictions are given concerning the parameterizations to be applied, where new developments are to be expected. Only some general physical preconditions, like continuity of the fluxes, are given. This is similar for the numerical grid and the solution technique to be used, since in both fields continuous new developments take place.

23.5.3.3 Test Cases

The test cases are specified in detail in the appendix of EGa. Five test cases use idealized domain setups (e.g., bell-shaped or Gaussian orography), and one of these compares with an analytic solution. To ensure some comparability of model results and this analytic solution, the domain setup and the model initialization are prescribed as similar as possible to the assumptions needed to derive the analytic solution. Nonetheless, the full model is used for this test case and no adjustments are to be made to the model code, since this approach turned out to be quite problematic in model evaluations performed in the past (Thunis et al. 2003).

All idealized test cases compare the model results with results of the very same model. Different model qualities are checked. These include determining the model's ability for reproducing the two-dimensionality of a solution, the correct consideration of orography, the correct simulation of flow fields, including katabatic winds in dependence of wind speed or model setup (grid size, shading). All these idealized cases provide many reference data. Therefore, a hit rate (Eq. 23.1) can reliably be used as MQI. Values suggested for A and D are given in Table 23.1; they agree with those suggested by Schlünzen et al. (2016). Hit rate H has to achieve 95% to check for sufficient domain height or for testing model result homogeneity.

Table 23.1 Absolute, A, and relative, D, deviation for surface measurements to calculate hit rate H (Eq. 23.1) as given by Schlünzen et al. (2016) and VDI (2017a)

Meteorological parameter	A	D
Component of wind vector	0.35 m/s	10%
Wind speed ff	0.50 m/s	10%
Wind direction (for ff > 1 m/s)	10°	Not applicable
Temperature	0.5 K	0.2%
Specific humidity	0.2 g/kg	2%

Three test cases prescribe realistic complex terrain and corresponding initial data taken from field measurements. The reference data of all three test cases are based on field experiments but have been specifically prepared for the model evaluation and are provided with the standard. The number of comparison data is small; therefore, MQOs include the uncertainty of the reference data (VDI 2017a).

For each test case, domain and grid size as well as the topography (orography plus land-use) are given, and the data sets are either given in the standard or available for download. Initial data, model starting time, and integration period, as well as the boundary conditions, are prescribed. The variables are summarized with their position and timing that are both essential for the model evaluation. The data are either directly given in the standard or provided via a web page where they can be downloaded. The test cases were all tested by the group of experts, applying up to six different models to test the test cases and the predetermined MQOs. Only if all test cases are passed, the evaluated model fulfills the developer-specific part of the EGa.

23.5.4 Some Detailed Evaluation Steps to be Performed by the Model User

Detailed information is given on how to select the model domain, e.g., vertical and horizontal extension, or the relation of the model domain to the focus area. For determining the independence of model results in the focus area on the domain size and grid, MQOs are given. It is also outlined where grid stretching is allowed and what needs to be considered to avoid an artificial flow acceleration within the domain.

Another important point for model users concerns the use of field data for model initialization. Here, another standard (VDI 2017c) is referenced. It is outlined in EGa how to interpolate model results at single grid points to observational field sites. Since applications in environmental meteorology often employ stationary solutions, a section of EGa is dedicated to the methods that have to be applied to determine (quasi)stationarity solutions.

As mentioned before, even an evaluated model is not perfect. Therefore, it is recommended in EGa to use the internal checks available in the model. These shall determine if wind speed or other meteorological variables remain within plausible ranges (values suggested in Schlünzen 1997). Furthermore, all results have to be checked for plausibility, at least in the focus area. If possible, quantitative comparisons have to be performed. The uncertainty of the reference data used for evaluation should be taken into account as outlined before. Last not least, the model simulations shall be documented to ensure third parties can reproduce the approaches taken.

One point in the user's part of the standard concerns test cases of EGa that have to be conducted by the model user. For the model users, three test cases are selected. This approach was taken to ensure that a model user has at least some basic skill to apply the model. The successful application includes the evaluation of the model results and may have to be provided on request to the environmental agency that uses the model results within an approval procedure.

23.6 Conclusions

Atmospheric models used in environmental meteorology are mostly computer programs, solving the conservation equations of mass, energy, and momentum with numerical methods in dependence on initial and boundary values. Evaluation of such models has a long history and is most relevant since costly investments depend on conclusions based on model results.

The generic structure of evaluation guidelines as outlined in Sect. 23.3 with its three parts, (A) specification of application area, (B) evaluation steps to be taken by the model developer, and (C) evaluation steps to be taken by the model user, can be applied to all fields of science. Specification for a specific application range helps not only one model user group but the whole scientific community, since it allows the detection of knowledge gaps in that field of science. If different models are not able to simulate a situation that is covered by the application range of an evaluation guideline, there are still three possible reasons: (1) the setup of the test case is wrong, or outside the application range, (2) the comparison data are not suitable, and (3) there is a gap in scientific knowledge, so that relevant processes are still unknown. This very last reason will lead to scientific progress in that science field. This progress can be systematically triggered, by attempting to falsify a model instead of choosing only those situations where a model has shown good performance in the past.

Specification of the generic structure of the evaluation guideline needs suitable test cases. These should be covering the application area. For this, it also needs to be determined, if forecasts, assessments, and statistics of model results or scenarios shall be the application focus that is to be assessed. Once this is determined, there is a need for reliable evaluation data. These have to include information on the uncertainties of the data used as reference data. The selection and preparation of a data set, including the test of the test case, easily takes several months and is scientifically challenging due to the uncertainties in the reference data. If the uncertainties are too large, the data set might not be sufficient to discriminate well-performing models from those which are insufficient. If the MQOs of a test case are only fulfilled by one of many models, then this model might have accidentally passed the test and this might not be repeatable even with the same model. Altogether, it is a scientific challenge to prepare test cases in such a way that a wider community can easily and successfully employ them.

When given the MQIs for a test case, model quality indicators assuming normal distributions should be avoided, since the differences of model results and reference data are mostly non-Gaussian distributed. This might be better considered by using hit rates. In any case, the uncertainty of the reference data needs to be included in the MQOs.

The further development from a guideline to a standard includes many steps, as discussed in Sect. 23.3. This time-consuming process involves the relevant stakeholders and a reviewing process. This is needed, if the application of a standard shall become more probable. However, it still does not mean a standard is applied. For this, it has to become part of legal regulation or a directive (e.g., EC 1980). As an example of an environmental regulation, Appendix 3 of the German regulation TA Luft (2002) is taken. The requirement to use a Lagrangian particle model for dispersion calculations that has to be consistent with VDI 3945 Part 3 (VDI 2000) is specified there. This ensures that environmental assessments are based on the same method at least for the dispersion simulation. Further steps are prepared concerning the meteorology fields (VDI 2017a, b), model use (VDI 2015), and assessment setup (VDI 2010) and might be taken up in a future update (BMUB 2016). However, how the formulations will be and if the standards developed will ever be part of a legally binding regulation is outside the science community but part of political negotiations and decisions.

Acknowledgements The guideline development in CCA-EM includes many experts in the field of environmental meteorology that support the development of standards since more than 60 years. We have to thank all those involved in this time-consuming and voluntary work, since without their contributions to the development of environmental meteorology standards the atmosphere would be less healthy and environmental friendly.

The research needed for this contribution is supported through the Cluster of Excellence "CliSAP" (EXC177) funded by the German Science Foundation, the research project UrbMod funded by the state of Hamburg, Germany, and last not least the German Environment Agency UBA via UFOPLAN project 3712 43 241.

The content of this paper is in the responsibility of the author.

References

Baklanov, A., Schlünzen, K. H., Suppan, P., Baldasano, J., Brunner, D., Aksoyoglu, S., et al. (2014). Online coupled regional meteorology chemistry models in Europe. Current status and prospects. *Atmospheric Chemistry and Physics, 14*, 317–398, https://doi.org/10.5194/acp-14-317-2014.

BMUB. (2016). Entwurf zur Anpassung der Ersten Allgemeinen Verwaltungsvorschrift zum Bundes–Immissionsschutzgesetz (Technische Anleitung zur Reinhaltung der Luft – TA Luft) Stand: 09.09.2016. Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit. Retrieved June 21, 2017, from www.bmub.bund.de/N53642/.

Cox, R., Bauer, B. L., & Smith, T. (1998). Mesoscale model intercomparison. *Bulletin of the American Meteorological Society, 87,* 167–196.

Deardorff, J. W. (1978). Efficient prediction of ground surface temperature and moisture, with inclusion of a layer of vegetation. *Journal Geophysical Research, 83,* 1889–1903.

Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., et al. (2010). A framework for evaluating regional-scale numerical photochemical modeling systems. *Environmental Fluid Mechanics*, *10*, 471–489.

Di Sabatino, S., Olesen, H. R., Berkowicz, R., Franke, J., Schatzmann, M., Leitl, B., et al. (2011a). Towards a model evaluation protocol for urban scale flow and dispersion models. *International Journal of Environment and Pollution, 47*, 326–336.

Di Sabatino, S., Buccolieri, R., Olesen, H. R., Ketzel, M., Berkowicz, R., Franke, J., et al. (2011b). COST 732 in practice: The MUST model evaluation exercise. *International Journal of Environment and Pollution, 44,* 403–418. https://doi.org/10.1504/ijep.2011.038442.

EC. (1980). Council Directive 80/779/EEC of 15 July 1980 on air quality limit values and guide values for sulphur dioxide and suspended particulates. *Official Journal of the European Communities*, *L 229*, 30–48.

EC. (2008). Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe (OJL 152, 11.6.2008, pp. 1–44). Retrieved November 09, 2011, from http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:152:0001:0044:EN:PDF.

Franke, J., Hellsten, A., Schlünzen, K. H., & Carissimo, B. (2011). The COST 732 best practice guideline for CFD simulation of flows in the urban environment - a summary. *International Journal of Environment and Pollution, 44,* 419–427. https://doi.org/10.1504/IJEP.2011.038443.

Geertsema, G., Schlünzen, K. H., ter Pelkwijk, H., Jalkanen, L., Baklanov, A., Fisher, B., et al. (2018). User training for mesoscale modelling applications to air pollution. In R. S. Sokhi, A. Baklanov, & K. H. Schlünzen (Eds.), Mesoscale modelling for meteorological and air pollution applications. Anthem Press, London. ISBN:9781783088263.

GLA. (2002). 50 years on. The struggle for air quality in London since the great smog of December 1952. Mayor of London, Greater London Authority.

Hay, J. S., & Pasquill, F. (1957). Diffusion from a fixed source at a height of a few hundred feet in the atmosphere. *Journal of Fluid Mechanics*, *2*, 299. https://doi.org/10.1127/0941-2948/2012/0356.

Letzel, M. O., Helmke, C., Ng, E., An, X., Lai, A., & Raasch, S. (2012). LES case study on pedestrian level ventilation in two neighbourhoods in Hong Kong. *Meteorologische Zeitschrift, 21,* 575–589.

Luft, T. A. (2002). Technical Instruction on Air Quality Control – Erste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz, June 24, 2002. GMBl. Nr. 25-29, S. 511.

Meroney, R., Ohba, R., Leitl, B., Kondo, H., Grawe, D. (2016). Review of CFD guidelines for dispersion modeling. *Fluids*, 1–14, https://doi.org/10.3390/fluids1020014.

Nordmann, S., Quass, U., Schlünzen, K. H., Müller, W. J., & Jäckel, S. (2017). CEN/EU Richtlinienaktivitäten zur Qualitätssicherung von Ausbreitungsrechnungen und Verursacheranalysen. *Gefahrstoffe-Reinhaltung der Luft, 7*(8), 303–308.

Olesen, H. R. (2017). Initiative on "Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes". Retrieved June 25, 2017, from http://www.harmo.org/.

Popper K. R. (1982). Logik der Forschung. Verlag J.C.B. Mohr (Paul Siebeck), Tübingen (pp. 450).

Schlünzen, K. H. (1996). Validierung hochauflösender Regionalmodelle. *Ber. aus dem Zentrum f. Meeres- und Klimaforschung*, Meteorologisches Institut, Universität Hamburg, A23, 184. http://www.bis.zmaw.de/fileadmin/Bib/Volltexte/ZMK-A23.pdf.

Schlünzen, K. H. (1997). On the validation of high-resolution atmospheric mesoscale models. *Journal of Wind Engineering and Industrial Aerodynamics*, *67 & 68*, 479–492.

Schlünzen, K. H., & Katzfey, J. J. (2003). Relevance of sub-grid-scale land-use effects for mesoscale models. *Tellus, 55A,* 232–246.

Schlünzen, K. H., & Sokhi, R. S. (Eds.) (2008). Overview of tools and methods for meteorological and air pollution mesoscale model evaluation and user training. Joint report of COST Action 728 and GURME. GAW Report No. 181 (115 pp).

Schlünzen, K. H., Grawe, D., Bohnenstengel, S. I., Schlüter, I., & Koppmann, R. (2011). Joint modelling of obstacle induced and mesoscale changes – current limits and challenges. *Journal of Wind Engineering and Industrial Aerodynamics, 99,* 217–225. https://doi.org/10.1016/j.jweia.2011.01.009.

Schlünzen, K. H., Conrady, K., & Purr, C. (2016). Typical performances of mesoscale meteorology models. In: D. G. Steyn & N. Chaumerliac (Eds.) *Air pollution modeling and its application XXIV, air pollution modeling and its application XXIV* (p 447–457). https://doi.org/10.1007/978-3-319-24478-5_72.

Schlünzen, K. H., Grawe, D., & Oettl, D. (2017). Qualitätssicherung in der Ausbreitungsrechnung - Evaluierungsrichtlinien für mesoskalige und mikroskalige Windfeldmodelle. *Gefahrstoffe-Reinhaltung der Luft, 7*(8), 298–302.

Schlünzen, K. H., Builtjes, P., Deserti, M., Douros, J., Galmarini, S., Miranda, A. I., Palau, J. L., & Schere, K. (2018). Evaluating the performance of mesoscale meteorology models used for air quality simulations. In: R. S. Sokhi, A. Baklanov, & K. H. Schlünzen (Eds.), Mesoscale modelling for meteorological and air pollution applications. Anthem Press, London. ISBN:9781783088263.

Thunis, R., Galmarini, S., Martilli, A., Clappier, A., Andronopoulos, S., Bartzis, J., et al. (2003). MESOCOM: An inter-comparison exercise of mesoscale flow models applied to an ideal case simulation. *Atmospheric Environment, 37*, 363–382. https://doi.org/10.1016/s1352-2310(02)00888-9.

VDI. (2000). VDI 3945 Part 3 Environmental meteorology - Atmospheric dispersion models - Particle model. Berlin, Beuth-Verlag. Retrieved July 27, 2017, from https://www.beuth.de/en/technical-rule/vdi-3945-blatt-3/36552631.

VDI. (2010). VDI 3783 Part 13: Environmental meteorology - Quality control concerning air quality forecast - Plant-related pollution control - Dispersion calculation according to TA Luft. Beuth-Verlag. Retrieved July 27, 2017, from https://www.beuth.de/en/technical-rule/vdi-3783-blatt-13/121969008.

VDI. (2015). VDI 3783 Part 16: Environmental meteorology - Prognostic mesoscale wind field models - Methods for licensing procedures according to TA Luft. Beuth-Verlag. Retrieved July 27, 2017, from https://www.beuth.de/en/technical-rule/vdi-3783-blatt-16/228625611.

VDI. (2017a). VDI 3783 Part 7: Environmental meteorology - Prognostic mesoscale wind field models – Evaluation for dynamically and thermodynamically induced flow fields. Beuth-Verlag. Retrieved July 27, 2017, from https://www.beuth.de/en/technical-rule/vdi-3783-blatt-7/267500583.

VDI. (2017b). VDI 3783 Part 9: Environmental meteorology - Prognostic microscale wind field models - Evaluation for flow around buildings and obstacles. Beuth-Verlag. Retrieved July 27, 2017, from https://www.beuth.de/en/technical-rule/vdi-3783-blatt-9/267500591.

VDI. (2017c). VDI 3783 Part 20: Environmental meteorology – Testing the transferability of meteorological data for application within the context of TA Luft. Berlin: Beuth Verlag. Retrieved January 16, 2019, from https://www.beuth.de/en/technical-rule/vdi-3783-blatt-20/261571898.

VDI. (2018). VDI 3783 Part 15.1: Environmental meteorology – Simplified method for estimating nitrogen deposition. In preparation, agreed draft published 2018. (https://www.beuth.de/en/draft-technical-rule/vdi-3783-blatt-15-1/288390435). publication of standard in 2019 (personal communication).

Chapter 24
The Management of Simulation Validation

Fei Liu and Ming Yang

Abstract In this chapter, we discuss the management of simulation validation for complex simulation systems. We first present nine principles for simulation validation, which are important to achieve good management and determine the success of simulation validation. By considering these principles, we present a management framework of simulation verification and validation (V&V), which includes four components: V&V process, V&V scheme, V&V metrics, and V&V tools. That is, we adopt a process-oriented, optimized, quantitative, and automatic management manner for simulation V&V of complex simulation systems. We then describe each component of the framework in detail and discuss the involved management issues. We hope this chapter could help the user to understand the management of simulation validation for complex simulation systems, and guide the user to manage the validation of practical simulation systems.

Keywords Simulation validation · Management · Verification and validation

24.1 Introduction

Modeling and simulation (M&S) has been widely and increasingly applied in many fields such as military, aerospace, manufacturing, transportation, economic, and biological areas (Hill et al. 2001; Robinson 2001; Mostafa et al. 2018; Liu et al. 2014), and played an essential role in some of these areas. By means of M&S techniques, a real (or even imaginary) system can be modeled as a simulation model (or simula-

F. Liu (✉)
School of Software Engineering, South China University of Technology,
Guangzhou 510006, China
e-mail: feiliu@scut.edu.cn

M. Yang
Control and Simulation Center, Harbin Institute of Technology,
Postbox 3006, Harbin 150080, China
e-mail: myang@hit.edu.cn

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_24

tion system). In order to build the confidence of a simulation system to be studied, verification and validation (V&V) has to be conducted, which can reveal whether the simulation system has gained sufficient credibility (Sargent 1991, 2013; Balci 2003). Among these two terms, verification confirms whether we correctly build a model with respect to its specification. Simulation verification is similar to software verification and thus can make use of plenty of software engineering techniques (IEEE 2012). In contrast, validation answers the question: do we build a right simulation model? That is, does a model represent the real system that it is supposed to model? A model is usually built for a specific purpose and validation assures that the specific purpose is achieved (Sargent 1991).

For simple simulation models, validation is easy. For example, a three-step approach (Naylor and Finger 1967) has been widely used for validating simple models: (1) build a model and make it have high face validity, (2) validate the model assumptions such as structural and data assumptions, and (3) compare the model outputs with real outputs to obtain the consistency of these two groups of data. If the consistency satisfies a predefined threshold for a specific purpose, we can say the model is valid enough. This means the management of validation for simple models is not essential.

However, currently simulation systems become more and more complicated (Fujimoto 2003). These systems exhibit at least the following characteristics:

- A simulation system is usually composed of different types of computer generated entities and physical devices, and the involved models may come from different areas, e.g., plane models, command and control models, electric models and different types of environment models.
- The models of a simulation system could be distributed among different computers, which forms a distributed simulation environment. In this case, time synchronization introduces more challenging issues for testing and validation.
- A simulation system has different levels of hierarchy, and thus can be broken into subsystems, submodels, and components. Moreover, many components could be highly coupled.
- The development of such a simulation system could take years and thus the management work is hard.

For such complex simulation systems, both the development and validation work is challenging (Shi et al. 2008; Liu et al. 2008). For such systems, it is not enough to depend on the validation methods used for simple models. Compared with simple models, validation of complex simulation systems is not only a technical issue any more, but involves many management issues (Liu et al. 2008). In some sense, validation is more like a management job.

- For a complex simulation system, we have to break the whole system to be studied into different levels, such as subsystems, models, and components. After that, we validate each component and the coupling of two components and finally accomplish the validation of the whole system in a bottom to top way. During this process, a variety of validation techniques could be adopted. For such a complicated task, we have to adopt good manners and a set of tools to manage it:

- We have to collect experts from different areas to accomplish the validation of models from different areas. This requires a good organization of all validation tasks, the validation experts and other resources such as documents.
- We need a good validation plan that facilitates the management of validation tasks during the long development time of complex simulation systems.

For the validation of complex simulation systems, many research topics have been proposed, such as constructing reasonable validation processes, considering appropriate validation metrics, and developing powerful tools. In order to achieve an effective validation of complex simulation systems, we have to carefully manage the simulation validation and make them work together well.

In this chapter, we will concentrate on the management of simulation validation for complex simulation systems. However, most of the time we discuss validation in conjunction with verification (V&V), as they are closely intertwined. We will describe our work on how we have achieved an effective validation during the past few years (Liu and Yang 2009; Shi et al. 2009a, b, 2008; Liu et al. 2006c, b; Liu and Yang 2005b; Liu et al. 2005; Liu and Yang 2005a; Liu et al. 2008).

The structure of the chapter is as follows. We first summarize the main principles of simulation validation. After that, we present our management framework of simulation V&V, followed by the description of each component of this framework. Finally, we provide the conclusions.

24.2 Simulation Terminology

This chapter defines some important terms, which will be used throughout the paper.

Modeling and simulation. Modeling is a process to build a mathematical or computational model for a specific real or imaginary system, while simulation is a process to dynamically run a computational model in a computer. Usually, these two terms are widely accepted in the simulation community.

Models. We have many types of models in the simulation community, and different people may give different explanations. In this chapter, we simply differentiate some of these models as follows. **Mathematical models** are those models that are described by mathematical formula, such as differential equations. **Computational models** are built with such techniques as finite state machine, Petri nets or DEVS (Hollmann et al. 2015). **Simulation models** are implementations of mathematical or computational models via coding.

A **conceptual model (CM)** is an intermediate step (or model) which links modeling requirements with simulation design. A conceptual model (Balci et al. 2008; Robinson 2017, 2013, 2012) describes what a simulation model looks like before it is implemented.

Besides, a **simulation system** (SS) is also used to denote a simulation model (SM) or a group of coupled simulation models. For a real system composed of many components, we can construct a simulation model for each component and those components can be combined into a single unified simulation system.

24.3 Principles of Simulation Validation

Setting up and understanding a set of principles is the primary issue of simulation validation. The principles can help people to manage complicated validation work, which may determine the success of simulation validation if they are well and appropriately used in a simulation development project. Based on the analysis of the literature (Balci 1995; Chew and Sullivan 2000; Robinson 1997; Balci et al. 2000; Balci 2003) and our practical experience, the following principles are most relevant to simulation validation:

(1) Validation of a model is conducted with respect to its application purpose.

Any model has an explicit and specific application purpose; validation of the model has to be done with respect to its application purpose. A model could be valid for one application purpose and invalid for another application purpose. Therefore, an accurate specification of the application purpose of a model is the most important issue for the validation work.

For a complex simulation system, we do not directly validate it against its application purpose. Instead, based on the application purpose, we develop validation requirements or criteria and make validation plans. After that, we validate the system against the defined validation requirements, which indirectly reflect the application purpose.

(2) Validation must be conducted throughout the whole life cycle of a simulation system.

Like verification, validation is not an activity of a phase, but a series of continuous activities in the whole life cycle of a simulation system. Validation starts by defining the validation requirements according to the application purpose of the simulation system. After that, we draw up a validation plan, which is continuously refined during several subsequent phases from simulation requirement analysis to system test. When a conceptual model is under development, the validation plan is used to perform the validation of the conceptual model. When the simulation development goes to the implementation phase, the validation of models or subsystems is performed. After the whole system is finished, the result validation is performed. Even when the simulation is executed, the result validation is always accompanied with it.

Besides, please note that the validation could be iterative due to the following reasons. For example, once simulation requirements change, we have to run validation for those affected parts. There may also be some severe faults during conceptual modeling or design phases, which cause the change of models. In this case, we also need to validate the relevant components again. Due to the multi-job, iterative characteristics of the validation work, we can see that there are many management issues involved in the simulation validation.

(3) Validation is a product/process/project-centered assessment.

For complex simulation systems, validation is not only product-centered assessment any more, but becomes product/process/project-centered assessment (Balci 1995).

Here product means the intermediate or final result of a simulation system, e.g., requirement analysis results, conceptual model, or the whole simulation system. Process means how to perform the development with appropriate methodologies and techniques. Project means those management characteristics such as personnel, resources, documents, planning, and control.

For complex simulation systems, we not only need to assure the correctness of simulation products, but also have to consider the correctness and appropriateness of both process and project. For example, when a conceptual model is validated, we at least consider the following aspects:

- Conceptual model validation. Check whether the conceptual model product is right;
- Conceptual model construction process assessment. Check whether the construction process of a conceptual model is correct and appropriate;
- Conceptual model project assessment. Check whether appropriate personnel and resources are used, and whether the documents are well written etc.

(4) The validation result should not be considered as a binary variable: Yes or No, but should be a quantitative metric of the validity.

A model is an abstraction of a real system, so we cannot expect to establish an absolutely valid model; even if we could, few people would like to pay the price (Shannon 1975). Therefore, a model should not be considered as passing the validation simply by answering yes or no for its validity, but should be done in a quantitative way, e.g., using a scale from 0 (absolutely incorrect) too 100 (absolutely correct) to describe the degree of validity.

In order to obtain a quantitative metric of the validity, we need to quantify all the validation results of a simulation system, which means we may have to develop a validation metrics for the simulation system. Therefore, we think constructing and managing a reasonable validation metric should be an important matter of validation management of complex simulation systems (*see also* Chap. 13 *by Marks* and Chap. 18 *by Saam in this volume*).

(5) Simulation validation should be independent from developers.

When a model gets finished, the developers usually validate it first to make sure they built the right model. However, this is usually not enough as they are often biased. For a large simulation system, the independent validation becomes much more essential. A good practice for this is to set up an independent validation team to perform all the validation work. Therefore, the formation of a verification and validation group is usually necessary for the development of complex simulation nowadays.

(6) Complete validation is impossible.

The validation of a model is constrained by time and budget, so a complete validation should not be expected. Therefore, we have to think of how to spend the limited time and budget on essential validation activities in order to achieve a better validation result.

For complex simulation systems, we have to carefully draw up validation plans by choosing those essential activities hopefully with the help of some optimization techniques. In the following, we will demonstrate how to do this.

(7) Simulation validation must be well planned and documented.

As described above, validation is a continuous activity throughout the entire life cycle of a simulation system. During this process, there are many specific validation activities and different types of documents, so we have to carefully make a plan on these activities and manage all the documents. For this, a computer aid tool is usually expected.

(8) The validity of each sub-model does not guarantee the validity of the whole model.

A simulation system is usually composed of a number of models. Even if we validate each model and they are all valid with respect to the application purpose, this does not imply that the whole model can be acceptable.

In fact, for a complex simulation system, the validation at the system level is a most important job, which can be seen from the validation process model that will be discussed in the following sections.

(9) Validation and error detection should be done as early as possible.

During the life cycle of a simulation system, correcting errors detected in later phases is much more expensive. Therefore, we need to try to validate a model and detect its possible errors as early as possible (*see* Chap. 5 *by Roy in this volume*).

When we make a validation plan, we need to balance the time and cost for each phase and should avoid to set the majority of the focus on result validation. Instead, we should allocate more time and resources on the validation of the conceptual model.

24.4 Management of Simulation V&V: A Framework

As described above, simulation V&V is a complicated task for complex distributed simulation systems, involved by many organizations. In order to achieve a successful V&V, a good management is essential. In the last decades, we have thoroughly researched this matter.

Before proceeding, we want to first clarify what is simulation V&V management. There are different definitions for management depending on different areas and people. Concerning our scenario, the definition of Fayol (1917) provides an accurate description. Namely, “to manage is to forecast and to plan, to organize, to command, to coordinate and to control”.

We adopt this definition to describe simulation V&V management, which operates through the aforementioned five basic functions: planning, organizing, coordinating, commanding, and controlling

- Planning. Generate V&V plans according to simulation purposes and V&V requirements. When doing this, the above-mentioned principles have to be considered.
- Organizing. Make sure that all the resources available for V&V are put into place. The usage of a well-defined V&V process model and a computer tool can improve organizing and the other three following functions mentioned below.
- Coordinating. Coordinate all the people to well accomplish their individual V&V tasks. This function is proposed from the point of view of different collaborating teams.
- Commanding. Ask and urge people to accomplish their respective tasks.
- Controlling. Check the V&V progress against the validation plan to make sure that the V&V plan is well performed.

For simple models, simulation V&V management usually does not matter. However, this is not the case for complex simulation systems. These years, we have presented a framework for guiding the management of simulation V&V (see Fig. 24.1 for an illustration), which can also be considered for some V&V management practices for complex simulation systems

- Developing a V&V process model to organize all the validation activities. This V&V process should consist of as many V&V activities as we can think of for a class of simulation systems. When a new simulation system is under study, we can tailor this V&V process model to obtain those V&V activities suitable for the current simulation system. In another word, simulation V&V can be considered as a process-oriented job. The V&V process model can help to perform the organizing

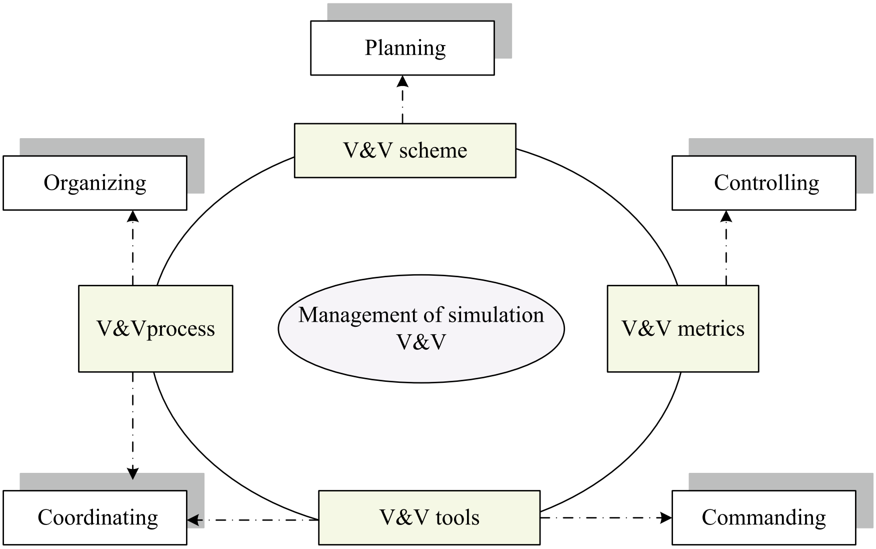


Fig. 24.1 Management of simulation V&V: a framework

and coordinating functions for a V&V management. In Sect. 24.5, we will discuss this issue in detail.

- Drawing up a V&V scheme (plan) to help to accomplish an effective V&V. Based on the V&V requirements and V&V process model given above, we can draw up a V&V scheme. This scheme will guide all the V&V work throughout the whole simulation development. Besides, in order to draw up a better scheme, optimization techniques have to be developed and used. This issue will be discussed in Sect. 24.6. This corresponds to the planning function of management.
- Quantifying simulation V&V results to give a measure of the overall credibility. We developed a V&V metric system to quantify the credibility of each component. By integrating the credibility of each component from bottom to top, we can obtain the whole metric of simulation V&V (Liu et al. 2006a). Thus, we can control the V&V results in a quantitative way.
- Developing software tools to manage all the documents, personnel and resources. This not only assures the traceability of all the V&V work, but also helps to find issues when something unexpected happens (Klock and Kemper 2010a); *see also* Chap. 25 *by Reinhardt et al. in this volume.* Using a set of tools, we can implement the organizing, coordinating and commanding functions of the simulation V&V management in an efficient way.

In summary, to achieve a successful V&V, we adopt a process-oriented, optimized, quantitative, and automatic (POQA) management approach. We have applied our management framework to several complex simulation systems, which show its effectiveness. In the following sections, we will in detail discuss each component of this framework and illustrate how it works.

24.5 Process-Oriented Simulation V&V Management

It has been shown that V&V is accompanied by the whole M&S life cycle (Jennifer and Cindy 2000). A well-established approach to managing V&V of complex simulation systems is to design a complete and feasible V&V process model for the whole M&S life cycle. That is, at each M&S phase, appropriate V&V activities have to be defined, executed, and managed. So far, several V&V processes have been developed for different simulation applications, such as those for DIS and HLA-based simulation systems (Jennifer and Cindy 2000; DMSO 1996).

In order to satisfy our requirements, we developed a more generic and detailed V&V process model (see, e.g., Liu et al. 2008), illustrated in Fig. 24.2. This process model includes the following main validation activities:

- Develop simulation validation requirements;
- Develop simulation validation scheme;
- Validate conceptual model;
- Validation subsystem models;
- Validate the whole simulation system.

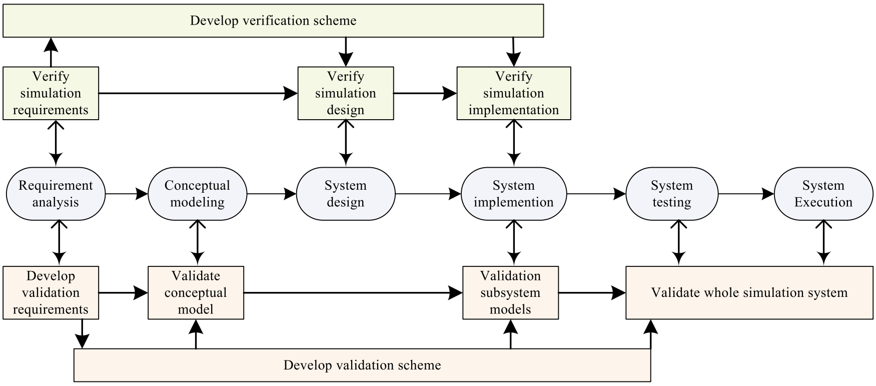

Develop verification scheme
Verify simulation requirements
Verify simulation design
Verify simulation implementation
Requirement analysis
Conceptual modeling
System design
System implemention
System testing
System Execution
Develop validation requirements
Validate conceptual model
Validation subsystem models
Validate whole simulation system
Develop validation scheme

Fig. 24.2 A simulation V&V process model

The last two items consist of the so-called result validation.

This process model also includes the following main verification activities:

- Develop simulation verification scheme;
- Verify simulation requirements;
- Verify simulation design;
- Verify simulation implementation.

24.5.1 Simulation Validation Steps

In this section, we will briefly describe these validation activities of the process model.

(1) Develop validation requirements.

The first step for a good simulation validation practice is to analyze and develop accurate and necessary validation requirements. In order to balance risk and cost, we should not try to validate all simulation requirements. Instead, we should focus on those that are essential in terms of the application purpose.

Here, we adopt the following principle to determine the validation requirements. We first consider the application purpose of a simulation system as a set of high-level validation requirements. We then break each high-level requirement into several small requirements that are operational. We repeat this step until we determine all the validation requirements. Here the software requirement analysis techniques can be applied for validation requirement analysis. Please note that we should also maintain the traceability between the validation requirements and simulation application purpose.

When we finish validation requirements, we need to further precisely specify acceptance criteria, in terms of which we check if validation passes or not (DMSO

1996). Validation activities and their intensity are always depending on those criteria which can differ in regard to e.g. usability criteria, suitability, or required precision of an M&S.

(2) Develop validation scheme.

When we finish the fist step, we can analyze the cost and risk of each validation requirement and then draw up the simulation validation scheme. The scheme usually consists of two main parts: conceptual model validation and result validation. In fact, throughout the whole simulation system development, the validation scheme is dynamically improved according to the change of simulation requirements. In Sect. 24.6, we will in detail discuss how to achieve a good validation scheme.

(3) Conceptual model validation.

As described above, a CM is the key to achieve a successful simulation development, and thus the validation of a CM is usually the most important job in a complex CS. The validation of a CM usually includes the following general validation activities:

- Validate the modeling hypothesis and assumptions. A model is a simplification of the corresponding real system, and thus some modeling hypothesis and assumptions have to be made for the simplification from the modeler's point of view. Therefore, we also need to verify whether these assumptions satisfy the user's application purpose.
- Validate the functions of the whole conceptual model. A real system may have more than one function, and we only need to concentrate on those functions that are relevant to the user's application purpose.
- Validate each entity including its parameters, behavior, internal interactions, inputs, outputs, etc. Each entity may correspond to a model, which may be developed by different institutions or people. A good specification and validation of each entity is the most difficult but important job.
- Validate the interaction among entities, including interface parameters, allowed precision and fidelity, which is also very important for current distributed simulation systems.

For a complex simulation system, its conceptual model could include many hierarchies, e.g., model, sub-model, module and sub-module (Robinson 2006). At each level, we need to perform sufficient validation activities, and thus the total amount of validation activities can be huge. This issue applies also to the other validation phases.

In regard to management, the following points need to be explicitly specified in the validation plan:

- what validation activities have to be conducted,
- what validation metrics should be taken for each validation activity,
- what techniques (see DMSO 1996) for a list of available techniques) should be employed for each validation activity,
- what experts should be invited and used for each validation activity,

- what kinds of documents should be produced, and
- how to communicate with developers about the errors or issues detected.

These issues described above also apply to subsystem validation or system validation.

(4) Result validation.

During result validation of a model, a key task is to select key outputs (*see also* Chap. 14 *by Currie in this volume*) or performance metrics of the model. The selection criteria have to consider at least two aspects. On the one hand, the output to be considered should reflect the confidence of the user in the model (*see also* Chap. 17 *by Saam in this volume)*. On the other hand, the practical experimental data is available for this output or the output can be judged directly by some experts (Sargent 2015, 2011, 2010; Sargent and Balci 2017); *see also* Chap. 15 *by Murray-Smith in this volume*.

The result validation of a system can be done in a similar way to individual model validation. The only difference is that system validation puts more emphasis on the coupling effects of different models or subsystems.

Besides, during model validation or system validation, another good practice is to use more than one method for validating a model output. Due to the use of several sources of data and several methods (see Sargent 2015, 2011 for a summary of these available validation methods), the subjectivity can be reduced and the validation result becomes more accurate (Liu et al. 2009). Using some semi-automated tool can be another useful method to improve result validation (Klock and Kemper 2010b).

24.5.2 Simulation Verification Steps

In this section, we will briefly describe the verification activities of the V&V process model (*see also* Chap. 11 *by Rider in this volume*).

(1) Develop verification scheme. Draw up a verification plan and then execute it to perform all verification activities. This is similar to the development of a validation plan.

(2) Simulation requirements verification. Verify simulation requirements to see whether they address all the functional and performance requirements of users and whether they are feasible, sufficient, and accurate.

(3) Simulation design verification. Verify simulation design to see if it faithfully reflects all the simulation requirements.

(4) Simulation implementation verification. Verify simulation implementation to see if the simulation codes correctly implement simulation design. Basically, verification is performed by comparing the product of the current phase with that of the previous phase.

In summary, the V&V process model can improve the V&V management in the following aspects:

- The process model can help us to determine necessary V&V activities at each simulation development phase, and thus facilitates the reasonable assignment of the available V&V resources.
- With the process model, we can also easily locate the V&V experts for each phase and find suitable ones for performing V&V tasks.
- With the process model, we can coordinate all the people to well accomplish their own V&V tasks.

We have developed a computer tool to manage and tailor the process model above. The details can be found in Fang et al. (2005). This tool is being kept updated and new functionalities are continuously being added to it.

24.6 Draw up an Optimized V&V Scheme

The starting point of the simulation validation management is to draw up a good validation scheme (or plan). As we know, any project is constrained by the financial budget. This budget issue is more challenging for simulation V&V as it has to compete with simulation development for a bigger share of the total money.

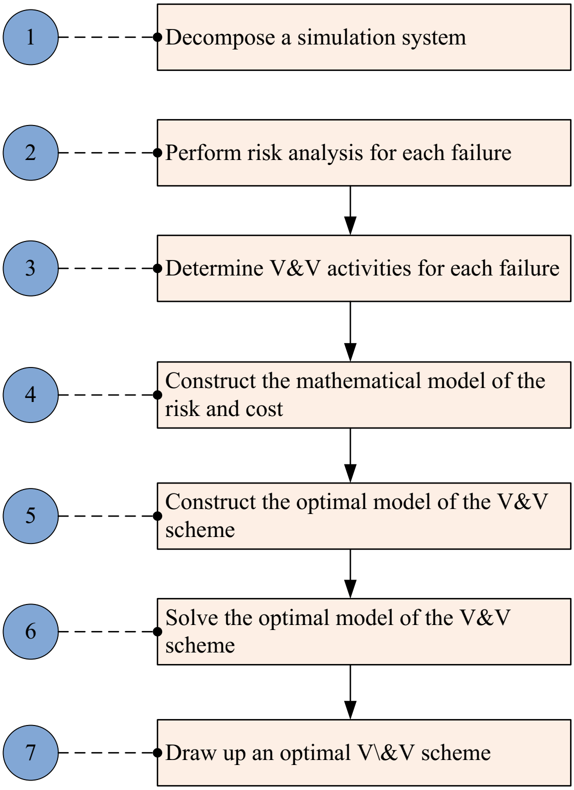
We may have to admit that more V&V activities may increase the credibility of the simulation system to be studied, which, however, may cost more money. So here is the question: how to spend the limited money for a better V&V result?

The traditional way is to ask experienced experts to select the most appropriate validation activities; however, although this applies to small simulation systems, it becomes hard to operate for large ones. This issue has been explored in Muessing and Laack (1997), but they only gave an informal risk assessment process for roughly selecting the needed V&V activities.

To address this issue, we presented a rigorous method for optimizing a V&V scheme of a complex simulation system in Liu and Yang (2009). The method is illustrated in Fig. 24.3, which involves the following steps:

1. Decompose a simulation system into subsystems, models, components, and key performance measures considering its hierarchies. Thus we obtain a tree-like structure of the simulation system. Throughout the entire V&V process, we may adopt the same decomposition method.
2. Perform risk analysis. We adopt fuzzy FMEA (failure mode and effects analysis) to assess the risk of each bottom node of the tree obtained by decomposing the simulation system. That is, for each bottom node, we determine possible failure modes. For each failure mode, we use linguistic variables to describe severity, occurrence and detection of the risk associated with this failure, and then employ a fuzzy rule base to yield the risk of this failure.

Fig. 24.3 A method for optimizing a V&V scheme of complex simulation systems, adapted from Liu and Yang (2009)


1
Decompose a simulation system
2
Perform risk analysis for each failure
3
Determine V&V activities for each failure
4
Construct the mathematical model of the risk and cost
5
Construct the optimal model of the V&V scheme
6
Solve the optimal model of the V&V scheme
7
Draw up an optimal V\&V scheme

3. Determine V&V activities and estimate the cost. We determine the necessary V&V activities for estimated risk and accepted maximum risk of each failure, and then estimate the cost for addressing this failure.
4. Construct the mathematical model for the relation between the risk and cost for each failure using the fuzzy linear regression analysis technique.
5. Construct the optimal model of the V&V scheme. In terms of the mathematical model of the relation between the risk and cost for each failure, we further construct an optimal model using the fuzzy linear programming technique.
6. Solve the optimal model with the maximum likelihood method.
7. Draw up an optimal V&V scheme. In terms of the solution of the optimal model, we can select the appropriate V&V activities and also assign reasonable cost to each V&V activity. Finally, an optimal scheme is obtained.

In Liu and Yang (2009), we also provided a detailed case study, which illustrates how to apply this method for real simulation systems. For the management of simulation V&V, we can employ the above-mentioned method to draw up a good validation plan.

From the point view of management, the approach above can be applied in the following way:

1. The total V&V task consists of three sub-tasks: conceptual model validation (*see also* Chap. 10 *by Gelfert in this volume*), subsystem validation, and system validation. So we first determine the budget for each sub-task. Besides, we also need to determine other available resources such as experts.
2. We then use the approach above for each sub-task. As a result, we will obtain a V&V plan for each sub-task.
3. After that, we perform V&V for each sub-task according to the V&V plan.

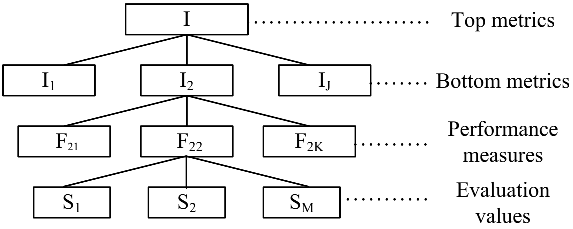
24.7 Quantify Simulation V&V Results

In the past, we usually adopted a qualitative method to evaluate whether a simulation system passed V&V. This traditional method is affected by the subjectivity of experts by evaluating a model in the binary way such as yes or no, or in the fuzzy way such as high, medium, or low. There are many drawbacks for this method as plenty of quantitative information is missing.

In order to improve the management of simulation V&V, we present an approach to quantifying V&V results. That is, we first break the whole simulation system into different levels, such as subsystem, model, and component, assuming that the component level is the bottom level. We then consider how many performance metrics are essential for the successful V&V of the considered component. As a result, we obtain a tree-like V&V metric system (Liu et al. 2006a), illustrated in Fig. 24.4, which at least consists of the following four levels:

1. Top metrics. At this level, we break a simulation system into subsystem and model. This level usually can be further divided into several sub-levels.
2. Bottom metrics, which correspond to the basic components of a model.
3. Performance measures. For each bottom component, we consider several essential performance measures, according to which we can judge whether this component is valid.

Fig. 24.4 Simulation V&V metrics



4. Evaluation values. That is, each performance measure gets some evaluation values either from experts or from different V&V methods like statistical methods (*see* Chap. 19 *by Robinson in this volume*).

When we obtain evaluation values for all the performance measures, we can compute the evaluation values for the top metrics in a bottom to top way. Thus, for the whole simulation system, we will have a quantitative metric of the credibility.

In our validation work, we require that all the models follow the above-mentioned quantitative V&V approach. When doing this, we usually adopt the following steps:

1. We develop specific V&V metrics for conceptual model validation, subsystem validation, and system validation, respectively. This has to be done by considering the validation plan.
2. Supervise the relevant people to execute their V&V work, and obtain the validation values.
3. Adopt an appropriate method such as the weighted summation or fuzzy methods to compute the evaluation values of top metrics (Liu et al. 2006a).
4. Analyze the evaluation results and send the issues to developer.

24.8 Computer Aided Management of Simulation V&V

Considering the fact that the validation of complex simulation systems usually takes a long time and involves many resources such as people and documents, we developed a set of V&V tools, which incorporates an integrated management environment, called HITVICE (Fang et al. 2005), and also several separated tools. Currently, these tools are still in progress.

24.8.1 Management Platform

This management platform, HITVICE, consists of several interlinked subsystems, each performing different management jobs. The general functionalities of the platform are shown in Fig. 24.5. In the following, we will briefly introduce each component:

1. Workflow manager. Workflow Manager is used to realize the management and tailoring of V&V activities for a V&V process model (e.g., the one given in Fig. 24.2). With the workflow manager, we can also coordinate and control different groups of people to accomplish their respective tasks.
2. Project manager. HITVICE offers project manager to manage multiple projects, which allows to create new V&V projects, monitor projects and control projects. This facilitates the reusability of existing work such as documents as well as existing experiences.

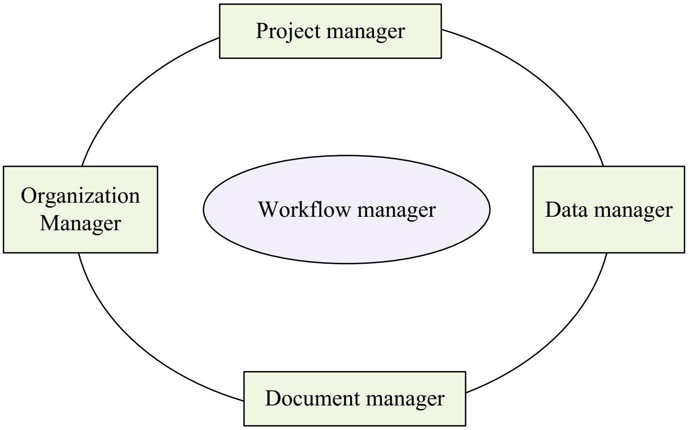


Fig. 24.5 The V&V management platform

3. Organization manager. With the organization manager, the user can appoint the V&V staff, and manage their V&V activities. Organization manager employs a role-team-staff structure to manage an organization.
4. Data manager. Data manager helps the user to collect and organize different types of data. The user can view, edit, and search data, and the import and export of data with other systems is also allowed.
5. Document manager. A V&V project usually has different types of documents, which need to be carefully dealt with. Document manager offers many functions for archiving and managing these documents, such as version control and document tracing.

24.8.2 Other Validation Tools

Besides the general management functions described above, we also developed several validation tools, which are briefly described as follows:

1. Expert systems-like validation tool (Liu et al. 2006b, 2009). The tool supports the validation of simulation systems with such methods as statistical methods, Turing testing and face validation. It offers an environment where we can step by step use these methods to accomplish the validation of a task.
2. V&V metrics tool (Qin et al. 2010). This tool implements the validation metrics e.g., described in Sect. 24.7, and also offers methods for computing the metrics at different levels. With this tool, the user can define a validation metric system, and automatically compute the values of different metrics if the evaluation values for the performance measures are given.

24.9 Discussion

(1) Validation for current complex simulation systems is more like a management issue than a technical issue.

For small models, validation is basically a technical issue, for which you usually choose appropriate techniques such as statistical methods or expert's judgement, and then you evaluate if the consistency between simulation outputs and real ones is satisfactory. However, during the validation of complex simulation systems, the management work makes up a large portion of the work overall. To reduce the validation cost, a best practice is to set up a set of standards, which should be complied with for any validation task (*see* Chap. 23 *by Schlünzen in this volume*). But it is also important that the sponsor of a simulation system recognizes the essence of validation and makes sufficient budget for validation.

(2) Model validation versus system validation.

Validation can be basically divided into two levels: system and model. Model validation is usually a technical issue, which has been discussed for more than three decades since the start of simulation research. System validation has been presented for complex simulation systems, which usually does not focus on the validity of each component but rather concentrates on the validity of the interoperability among components. A clear distinction of these two levels of validation can help to effectively manage validation and reasonably assign cost to different jobs.

(3) Sufficient validation versus limited budget.

This is an issue that is never resolved for many projects. To balance these conflicting requirements, one has to carefully estimate the whole workload necessary to achieve sufficient validation. Based on this estimate we try to convince the sponsor to offer more money for validation. Never be positive about this issue, as anyone thinks development is much more important than validation.

24.10 Conclusions

In this chapter, we discussed the management of simulation validation for complex simulation systems. We first presented nine principles for simulation validation; understanding and following these principles is important to achieve good management of simulation validation and determine the success of validation.

By considering these principles, we presented a management framework of simulation V&V, which includes four components: V&V process, V&V scheme, V&V metrics, and V&V tools. From this framework, we could see that we adopted a process-oriented, optimized, quantitative, and automatic management manner for simulation V&V. We then in detail described the four components in the framework and discussed the management issues. We hope this chapter may help the user to understand the management of simulation validation for complex simulation systems.

As the validation of complex simulation systems is a very complicated task, we have been researching this matter for a long time. We hope that the results described in this paper will advance the state of the simulation validation.

Acknowledgements This work has been supported by National Key R&D Program of China (2018YFC0830900), National Natural Science Foundation of China (61873094), Science and Technology Program of Guangzhou, China (201804010246), and Natural Science Foundation of Guangdong Province of China (2018A030313338).

References

IEEE. (2012). IEEE standard for system and software verification and validation. In *IEEE Std 1012-2012 (Revision of IEEE Std 1012-2004)*, pp. 1–223.

Balci, O. (1995). Principles and techniques of simulation validation, verification, and testing. *Winter Simulation Conference Proceedings*, *1995*, 147–154.

Balci, O. (2003). Verification, validation, and certification of modeling and simulation applications. In *Proceedings of the 2003 Winter Simulation Conference, 2003* (Vol. 1, pp. 150–158).

Balci, O., Arthur, J. D., & Nance, R. E. (2008). Accomplishing reuse with a simulation conceptual model. In *2008 Winter Simulation Conference* (pp. 959–965).

Balci, O., Ormby, W. F., Carr, J. T., & Saadi, S. D. (2000). Planning for verification, validation, and accreditation of modeling and simulation applications. In *2000 Winter Simulation Conference Proceedings (Cat. No.00CH37165)* (Vol. 1, pp. 829–839).

Chew, J. & Sullivan, C. (2000). Verification, validation, and accreditation in the life cycle of models and simulations. In *2000 Winter Simulation Conference Proceedings (Cat. No.00CH37165)* (Vol. 1, pp. 813–818).

DMSO (1996). Verification, validation and accreditation recommended practice guide.

Fang, K., Yang, M., & Wang, Z. (2005). The Hitvice VV&A environment. *Proceedings of the Winter Simulation Conference*, *2005*, 1220–1227.

Fayol, H. (1917). *Administration industrielle et generale; prevoyance, organisation, commandement, coordination, controle*. Pinat, Paris: H. Dunod et E.

Fujimoto, R. M. (2003). Distributed simulation systems. In *Proceedings of the 2003 Winter Simulation Conference, 2003* (Vol. 1, pp. 124–134).

Hill, R. R., Miller, J. O., & McIntyre, G. A. (2001). Applications of discrete event simulation modeling to military problems. In *Proceeding of the 2001 Winter Simulation Conference (Cat. No.01CH37304)* (Vol. 1, pp. 780–788).

Hollmann, D. A., Cristi, M., & Frydman, C. (2015). CML-DEVS: A specification language for devs conceptual models. *Simulation Modelling Practice and Theory*, *57*, 100–117.

Jennifer, C. & Cindy, S. (2000). Verification, validation, and accreditation in the life cycle of models and simulations. In *Proceedings of the Winter Simulation Conference* (pp. 813–818). Orlando, FL, USA: IEEE.

Klock, S. K. & Kemper, P. (2010a). An automated technique to support the verification and validation of simulation models. In *2010 IEEE/IFIP International Conference on Dependable Systems Networks (DSN)* (pp. 595–604).

Klock, S. K. & Kemper, P. (2010b). An automated technique to support the verification and validation of simulation models. In *2010 IEEE/IFIP International Conference on Dependable Systems Networks (DSN)* (pp. 595–604).

Li, Y., Liu, F., & Yang, M. (2009). Research of knowledge-based method to simulation model validation. *Journal of System Simulation*, *21*(8).

Liu, F., Blätke, M.-A., Heiner, M., & Yang, M. (2014). Modelling and simulating reactiondiffusion systems using coloured petri nets. *Computers in Biology and Medicine*, *53*, 297–308.

Liu, F., Ma, P., Yang, M., Sun, G., & Wang, Z. (2006a). Study on the credibility quantification of large complex smimulation systems. *Journal of Sichuan University (Engineering Science Edition)*, *38*(5), 169–174.

Liu, F., Ma, P., Yang, M., & Wang, Z. (2009). Key problems in validation of intelligent models. *Journal of Harbin Institute of Technology*, *16*(3), 371–375.

Liu, F., & Yang, M. (2005a). Validation of system models. In *IEEE International Conference Mechatronics and Automation, 2005* (Vol. 4, pp. 1721–1725).

Liu, F., & Yang, M. (2005b). *Verification and validation of artificial neural network models* (pp. 1041–1046). Heidelberg: Springer.

Liu, F., & Yang, M. (2009). An optimal design method for simulation verification, validation and accreditation schemes. *Simulation: Transactions of The Society for Modeling and Simulation International*, *85*(6), 375–386.

Liu, F., Yang, M., & Wang, Z. (2005). Study on simulation credibility metrics. In *Proceedings of the Winter Simulation Conference, 2005* (p. 7).

Liu, F., Yang, M., & Wang, Z. (2006b). Design and development of an expert system-like validation tool for distributed simulation systems. *Journal of Jiangsu University (Natural Science Edition)*, *27*(3).

Liu, F., Yang, M., & Wang, Z. (2006c). Formal verification method of simulation scenario based on high-level petri nets. *Control and Decision*, *21*(11).

Liu, F., Yang, M., & Wang, Z. (2008). VV&A solution for complex simulation systems. *International Journal of Simulation: Systems, Science and Technology*, *9*(1), 10–18.

Mostafa, H., Liu, F., & Heiner, M. (2018). Efficient modelling of yeast cell cycles based on multisite phosphorylation using coloured hybrid petri nets with marking-dependent arc weights. *Nonlinear Analysis: Hybrid Systems*, *27*, 191–212.

Muessing, P., & Laack, D. (1997). Optimizing the selection of VV&A activities: A risk/benefit approach. In *Proceedings of the 1997 Summer Computer Simulation Conference* (pp. 60–66).

Naylor, T. H., & Finger, J. M. (1967). Verification of computer simulation models. *Management Science*, *2*, 92–101.

Qin, L., Fang, K., & Yang, M. (2010). Research on the simulation credibility evaluation assistant tool based on hierarchical evaluation. *Computer Simulation*, *27*(6).

Robinson, S. (1997). Simulation maidel verification and validation: Increasing the users' confidence. *Winter Simulation Conference Proceedings*, 53–59.

Robinson, S. (2001). Modes of simulation practice in business and the military. In *Proceeding of the 2001 Winter Simulation Conference (Cat. No.01CH37304)*, (Vol. 1, pp. 805–811).

Robinson, S. (2006). Conceptual modeling for simulation: Issues and research requirements. In *Proceedings of the 2006 Winter Simulation Conference* (pp. 792–800).

Robinson, S. (2012). Tutorial: Choosing what to model; conceptual modeling for simulation. In *Proceedings of the 2012 Winter Simulation Conference (WSC)* (pp. 1–12).

Robinson, S. (2013). Conceptual modeling for simulation. In *2013 Winter Simulations Conference (WSC)* (pp. 377–388).

Robinson, S. (2017). A tutorial on simulation conceptual modeling. In *2017 Winter Simulation Conference (WSC)* (pp. 565–579).

Sargent, R. G. (1991). Simulation model verification and validation. In *1991 Winter Simulation Conference Proceedings* (pp. 37–47).

Sargent, R. G. (2010). Verification and validation of simulation models. In *Proceedings of the 2010 Winter Simulation Conference* (pp. 166–183).

Sargent, R. G. (2011). Verification and validation of simulation models. In *Proceedings of the 2011 Winter Simulation Conference (WSC)* (pp. 183–198).

Sargent, R. G. (2013). An introduction to verification and validation of simulation models. In *2013 Winter Simulations Conference (WSC)* (pp. 321–327).

Sargent, R. G. (2015). An introductory tutorial on verification and validation of simulation models. In *2015 Winter Simulation Conference (WSC)* (pp. 1729–1740).

Sargent, R. G., & Balci, O. (2017). History of verification and validation of simulation models. In *2017 Winter Simulation Conference (WSC)* (pp. 292–307).

Shannon, R. E. (1975). *Systems simulation: the art and science*. Englewood Cliffs, New Jersey: Prentice-Hall.

Shi, P., Liu, F., & Yang, M. (2008). Research on validation method for complex simulation systems. In *2008 Asia Simulation Conference—7th International Conference on System Simulation and Scientific Computing* (pp. 888–892).

Shi, P., Liu, F., & Yang, M. (2009a). Quantify simulation verification and validation. In *2009 11th International Conference on Computer Modelling and Simulation* (pp. 123–128).

Shi, P., Liu, F., Yang, M., & Wang, Z. (2009b). A fuzzy rules-based approach to analyzing human behavior models. In *2009 11th International Conference on Computer Modelling and Simulation* (pp. 346–351).

Chapter 25
Valid and Reproducible Simulation Studies—Making It Explicit

Oliver Reinhardt, Tom Warnke, Andreas Ruscheinski and Adelinde M. Uhrmacher

Abstract The validation of complex simulation models is a challenging task. To increase the trust into the model, diverse simulation experiments are executed to explore the behavior of the model and to check its plausibility. Thus, these simulation experiments present an important information about the validity of the model, similarly as the data used for calibration, as input for the model, and for testing its predictiveness. Simulation models are rarely developed from scratch but by reusing existing models, e.g., by extending or composing them, or for cross-validation. These models and their validity provide further details about the validity of a model. Thus, a multitude of artifacts contribute intricately related to the final simulation model and our "gut feelings" about it. To make these artifacts and their relations explicit and accessible, we will apply a declarative formal modeling language, a declarative language for specifying and executing diverse simulation experiments, and a provenance model to relate the diverse artifacts in telling the validation tale of an agent-based migration model.

Keywords Validation · Multilevel modeling · Demography · Provenance

25.1 Introduction

Validation is an important part of the modeling and simulation life cycle, as validation helps to decide whether a useful approximation of the system has been achieved, and directs a model's further refinement and enrichment, or as stated by Osman Balci:

O. Reinhardt (✉) · T. Warnke · A. Ruscheinski · A. Uhrmacher
University of Rostock, Institute of Computer Science,
Albert-Einstein-Straße 22, 18059 Rostock, Germany
e-mail: oliver.reinhardt@uni-rostock.de

T. Warnke
e-mail: tom.warnke@uni-rostock.de

A. Ruscheinski
e-mail: andreas.ruscheinski@uni-rostock.de

A. Uhrmacher
e-mail: adelinde.uhrmacher@uni-rostock.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_25

"Model *Validation* is substantiating that the model, within its domain of applicability, behaves with satisfactory accuracy consistent with the [Modeling and Simulation] objectives. Model validation deals with building the right model" (Balci 1997, p. 135). Clearly, validation is an approximative process, with which the trust into a model is successively increased and, correspondingly, different approaches exist. Zeigler defines three levels of validity (Zeigler et al. 2000): *replicative validity* or *historical validity*, i.e., the model reproduces data which has been observed from the real system (retrodiction), *predictive validity*, where a model produces data before it is observed from the real system, and *structural validity*, where the model reflects the structural relations of the real system. Troitzsch's discussion (Troitzsch 2004) about the question, "whether a theory which predicts empirical observations correctly at the same time explains what it predicts", deals with the difference between the second and third level of Zeigler's approach: explaining being interpreted as "showing how things work". Predictive validity refers to testing whether the model is able to reproduce data it has not seen, i.e., been trained with, before. Calibration contributes to replicative validity, by finding a parameterization of the model which can reproduce the observed behavior of the real system. As the target of the validation is not the concrete implementation of the simulation model, but the conceptual model, verification, the "testing process to establish whether a computer-based representation correctly describes the underlying mathematical, logical and theoretical structure of the model" (see Chap. 4 by Murray-Smith in this volume), is an important prerequisite of validation. Typically, the validity of a simulation model is tested in a *simulation experiment* (or to make the goal of validating the model more explicit, a *validation experiment*), where simulation runs are executed systematically, and the results are analyzed and, typically, compared to real-world data.

Figure 25.1 shows the layered structure of a typical validation experiment (Rybacki et al. 2012). The top layer is the validation experiment as a whole. It consists of multiple simulation configurations. In each of these, a point of the model's parameter space is selected, investigated, and then evaluated. In a stochastic model, for each configuration, a number of simulation runs (or replications) are executed. These sets of runs have to be analyzed, e.g., by calculation of mean values or confidence intervals. In a deterministic model, this layer is omitted, as it is not necessary to execute multiple runs for a single configuration. The bottom layer is formed by a single simulation run. There, apart from the model execution per se, the model state must be observed, and the observations of the single run may be analyzed, e.g., for steady-state detection. Considering the complexity of this experimental process, and the need to reproduce results of modeling and simulation more easily (Uhrmacher et al. 2016), modeling and simulation research will benefit by an unambiguous and sound specification of all these experimental steps (marked with (a) in the figure).

Domain-Specific Languages (DSLs) are programming languages which are not designed as general-purpose tools, but to solve more easily specific problems of a defined application domain (van Deursen et al. 2000). Consequently, domain-specific languages have been developed for specifying all steps of a simulation experiment (marked a) as well as for specifying the simulation model itself (marked b). Specification languages for specific parts of the simulation experiment, e.g., for querying

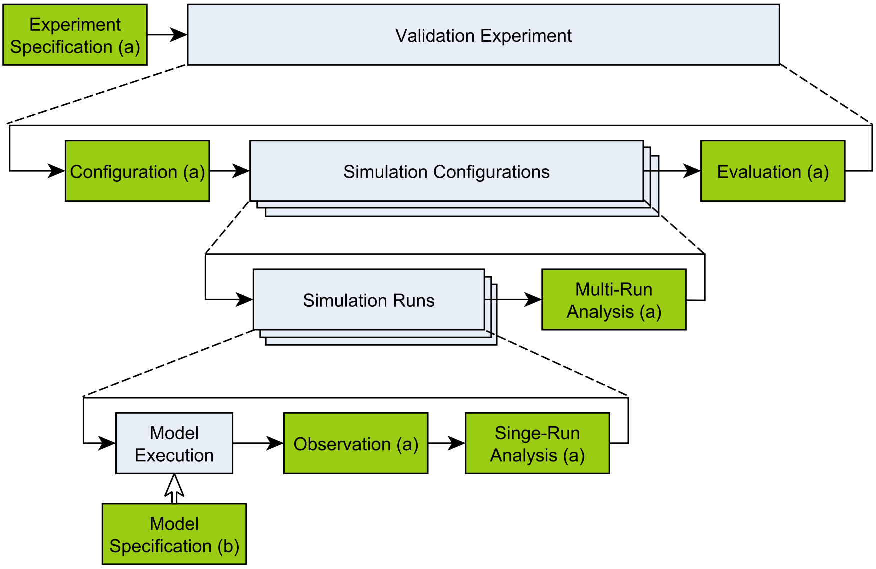


Fig. 25.1 Structure of a validation experiment. Specification of objects or processes in green. Examples for specification languages are given in Sects. 25.3 (for b) and 25.4 (for a), and (Schützel et al. 2014). See Rybacki et al. 2012

traces (Laurent et al. 2018) or for the whole simulation experiment (Waltemath et al. 2011; Ewald and Uhrmacher 2014), are a way to simplify the experimentation process and to support reproducibility. Domain-specific modeling languages (DSMLs) are DSLs for describing simulation models in a certain application domain, e.g., systems biology (Harris et al. 2016; Maus et al. 2011), digital systems (IEEE 2009), or demography (Warnke et al. 2017). DSMLs allow a more clear and compact implementation of the simulation model closer to the conceptual model, aiding verification as well as the reasoning about the model.

Going beyond a single validation experiment to a whole simulation study, model creation and validation must be well documented (see the nine principles of simulation validation proposed by Liu et al., Chap. 24 in this volume). Thereby, not only the validated model as a product but also the process of model creation and validation is of interest. Provenance provides "information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability, or trustworthiness" (Groth and Moreau 2013). Provenance models (Moreau et al. 2011) can be employed to explicitly describe a simulation model's history, e.g., the theories and data that contributed to it, and the experiments conducted with it, allowing to reason about the model's validity (Ruscheinski and Uhrmacher 2017).

Throughout the rest of this chapter, we will demonstrate all three kinds of explicit specification using an agent-based model of migration from Senegal to Europe (Klabunde et al. 2017) as an example, which is introduced in Sect. 25.2. In Sect. 25.3, we demonstrate how domain-specific modeling languages aid the modeler in implementing, documenting, and validating the model. In Sect. 25.4, we show the advantages of using an experiment specification language for describing and executing simulation experiments. Finally, in Sect. 25.5, how a provenance model can aid in ensuring the validity of models over a whole simulation study.

25.2 Example: A Model of the Decision to Migrate

The model presented by Klabunde et al. (2016, 2017) explores the hypothesis that in a critical phase approximately between the ages of 18 and 40, individuals make a series of important life decision, e.g., to get married or to have children, with which the decision to migrate competes. In simulations, it is tested whether based on this microlevel hypothesis, the observed age pattern of migrants can be explained. Thereby, the linked life courses of individuals are in the focus. This includes marriage, fertility, and mortality of individuals, which are governed by stochastic rates, as well as income and expenses. The migration decision process itself is modeled based on the Theory of Planned Behavior (Ajzen 1991). The assumption is, that the decision to migrate is made in multiple stages, through which every potential migrant goes (see Fig. 25.2): an intention is formed, plans and then preparations are made, and finally, the migration is attempted. Each agent has an intention to migrate, which, in accordance with the Theory of Planned Behavior, is derived from their attitude toward migration, their beliefs about social norms regarding migration, and their beliefs about behavioral control regarding migration. Those three factors are influenced by the agent's personal situation and his or her environment. A total of eight free weighting parameters determines the strength with which different aspects influence the migration intention. Finally, the migration intention governs how fast the agent proceeds through the stages of the decision process, as shown in Fig. 25.2.

The model was then applied to the case of migration from Senegal to Europe. To this end marriage, fertility, mortality, income, and expenses were estimated from data. For marriage a Coale–McNeill model (Coale and Mcneil 1972) was fitted, using data from the Demographic and Health Survey of Senegal (DHS) for individuals in Senegal, and from the MAFE survey (Migration between Africa and Europe) (Beauchemin 2015) for individuals who migrated. The individuals are then paired by employing a marriage market (Zinn 2012). Fertility was also estimated from DHS and MAFE data. For mortality, a Heligman–Pollard model (Heligman and Pollard 1980) was fitted to data from the UN World Population Prospects 2015. Income is taken from IMF data, consumption from World Development Indicators. An initial population was sampled from the 1988 Senegal census. Initial wealth was estimated from data by Davies et al. (2011). By adjusting the eight free parameters, the model was then calibrated to reproduce the distribution of the age at migration and the distribution of the time passed between starting to plan migration and the actual migration attempt observed in the MAFE survey. Furthermore, a sensitivity analysis

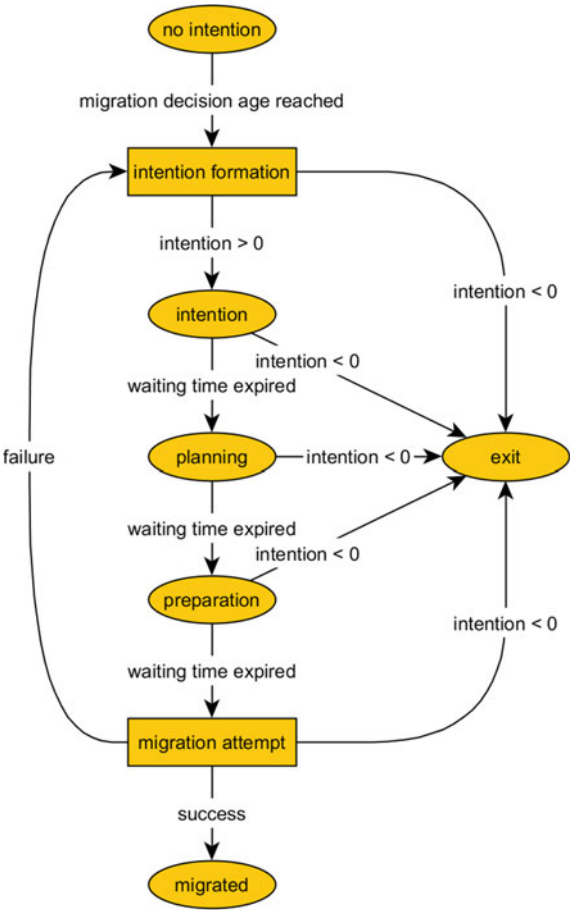


Fig. 25.2 The stages of the migration decision process. When people are born, they start in the "no intention" stage. When they reach a certain age, they enter the intention formation process. As long as the migration intention is positive, they advance through the stages, until they attempt migration. The waiting times until they advance are stochastic and depend on the strength of their intention. When their intention gets negative, they leave the decision process. See Warnke et al. (2017)

of the model was performed, to determine how changes of the free parameters affect the result of the calibration.

A preliminary analysis of 213 papers in the Journal of Artificial Societies and Social Simulation (JASSS) since 2011 by Troitzsch (2017) revealed that 19.2% of them compare quantitative simulation results to quantitative empirical data, while another 17.4% discuss the necessity of such comparison. This model clearly belongs to the current minority of papers published on agent-based social (or demographic) simulation as it relies on diverse data sets for calibration. In addition, it uses theories and other models. Although to test the assumed decision-making mechanisms at microlevel further efforts are required, e.g., controlled cognitive experiments (Conte et al. 2012), a lot went already into developing the model and substantiating the claims made. The rest of the paper will be on methods to make these efforts more easily accessible and, thus, assessable, using explicit declarative specifications of the model, its foundation in demographic theory and data, its history, and the validation experiments conducted with it.

25.3 Managing the Model: Domain-Specific Modeling Languages

To validate a simulation model, i.e., to ensure the model is the right model for its purpose, it is necessary to have a precise and deep understanding of it. Therefore, a thorough and accessible description of the model is needed.

Often, the model is implemented in a general programming language such as Python (e.g., Noble et al. 2012). However, the implementation of complex models is often difficult to understand, as models are rather lengthy and burdened with simulation details, which dilute the essential model mechanisms. Due to their length and technical nature, they cannot be directly included in publications about the model. They are hard to understand for domain experts, who are not necessarily familiar with programming in general, or the used programming language in particular (Steiniger et al. 2014).

Grimm et al. (2006) proposed a standardized protocol, the ODD protocol, for the structured textual description of agent-based simulation models. The protocol is widely adopted and is also recommended for uploading agent-based models in model repositories such as the OpenABM model repository. Similar approaches exist in other application domains, e.g., PMRR (Preferred Model Reporting Requirements; Rahmandad and Sterman 2012) in the social sciences or the MIRIAM (Minimal Information Required in the Annotation of Models; (Novère et al. 2005) for biochemical models. While these standards facilitate an accessible and assessable documentation of the model, the resulting documents are not yet readily executable, and often crucial details to implement the model based on the documentation are missing.

Domain-Specific Modeling Languages (DSMLs) are aimed at bridging the gap between documentation and implementation of the model, with the ultimate goal to provide an executable documentation. DSMLs are designed to be used in one specific application domain. The use of domain metaphors (e.g., a social network in the social sciences or molecular bindings in biochemistry) allows for tailoring the language to the typical problems of the domain. Therefore, it is easier for the modeler to implement the model, and for a domain expert to understand the implementation. Practical expressiveness, i.e., how easy is it to specify a model of a domain in the language and can also more complex mechanisms be expressed, and succinctness are central requirements for the design of domain-specific modeling languages. Whereas the former is difficult to measure, requiring dedicated user studies (Kossow et al. 2016), an indication for the later is the used lines of code.

To demonstrate the advantages of domain-specific modeling languages, we have re-implemented the migration decision model in the Modeling Language for Linked

```
1  Person(
2     sex: {"m", "f"},
3     income: real := 0,
4     mortalityModifier: real := 1,
5     migrationStage: {"not viable", "intention" , "planning", "preparation", "
          migrated", "exit"} := "not viable",
6     migrationAttempts: int := 0,
7     failedMigrationAttempts: int := 0,
8     status: {"child", "adult", "retired"} := "child",
9     canAffordMonths: int := 0,
10    canNotAffordMonths: int := 0,
11    migrationAge: real := 0,
12    migrationStartAge: real := 0
13 );
14
15 parents:Person[2] <-> [0-]Person:children;
16 partner:Person[0-1] <-> [0-1]Person:partner;
17 friends:Person[0-] <-> [0-]Person:friends;
18
19 Person
20 | ego.inMigrationProcess(), ego.migrationIntention() >= 0, ego.migrationStage
       != "preparation"
21 @ ego.migrationAdvancementRate()
22 -> ego.advanceMigrationStage();
23
24 | ego.inMigrationProcess(), ego.migrationIntention() >= 0, ego.migrationStage =
        "preparation", ego.canAffordMigration()
25 @ ego.migrationAdvancementRate() * ego.borderEnforcementFactor()
26 -> ego.migrate();
27
28 Person.migrationAdvancementRate() := ?rho * e^(?a7 * ego.migrationIntention())
29 where ?rho := advancementRateBaseline,
30       ?a7 := advancementRateIntentionWeight;
31
32 Person.migrationIntention() := ?a4 * ?MA + ?a5 * ?SN + ?a6 * ?PBC
33 where ?a4 := attitudeWeight,
34       ?a5 := socialNormsWeight,
35       ?a6 := perceivedBehavioralControlWeight,
36       ?MA := ego.migrationAttitude(),
37       ?SN := ego.socialNorms(),
38       ?PBC := ego.perceivedBehavioralControl();
```

Fig. 25.3 The ML3 declaration of the agent type `Person` (lines 1–13), and the three possible kinds of links between persons (lines 15–17). Two of the ML3 rules concerned with the migration decision process (lines 19–26). Each rule corresponds with one arrow, or kind of equivalent arrows, in Fig. 25.2. The first rule corresponds the three arrows labeled with "waiting time expired". The second rule describes a successful migration attempt. Below that, the definition of the rate with which agents progress through the stages of the decision process (lines 28–30), and the definition of the migration intention based on the Theory of Planned Behavior (lines 32–38) are shown

Lives (ML3). ML3 is a domain-specific modeling language specifically designed to allow a succinct and understandable implementation of agent-based models with dynamic social networks (Warnke et al. 2017). The main entities of models implemented in ML3 are *agents*, which are interconnected via dynamic *links*. The behavior of agents is described by *stochastic rules*. While most agent-based models are executed with discrete time steps (events have a certain chance to happen each unit of time, e.g., each day or year), time in ML3 is continuous. This property makes it a good fit for the migration decision model, which is designed with continuous time in mind. An excerpt of the ML3 implementation of the model is shown in Fig. 25.3.

In comparison with the original model implementation using NetLogo, the implementation in ML3 has several advantages. First, model and simulation algorithm are strictly separated. In ML3, the modeler only has to define stochastic rules. The actual scheduling, the selection when which rule is applied, is done by a separate simulator, that has been implemented by the developers of the language. In the NetLogo implementation, this scheduling had to be done manually, using a continuous time extension (Sheppard and Railsback 2015). The separation of concerns in ML3 makes the model description much more succinct (about a seventh of the length) and readable. As less implementation has to be written, there is less room for errors. Also, it makes the scheduling logic reusable, allowing to put more effort into implementing and testing simulation algorithms, enabling the implementation of more efficient advanced simulation algorithms (e.g., Reinhardt and Uhrmacher 2017).

Separation of concerns does not only apply to model and simulation logic, but it also applies to the different components of the model. The stochastic rules that describe behavior in ML3 operate separately from each other in parallel. The different model components that are concerned with different processes operating in parallel, e.g., the fertility, mortality, and the migration decision process, can be implemented as separate sets of rules. This allows the modeler, to implement these different processes separately as separate component models, validate them separately, and then compose them to a complex model (Peng et al. 2017; Pierce et al. 2018). Not only does this again contribute to an easier identification and validation of specific mechanisms, but it is also feasible to exchange component models by simply exchanging the corresponding rules. That way one could, for example, compare different decision process components in a multi-model approach, as proposed by Gray et al. (2017). Additionally, the stochastic rules allowed us to implement the transitions through the stages of the decision process very naturally. The conceptual model of the decision process (Fig. 25.2) defines stages an agent can occupy, and the transitions to other stages they can make from each of those. Our implementation (Fig. 25.3) reflects this structure very directly, as it consists of one rule for every possible kind of transition.

The aspects of ML3 that make it domain-specific to the area of demography also play a major role in making the model description more succinct while enabling an easier implementation. For this model, two were of particular importance: the time-dependent transition rates and parameter maps. Many of the transition rates in demographic models depend on the age of the agent. For example, mortality is strongly age-dependent, as are fertility and marriage. ML3 allows such time dependency in transition rates and makes the aging of agents directly part of the language, which implies a more costly scheduling of events. Parameter maps are a way to deal with time series data, which is also used excessively in this and other demographic models.

The example shows the power of domain-specific modeling languages for assessing the structural validity of models and their role in modeling for explanation (Conte et al. 2012).

25.4 Managing an Experiment: Experiment Specification Languages

To gain confidence in the validity of a complex model, it is necessary to probe and explore its behavior thoroughly. Therefore, diverse simulation experiments are required (Klügl 2008; Leye et al. 2009). The results of simulation experiments, however, depend not only on the model but also on the context in which they are executed. This insight has led to formalizations of this context, most prominently Zeigler's *experimental frame* (Zeigler et al. 2000). Zeigler proposes to embed the model in an experimental frame, explicitly described as DEVS models, that generates the model inputs and analyzes the model outputs.

By enabling model users to access and repeat the simulation experiments conducted to validate a model, their confidence in the model's validity is increased. However, the two most frequently used ways to provide information about simulation experiments do not facilitate assessing the experiments done. First, executed experiments can be described informally, e.g., textually, or semi-formally as proposed in Grimm et al. (2006, 2010). Repeating the experiments is typically hindered by unambiguous or missing information or unavailable software or data. A better approach is to provide software artifacts to execute the experiments. However, such executable software is typically hard or even impossible to inspect, leaving unclear what experiment is getting executed by it. Additionally, technical issues, such as dependencies on third-party software, make running experiments difficult (or even impossible, for example, if the dependencies are not available anymore some time after the experiments have been published). To address the challenges of accessing, repeating, and thus assessing simulation experiments, explicitly specified simulation experiments that allow the replication of experiments are needed.

As shown in Fig. 25.2, simulation experiments consist of a multitude of experimental steps. For many of them, explicit specification languages exist. Observations could be specified using an instrumentation language (Helms et al. 2012). Single-run analysis might involve checking if the result trajectory of a simulation run fulfills a certain property. Temporal logic, such as LTL (Rozier 2011), allows specifying and checking such properties. Trace query languages, e.g., Laurent et al. (2018), can be used to specify and find specific transitions in a simulation run. In a stochastic model, properties might only hold with a certain probability. The analysis of such probabilistic properties is part of multi-run analysis, and formalisms such as MITL (Maler and Nickovic 2004) allow to specify them. Further examples and perspectives are given in Schützel et al. (2014).

Other approaches allow specifying the whole simulation experiment, integrating the description of all parts of the experiment into a single language, tying together different approaches into a unifying framework. SED-ML (Waltemath et al. 2011) has been developed in the SBML ecosystem in systems biology, originally to replicate published outputs of simulation experiments. Based on XML and cultivated by a standardization committee, SED-ML can be processed by many tools. These tools can interpret specifications with the standardized syntax and semantics, which enables tool-independent reproducible experiments. This makes SED-ML an effective exchange format for experiments. However, due to being a standard, new features can not easily be introduced in SED-ML. For example, parameter sweeps were not supported in the first release but introduced in Level 1 Version 2 (Bergmann et al. 2015).

The Simulation Experiment Specification on a Scala Layer (SESSL) (Ewald and Uhrmacher 2014) aims to mitigate this lack of flexibility. SESSL is an internal domain-specific language, which enables on-the-fly addition of features through its host language Scala, for example, to process simulation output data for further analysis (Peng et al. 2016). The resulting experiment specifications are valid Scala code with a declarative feel. Thus, SESSL experiments are readable as well as executable. Further, by using Maven (https://maven.apache.org) for artifact persistence and management, SESSL experiments can be reproduced across machines. This way, model users can access and repeat validation experiments more easily.

Before a model can be validated, values for its input parameters must be found—the model is *calibrated*. The migration model, for example, has eight weighting parameters that control the decision process of individuals. To find valid parameterizations of the model, methods to optimize a certain quality criterion can be employed, for example, to minimize the difference between the model output and a given observation. Choosing a metric to calculate the difference is not a trivial task and depends heavily on the data to compare. Similarly, diverse optimization algorithms are available. Typically, these have to be parameterized as well. To make such a calibration experiment accessible and replicable, this information has to be included. Figures 25.4 and 25.5 show how this can be realized with SESSL and its bindings for ML3 and Opt4j (Lukasiewycz et al. 2011).

Once calibrated, behavioral characteristics of the model can be checked by simulating the model and making sure that the observations from the simulation match some defined expectations, for example, derived from data. A formal framework for this approach is *Statistical Model Checking* (SMC) or simulation-based verification. SMC answers whether a random simulation run of a model satisfies a given property with at least a given probability (Agha and Palmskog 2018). Applying SMC to a model implies executing simulation runs, checking the property on each of the runs, and using hypothesis testing to infer statistically valid statements about the model's behavior. Thus, the properties to investigate must be defined on model outputs that are observable in a simulation run. Typically, temporal logics are used to express statements about the development of the model outputs in time. SMC experiments can be specified reproducibly by including the property to check as well as the statistical parameters in the experiment setup.

Apart from calibration and statistical model checking, many more types of experiments are required to validate a model. To make such experiments accessible, replicable and, thus, assessible, experiment specification languages need to support a wide variety of experiments, the set of which might be constantly growing. Particularly the later poses a challenge for the design and development of these languages, as they must satisfy constantly changing requirements while supporting a succinct and understandable description of experiments.

```
1  class MigrationExperiment extends Experiment with ParallelExecution with
       ParameterMaps {
2    model = "migration.ml3"
3    simulator = NextReactionMethod()
4    parallelThreads = -1
5    replications = 1
6    initializeWith(new JsonStateBuilder("initialstate.json"))
7    startTime = 1982
8    stopTime = 2050
9
10   fromFile("maleMortality.csv")()
11   fromFile("femaleMortality.csv")()
12   fromFile("fertility.csv")()
13   fromFile("income.csv")()
14   fromFile("ageDifferenceModifier.csv")()
15   fromFile("baseMarriageRate.csv")()
16   fromFile("borderEnforcement.csv")()
17   fromFile("disc.csv")()
18
19   set("minFertilityAge" <~ 12, "maxFertilityAge" <~ 49)
20   set("ageOfAdulthood" <~ 16, "ageOfRetirement" <~ 65)
21   set("minMarriageAge" <~ 9, "maxMarriageAge" <~ 60)
22   set("meanMigrationStartAge" <~ 17)
23   set("spouseAgeModifier" <~ -0.01301431)
24   set("intercept" <~ -0.490129556)
25   set("homeCountryGini" <~ 0.4, "hostCountryGini" <~ 0.3)
26 }
```

Fig. 25.4 An experiment with the migration model defined in SESSL 0.14. Line 1 declares a Scala class that represents a ML3 simulation experiment that uses parallel execution and parameterization of models with parameter maps (e.g., age-indexed). Lines 2–8 specify the model file, the simulation algorithm, the number of parallel threads to use, the number of replications to execute, the method to build the initial state, as well as the start and stop time of the simulation. In lines 10–17, files from which to read in parameters that are stored in CSV files with parameter maps are stated. Lines 19–25 specify some further scalar model parameters

25.5 Managing a Simulation Study: Provenance Models

Whereas the above domain-specific languages allow the user to succinctly specify a model and specify and execute a single simulation experiment, what (Rahmandad and Sterman 2012) requested in his *Preferred Model Reporting Requirements* (PMRR) to include in addition, i.e., information on the sources of data for the model's equations and algorithmic rules, has not been considered yet. However, given that not only data as input and data for calibration but also theories, such as the Theory of Planned Behavior (Ajzen 1991), and existing models such as the Heligman–Pollard mortality model (Heligman and Pollard 1980), and real-world experiments, contribute to a complex simulation model, only focusing on data will not suffice. Therefore, a more systematic inspection of a simulation model's provenance is asked for Ruscheinski and Uhrmacher (2017).

"Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness" (Groth and Moreau 2013). The Open Provenance Model (OPM) (Moreau et al. 2011) is a formalism to describe this provenance information as a directed graph. Nodes in this graph represent *artifacts*, *processes*, or

```
 1  val referenceMean = 2
 2
 3  minimize { (params, objective) =>
 4    execute {
 5      new MigrationExperiment with Observation {
 6        set("incomeEvaluationConstant" <~ params("a1"))
 7        set("incomeEvaluationCapitalWeight" <~ params("a2"))
 8        set("familyEvaluationWeight" <~ params("a3"))
 9        set("attitudeWeight" <~ params("a4"))
10        set("socialNormsWeight" <~ params("a5"))
11        set("perceivedBehavioralControlWeight" <~ params("a6"))
12        set("advancementRateIntentionWeight" <~ params("a7"))
13        set("advancementRateBaseline" <~ params("rho"))
14
15        observeAt(Change(agentType = "Person", field = "migrationStage",
16          filter = "ego.migrationStage = ’migrated’ && ego.planningTime != 0")) {
17          observe("decisionDuration" ~ expression("ego.planningTime"))
18        }
19
20        var ages = List.empty[Double]
21
22        withRunResult { result =>
23          if (result ? "decisionDuration")
24            ages ++= result.values("decisionDuration").asInstanceOf[Iterable[
                     Double]]
25        }
26
27        withReplicationsResult { result =>
28          val mean = if (ages.nonEmpty) math.abs(referenceMean - (ages.sum / ages
                   .size))
29                     else               Double.PositiveInfinity
30          objective <~ mean
31        }
32      }
33    }
34  } using new Opt4JSetup {
35    param("a1", 100.0, 5, 300)
36    param("a2", 0.0, 0.1, 2)
37    param("a3", 0.0, 5, 200)
38    param("a4", 0.0, 0.1, 2)
39    param("a5", 0.0, 0.1, 2)
40    param("a6", 0.0, 0.1, 2)
41    param("a7", 0.0, 1E-5, 1E-4)
42    param("rho", 0.0, 0.01, 1)
43    optimizer = ParticleSwarmOptimization(particles = 10, iterations = 20)
44
45    withOptimizationResults { results =>
46      println("Overall results: " + results.head)
47    }
48  }
```

Fig. 25.5 A calibration experiment defined in SESSL 0.14, using the migration experiment class from Fig. 25.4. The experiment uses a particle swarm optimization algorithm (line 43) from the Opt4j package (Lukasiewycz et al. 2011) that tunes the parameters specified with ranges in lines 35–42. The parameter ranges can be chosen based on model assumptions, but here we set them based on the known optimum (Klabunde et al. 2017). Lines 6–13 then read the parameters set by the optimizer and apply them to the model when running simulations. The observation of the model is configured in lines 15–18: the time between starting to plan a migration and actually migrating is recorded for every agent that actively migrates, i.e., is not brought with another migrating agent. The calculation of the target function to minimize is realized in lines 20–31. Specifically, line 20 declares a variable in which the planning times observed in single runs are stored (line 24). Lines 28–30 aggregate the durations to the mean and compute the difference to a reference mean (specified in line 1)

agents, while edges indicate dependencies between them. *Artifacts* are digital representations of entities within a computer system, e.g., a simulation model, a data set, or a simulation result. *Processes* are activities performed with artifacts to generate new artifacts, e.g., extending a model, or performing simulation experiment. Finally, *agents* are the persons enabling and controlling the processes. Between these elements, five dependencies are distinguished: 1. An artifact was *used* in a process, 2. An artifact *was generated by* a process, 3. A process *was controlled by* an agent, 4. A process *was triggered by* another process, and 5. An artifact *was derived from* another artifact.

As an example, we have reconstructed the provenance information about the migration decision model, using the publication about the model (Klabunde et al. 2017), the ODD description (Klabunde et al. 2015), and the information provided together with the model in the OpenABM model repository (Klabunde et al. 2016). We focus on the information about artifacts and processes and leave out information about agents, as we have little information about who exactly was involved in each process. A general approach is outlined in more detail in Ruscheinski and Uhrmacher (2017).

The result (published in Reinhardt et al. 2018b) is shown in Fig. 25.6. The top left of the figure shows the decision model itself as an artifact (*mig. mod.*). It was produced by composing (process *comp. model*) a set of submodels, or model components. While the available information does not allow for a more detailed description of that process, we can present a lot of information about the creation of the components. We will look at the top one, the income model (*inc.*), in more detail. The income model is based on assuming a log-normal income distribution (artifact *LN inc.*). It is parameterized with two parameters, describing the mean income (*mean inc.*) and the variation of income (*gini*). The former is estimated (*est. mean*) using GDP and employment data published by the International Monetary Fund for Senegal (*GDP* and *empl.*). The latter is derived from the *World Development Indicators* (*WDI*). Finally, the complete income model was compiled by parameterizing the log-normal distribution accordingly (*comb.inc.*). Please note, that the figure only shows the provenance model's structure. In the complete provenance model, each of the artifacts and processes is annotated with further information about them.

The provenance model shows the relations between the processes and artifacts contributing to the simulation model, making their interdependencies become explicit. This explicitness can be used to improve trust into the model. We have now made it explicit, that the income component of the model is based on a log-normal income distribution. That model is widely applied and its validity for different applications has been assessed (Bandourian et al. 2002). At the same time, the provenance model tells us, which data was used to fit the log-normal distribution to the Senegal case. This provides an overview about which data sets were used for which parts of the simulation model—an information especially important when it comes to validating the model, as one has to make sure that the model is not validated using the same data it was calibrated to. Furthermore, using additional information about assumptions, the established models, and theories annotated to the artifacts and the way the data was collected, the provenance model allows to reason about the adequacy of the

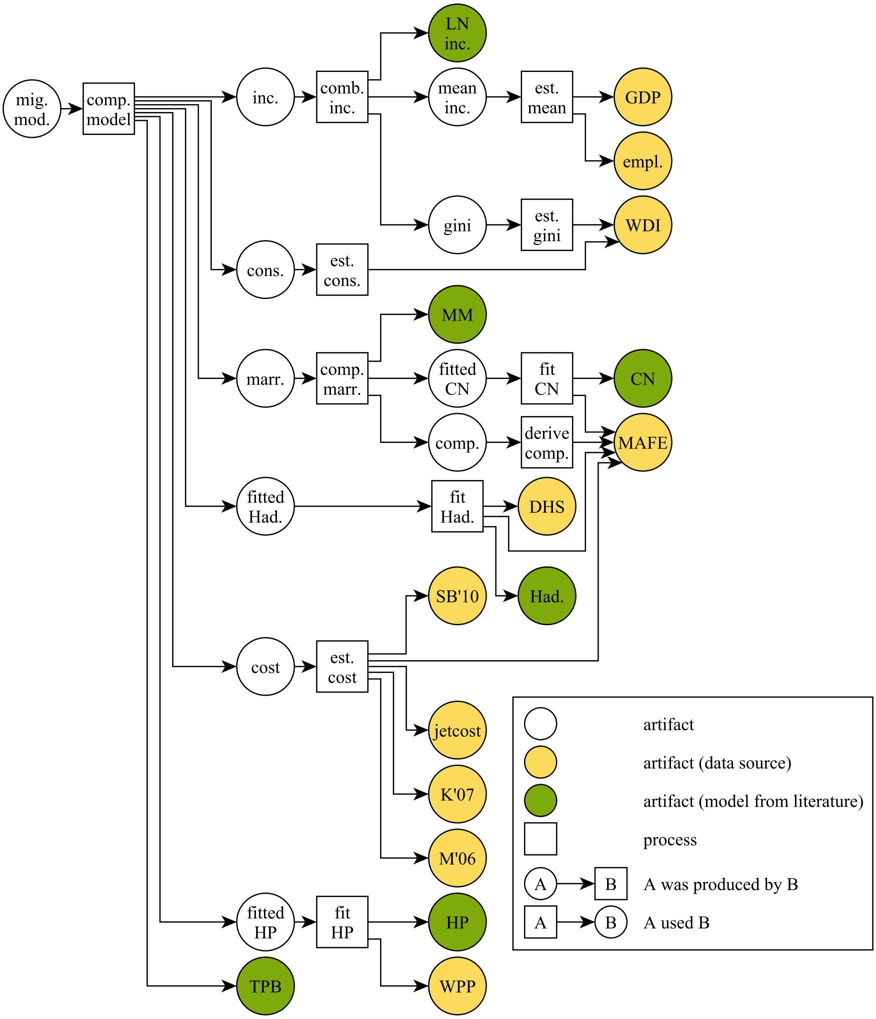


Fig. 25.6 Open Provenance Model for the migration decision process model, as we have derived it from the publications about the model. *Source* (Reinhardt et al. 2018b). See there for additional information on the artifacts and processes

used theories and data-sources. But the provenance model does not only consider the artifacts but also the processes through which they were derived. It makes the process of fitting explicit, pointing the modeler to the need to document it, and making a later reader aware of it. Further annotations can reveal information about the methods used. All together, this information gives us trust in the income component of the migration model. In general, the provenance information enables us to reconstruct assumptions made in the components, the theories they were derived from,

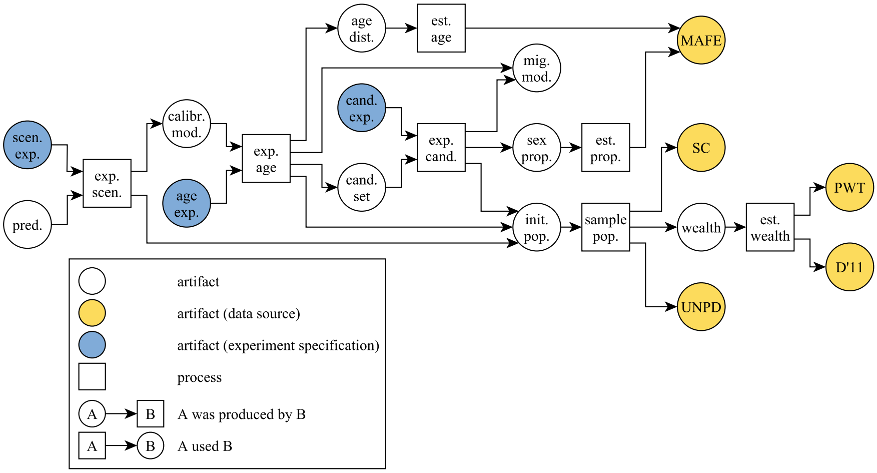


Fig. 25.7 Open Provenance Model for the calibration experiment. Note that the artifacts *mig. model* and *MAFE* are identical to the artifacts of the same name in Fig. 25.6. *Source* (Reinhardt et al. 2018b). See there for additional information on the artifacts and processes

methods used for developing them, and data sources used for fitting the components, and allows for tracing them to their origin. We can use provenance to reason about the artifacts used and processes executed for developing the model components. For example, if we identify a methodological error in the collection of a data set, we can use the provenance model to infer affected model components which need to be revised. For this, inference mechanisms such as OPQL (Lim et al. 2011) can be employed.

To the provenance of the simulation models belongs also information about the simulation experiments. In Fig. 25.7, the migration process model artifact (*mig. mod.*) is shown once again. The model was calibrated in two steps (Klabunde et al. 2017). In the first step, a set of candidate configurations (*cand. set*) was searched, which are able to reproduce the sex proportion of migrants estimated from the MAFE data (*sex prop.*). This is captured by the process *exp. cand.*, which produces the candidate set, as well as an experiment specification (*cand. exp.*). In this context, an experiment specification is anything that allows to reproduce the experimental steps to recreate the produced data, e.g., a textual description or an SESSL script as in Fig. 25.5. In the second calibration step, the candidate configuration that most closely reproduces the distribution of migrant ages observed in MAFE is chosen (*exp.age*). The result is the calibrated model (*calibr. mod.*).

Provenance information about the simulation experiment can be used, similar to provenance of the model, to trace the origin of data and methods used in the experiment. Further, we can use the provenance information to find all artifacts used for the execution of an experiment: the experiment specification produced by the experimentation process, and all artifacts used by this process. This information can be used to bundle these artifacts into a container, which can be shared, allowing to replicate the experiment result.

The reconstruction of the provenance data by tracing publications using the model documentation is cumbersome, as is the manual documentation of provenance information during the modeling process. Therefore, (semi-)automatic methods for retrieving provenance data are needed. While tool support is not yet readily available, different techniques were proposed (Ruscheinski and Uhrmacher 2017), where each method allows retrieving different parts of the provenance information.

First, we can annotate parts of the simulation model and experiments, which allows to specify all information related to the process of creating the simulation model or executing the simulation experiment.

To derive the provenance information from these annotations, the annotations need to be parsed and analyzed. This approach demands a close annotation of all artifacts to derive the full provenance model. However, it is up to the user to annotate the simulation model and experiments as part of the documentation and therefore can only be seen as a semiautomatic approach.

When scripts, scientific workflow, or domain-specific languages are employed to execute automatic tasks, those can also be used as source of provenance information, by integrating the derivation of provenance information into them. All these approaches allow to describe data-driven processes, but differ in the way the process is described and which features are provided by the execution environment. In scripting environments, like Python or R, the script is executed by the environment and it is up to the user to implement management procedures for the data by himself. Scientific workflows are often described by a graph containing nodes representing data-processing activities whereas the edges represent the flow of the data.

Finally, domain-specific languages can be used to describe a simulation experiment, as we have demonstrated in Sect. 25.4. First, the execution of the script or created simulation experiment can be observed to determine all read and created files (Murta et al. 2014), which become individual artifacts in the provenance model. And second, the script itself can call methods to store provenance information (Bochner et al. 2008). Scientific workflow environments often provide features to retrieve the provenance information directly after the execution of the workflow (Scheidegger et al. 2008).

Finally, a version control system can be used to derive provenance information by tracking changes to document. For example, if the migration model artifact is stored in a document, we can track how the model changes over time. However, the version control system can only capture the changes to artifacts, but not the processes that produce the changes. This can be supported using tools like Git2PROV (De Nies et al. 2013), which retrieves a provenance model from the commit history of a git repository.

25.6 Discussion

While the approach of "making it explicit" brings significant advantages for modeling, conducting simulation experiments, and the whole simulation study, it is not always without a cost. Domain-specific languages, be it for modeling or for simulation, make the task they are designed for easier. Some model aspects, e.g., continuous time, might even be infeasible without dedicated support from a simulation system. However, they are also putting restrictions on the user, compared to a general-purpose programming language. A domain specific modeling language is designed for a specific kind of simulation model, with specific properties. Models without these properties, or with features not accounted for by the DSML, will be difficult or even impossible to realize. Hence, the right DSML must be chosen carefully. Furthermore, one must take care not to choose the language too early in the modeling process, as "the limits of [one's] language mean the limits of [one's] world" (Wittgenstein 1922, 5.6). In the worst case, this might result in a model that is designed to fit the language, not the questions it shall answer. This can be partly avoided in an internal DSL such as SESSL, which allows the addition of features by directly using its host language Scala (see Reinhardt et al. 2018a, Sect. 6) for a demonstration). However, the additional freedom comes at a cost as well, e.g., making it impossible to give a formal semantics of the language.

At the same time, the user of a language is always dependent on the available support for using the language. A mature and commonly used programming language comes with dedicated development tools, e.g., editors or debuggers, ample documentation, and a large community of language users. A prototypical DSL for niche applications will lack this level of support. In addition to the language features, this language ecosystem must be considered to choose the right languages for a simulation study.

25.7 Conclusion

The validation of simulation models provides many challenges. The modeler's idea of the mechanisms of the model, the conceptual model, must be implemented in some programming language or modeling framework to gain an executable simulation model. Complex simulation experiments must be conducted to calibrate and validate the model, relating it to data collected from the modeled real-world system. Domain-specific modeling language provides an executable, yet succinct model representation, reflecting metaphors of the application field and supporting structures and dynamics considered essential for modeled systems in a particular area of modeling and simulation application. Thereby, domain-specific modeling languages facilitate a model's development, its maintenance, including its reuse, as well as its validation, e.g., by inspection. Domain-specific languages for specifying experiments support not only the documentation but also the replication of simulation experiments. This equally refers to exploratory experiments, and experiments done for calibration or validation. Therefore, domain-specific languages for specifying experiments allow to retrace and replicate research efforts invested in validating a simulation model and thus give insight into a model's behavioral repertoire and validity. Crucial at this point are the flexibility and extensibility of the languages to account for the increasing number of simulation methods that, for example, a demographic multilevel model should be subjected to. Statistical model checking methods are based on explicitly specifying expectations for the behavior of the model in a formal domain-specific language, which makes the behavioral assumptions that a simulation model should adhere to explicit, unambiguous, and automatically verifiable.

Even with a lack of data, evidence for the validity of a simulation model can be found by examining the history, or provenance, of the simulation model. Exploiting a provenance model allows to relate and query the diverse artifacts and processes that contributed to a simulation model. This way, experiment specification, theories that underlie the model, different variants of a simulation model, as well as data used as input, for calibration, or for validation can be linked even beyond individual simulation studies. All together reveal not the whole, but a crucial part of the tale about the science and art of modeling.

Acknowledgements This research was supported by the German Research Foundation (DFG) via research grant UH-66/15 and research grant UH-66/18.

References

Agha, G., & Palmskog, K. (2018). A survey of statistical model checking. *ACM Transactions on Modeling and Computer Simulation*, *28*(1), 6:1–6:39.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211.

Balci, O. (1997). Verification, validation and accreditation of simulation models. In *Proceedings of the 1997 winter simulation conference*, pp. 135–141. IEEE.

Bandourian, R., McDonald, J., & Turley, R. S. (2002). A comparison of parametric models of income distribution across countries and over time. SSRN Scholarly Paper ID 324900, Social Science Research Network, Rochester, NY, June 2002.

Beauchemin, C. (2015). Migration between Africa and Europe (MAFE): Looking beyond immigration to understand international migration. *Population*, *70*(1), 13–38.

Bergmann, F. T., Cooper, J., Le Novere, N., Nickerson, D., & Waltemath, D. (2015). Simulation experiment description markup language (SED-ML) Level 1 Version 2. *Journal of Integrative Bioinformatics (JIB)*, *12*(2), 119–212.

Bochner, C., Gude, R., & Schreiber, A. (2008). A python library for provenance recording and querying. *Provenance and Annotation of Data and Processes*, pp. 229–240.

Coale, A. J., & Mcneil, D. R. (1972). The distribution by age of the frequency of first marriage in a female cohort. *Journal of the American Statistical Association*, *67*(340), 743–749.

Conte, R., et al. (2012). Manifesto of computational social science. *European Physical Journal-Special Topics*, *214*, 325.

Davies, J. B., Sandström, S., Shorrocks, A. B., & Wolff, E. N. (2011). The level and distribution of global household wealth. *The Economic Journal*, *121*(551), 223–254.

De Nies, T., et al. (2013). Git2PROV: Exposing version control system content as w3c prov. In *Proceedings of the 2013th International Conference on Posters & Demonstrations Track-Volume 1035* (pp. 125–128). CEUR-WS. org.

Ewald, R., & Uhrmacher, A. M. (2014). SESSL: A domain-specific language for simulation experiments. *ACM Transactions on Modeling and Computer Simulation*, *24*(2), 11:1–11:25.

Gray, J., Hilton, J., & Bijak, J. (2017). Choosing the choice: Reflections on modelling decisions and behaviour in demographic agent-based models. *Population Studies*, *71*(sup1), 85–97

Grimm, V., et al. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, *198*(1), 115–126.

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The ODD protocol: A review and first update. *Ecological Modelling*, *221*(23), 2760–2768.

Groth, P., & Moreau, L. (2013). PROV-overview. An overview of the PROV family of documents.

Harris, L. A., et al. (2016). Bionetgen 2.2: Advances in rule-based modeling. *Bioinformatics*, *32*(21), 3366–3368.

Heligman, L., & Pollard, J. H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, *107*(1), 49–80.

Helms, T., Himmelspach, J., Maus, C., Röwer, O., Schützel, J., & Uhrmacher, A. M. (2012). Toward a language for the flexible observation of simulations. In *Proceedings of the 2012 Winter Simulation Conference*, pp. 418:1–418:12. IEEE.

IEEE. (2009). IEEE Standard VHDL Language Reference Manual. *IEEE Std 1076-2008 (Revision of IEEE Std 1076-2002)* (pp. c1–626).

Klabunde, A., Willekens, F., Zinn, S., & Leuchter, M. (2015). *An agent-based decision model of migration, embedded in the life course—Model description in ODD+D format*. Max Planck Institute for Demographic Research, Rostock, Germany: Technical report.

Klabunde, A., Zinn, S., Willekens, F., & Leuchter, M. (2016). Multistate modeling extended by behavioral rules (Version 6). https://www.openabm.org/model/5146/version/6/view.

Klabunde, A., Zinn, S., Willekens, F., & Leuchter, M. (2017). Multistate modeling extended by behavioral rules—an example of migration. *Population Studies*, *71*(sup1), 51–67

Klügl, F. (2008). A validation methodology for agent-based simulations. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC'08 (pp. 39–43). New York: ACM.

Kossow, C., Helms, T., Kreutzer, J. M., Martens, A., & Uhrmacher, A. M. (2016) Evaluating different modeling languages based on a user study. In *Proceedings of the 49th Annual Simulation Symposium*, ANSS '16 (pp. 18:1–18:8). Society for Computer Simulation International, San Diego, CA, USA.

Laurent, J., Medina-Abarca, H. F., Boutillier, P., Yang, J., & Fontana, W. (2018). A trace query language for rule-based models. In *Computational Methods in Systems Biology (CMSB 2018)*, Lecture Notes in Bioinformatics. Cham: Springer.

Leye, S., Himmelspach, J., Uhrmacher, A. M. (2009). A discussion on experimental model validation. In *2009 11th International Conference on Computer Modelling and Simulation* (pp. 161–167).

Lim, C., Lu, S., Chebotko, A., & Fotouhi, F. (2011). OPQL: A first OPM-level query language for scientific workflow provenance. In *2011 IEEE International Conference on Services Computing*, pp. 136–143. IEEE.

Lukasiewycz, M., Glaß, M., Reimann, F., & Teich, J. (2011). Opt4J—a modular framework for metaheuristic optimization. In *Proceedings of the Genetic and Evolutionary Computing Conference (GECCO 2011)* (pp. 1723–1730), Dublin, Ireland.

Maler, O., & Nickovic, D. (2004). Monitoring temporal properties of continuous signals. *Formal Techniques* (pp. 152–166), Modelling and Analysis of Timed and Fault-Tolerant Systems, Lecture Notes in Computer Science. Heidelberg: Springer.

Maus, C., Rybacki, S., & Uhrmacher, A. M. (2011). Rule-based multi-level modeling of cell biological systems. *BMC Systems Biology*, *5*, 166.

Moreau, L., et al. (2011). The open provenance model core specification (v1. 1). *Future Generation Computer Systems*, *27*(6), 743–756.

Murta, L., Braganholo, V., Chirigati, F., Koop, D., & Freire, J. (2014). Noworkflow: Capturing and analyzing provenance of scripts. In *International Provenance and Annotation Workshop* (pp. 71–83). Springer.

Noble, E. et al. (2012). Linked lives: The Utility of an agent-based approach to modeling partnership and household formation in the context of social care. In *Proceedings of the 2012 Winter Simulation Conference*, pp. 93:1–93:12. IEEE.

Novère, N. L., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., et al. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology*, *23*(12), 1509–1515.

Peng, D., Warnke, T., Haack, F., & Uhrmacher, A. M. (2016). Reusing simulation experiment specifications to support developing models by successive extension. *Simulation Modelling Practice and Theory*, *68*, 33–53.

Peng, D., Warnke, T., Haack, F., & Uhrmacher, A. M. (2017). Reusing simulation experiment specifications in developing models by successive composition—a case study of the wnt/β-catenin signaling pathway. *SIMULATION*, *93*(8), 659–677.

Pierce, M. E., Krumme, U., & Uhrmacher, A. M. (2018). Building simulation models of complex ecological systems by successive composition and reusing simulation experiments. In *Proceedings of the 2018 Winter Simulation Conference*. IEEE.

Rahmandad, H., & Sterman, J. D. (2012). Reporting guidelines for simulation-based research in social sciences. *System Dynamics Review*, *28*(4), 396–411.

Reinhardt, O., & Uhrmacher, A. M. (2017). An efficient simulation algorithm for continuous-time agent-based linked lives models. In *Proceedings of the 50th Annual Simulation Symposium*, ANSS'17, pp. 9:1–9:12, San Diego, CA, USA. Society for Computer Simulation International.

Reinhardt, O., Hilton, J., Warnke, T., Bijak, J., & Uhrmacher, A. M. (2018a). Streamlining simulation experiments with agent-based models in demography. *Journal of Artificial Societies and Social Simulation*, *21*(3), 9.

Reinhardt, O., Ruscheinski, A., & Uhrmacher, A. M. (2018b). Odd+p: Complementing the odd protocol with provenance information. In *Proceedings of the 2018 Winter Simulation Conference*. IEEE.

Rozier, K. Y. (2011). Linear temporal logic symbolic model checking. *Computer Science Review*, *5*(2), 163–203.

Ruscheinski, A., & Uhrmacher, A. M. (2017). Provenance in modeling and simulation studies-bridging gaps. In *Proceedings of the 2017 Winter Simulation Conference*. IEEE Press.

Rybacki, S., Leye, S., Himmelspach, J., & Uhrmacher, A. M. (2012). Template and frame based experiment workflows in modeling and simulation software with worms. In *2012 IEEE Eighth World Congress on Services (SERVICES)* (pp. 25–32). IEEE.

Scheidegger, C., Koop, D., Santos, E., Vo, H., Callahan, S., Freire, J., et al. (2008). Tackling the provenance challenge one layer at a time. *Concurrency and Computation: Practice and Experience*, *20*(5), 473–483.

Schützel, J., Peng, D., Uhrmacher, A. M., & Perrone, L. F. (2014). Perspectives on languages for specifying simulation experiments. In *Proceedings of the 2014 Winter Simulation Conference* (pp. 2836–2847). IEEE Press (2014).

Sheppard,C. J., & Railsback, S. (2015). Time Extension for NetLogo (Version 1.2) [Software]. https://github.com/colinsheppard/time.

Steiniger, A., Uhrmacher, A. M., Zinn, S., Gampe, J., & Willekens, F. (2014). The role of languages for modeling and simulating continuous-time multi-level models in demography. In *Proceedings of the 2014 Winter Simulation Conference*, pp. 2978–2989. IEEE.

Troitzsch, K. G. (2004). Validating simulation models. In G. Horton (ed.), *18th European Simulation Multiconference. Networked Simulations and Simulation Networks* (pp. 265–270).

Troitzsch, K. G. (2017). Using empirical data for designing, calibrating and validating simulation models. *Advances in Social Simulation 2015* (pp. 413–427). Advances in Intelligent Systems and Computing. Cham: Springer.

Uhrmacher, A. M., Brailsford, S., Liu, J., Rabe, M., & Tolk,A. (2016). Panel—reproducible research in discrete event simulation—a must or rather a maybe? In *Proceedings of the 2016 Winter Simulation Conference*, pp. 1301–1315. IEEE.

van Deursen, A., Klint, P., & Visser, J. (2000). Domain-specific languages: An annotated bibliography. *SIGPLAN Notices*, *35*(6), 26–36.

Waltemath, D., Adams, R., Bergmann, F. T., Hucka, M., Kolpakov, F., Miller, A. K., et al. (2011). Reproducible computational biology experiments with SED-ML—the simulation experiment description markup language. *BMC Systems Biology*, *5*(1), 198.

Warnke, T., Reinhardt, O., Klabunde, A., Willekens, F., & Uhrmacher, A. M. (2017). Modelling and simulating decision processes of linked lives—an approach based on concurrent processes and stochastic race. *Population Studies*, *71*(sup1), 69–83.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. London: Routledge & Kegan Paul.

Zeigler, B. P., Praehofer, H., & Kim, T. G. (2000). *Theory of Modeling and Simulation* (2nd ed.). San Diego: Academic Press.

Zinn, S. (2012). A mate-matching algorithm for continuous-time microsimulation models. *International Journal of Microsimulation*, *5*(1), 31–51.

Part VII
Validation at Work—Best Practice-Examples

Chapter 26
Validation of Particle Physics Simulation

Peter Mättig

Abstract The procedures of validating computer simulations of particle physics events at the LHC are summarized. Because of the strongly fluctuating particle content of LHC events and detector interactions, particle-based Monte Carlo methods are an indispensable tool for data analysis. Simulation in particle physics is founded on factorization and thus its global validation can be realized by validating each individual step in the simulation. This can be accomplished by drawing on results of previous measurements, in situ studies, and models. What is particularly important in particle physics is to quantify how well a simulation is validated such that a systematic uncertainty can be assigned to a measurement. The simulation is tested for a wide range of processes and agrees with data within the assigned uncertainties.

Keywords LHC · Simulation · Validation · Higgs

26.1 Introduction

Computer simulations are an important tool for present-day particle physics. The aim of this chapter is to characterize the procedures of validating one type of simulations in this area, viz., the simulation of particle physics events. As an example, the ATLAS experiment at the Large Hadron Collider (LHC) is considered. The simulation for other LHC experiments, e.g., CMS (2008), and the specialized experiments ALICE (2008) and LHCb (2008) work along fairly similar principles.

The chapter is organized as follows. After a short introduction into particle physics in Sect. 26.2, it will start with a short overview of the principles of data analysis and the use of simulations in this area (Sect. 26.3). It will then discuss the two basic

P. Mättig (✉)
Physikalisches Institut der Universität Bonn, Bonn, Germany
e-mail: peter.mattig@cern.ch

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_26

ingredients that enter simulation, the modeling of the physics process (Sect. 26.4) and that of the detector (Sect. 26.5). It will then characterize the principles of validation (Sect. 26.6) to motivate validation procedures in particle physics (Sects. 26.7–26.9) before discussing examples of validation in some detail and how simulation and validation works in typical analyses (Sect. 26.10). Finally, some points raised in the philosophical literature on simulations are commented (Sect. 26.11).

26.2 What Particle Physics is About: Example LHC

During the 1960s and 1970s, the descriptions of the electromagnetic, weak and strong interactions were put on a common ground leading to a theoretical framework, the Standard Model (SM) (Quigg 2013; Pich 2012). It is based on Quantum Field Theory and arguably the most encompassing and precisely tested theory of nature ever. It accommodates all measurements in an energy range of several 100 GeV[1] with just a few fundamental particles, which can be classified in three sectors:

- Twelve spin 1/2 particles (fermions) that can be separated in quarks and leptons are interpreted as "matter" particles.
- Three kinds of spin 1 particles (vector bosons), the photon (γ), the W, Z^0, and gluons, that transmit the electromagnetic, weak, and strong interactions of the matter particles, respectively.
- One spin 0 particle (scalar boson), the Higgs boson, to generate the masses of the fermions, force carriers and itself.

The SM predicts the complete dynamics of these particles based on 19 free parameters.[2] These particles and their interactions appear to be the basis for nucleons, atoms and eventually all phenomena beyond and are crucial for our current understanding of cosmology.

26.2.1 The Status of the Standard Model

The development of the conceptual framework was followed by an intensive experimental and theoretical program to establish the existence of all SM components, measure the free parameters (Patrignani et al. 2016) and test their dynamical properties. All components of the SM have been observed; the last one, the Higgs boson, has been found in 2012 at the Large Hadron Collider (LHC) (Evans and Bryant 2008) of the European Center for Particle Physics (CERN, Geneva), and almost all allowed

[1] Using quantum mechanical relations this energy range can be interpreted as 10^{-18} m, which is about 100 million times smaller than an atom.

[2] The number increases if the masses of the neutrinos are considered. Not all of the parameters related to neutrinos have been measured. Since they do not affect physics at the LHC, the subject of this paper, they will not be considered further.

interactions have been confirmed. The free parameters are measured to very high precision, e.g., the values of the coupling strength of the strong interaction and the mass of the Z^0, which will be used later, are determined as

$$\alpha_s(M_Z) = 0.1181 \pm 0.0011, \quad M_Z = 91.1876 \pm 0.0021 \text{ GeV} \tag{26.1}$$

Despite intensive searches, no effect has been observed at accelerators that does not agree with the SM. Therefore, physicists consider the SM as being confirmed as a theory for an energy scale of up to several 100 GeV. The SM is also internally consistent, i.e., void of any infinities even at energies many orders of magnitude beyond what can be probed in any foreseeable future.

Still, there is a reluctance to accept the SM as a final theory. For one thing, the SM does not explain how many particles there are, why they are organized in certain families and what the values of the parameters are. In addition, the SM does not contain gravity, which should only become relevant at $\mathrm{M}_{\mathrm{Planck}} = 10^{19}$ GeV, and it is not able to explain astrophysical observations like the existence of Dark Matter (Gelmini 2015) and Dark Energy (Planck Collaboration 2015). To address these issues, models have been devised that lead to physics beyond the SM (BSM). Searches for BSM physics have so far been futile, but they have moved more and more into the focus of experimental and theoretical activities.

26.2.2 The Forefront Experiment: LHC

Searches for BSM physics require higher precision on SM processes and in particular higher energies. The energy frontier of particle physics is the LHC, where every 25 nanoseconds a bunch of 10^{11} protons crosses another bunch flying in the opposite direction. Currently, each of the protons has an energy of 6.5 TeV leading to the highest collision energies reached in an accelerator and a huge rate of collisions. Protons are composed of subcomponents, quarks, or gluons, denoted together as partons. At the LHC, the protons serve only as vehicles for the partons and high energetic parton collisions are of the main interest. They produce a large variety of very different physics processes.

Each LHC bunch crossing leads to “an event” consisting of a spray of some 1000 particles, which are recorded by highly precise and sophisticated detectors covering almost the whole solid space angle and organized in different components. Their electronic recordings are translated into “physics objects”, i.e., candidates for electrons, quarks, photons, etc.—each one a stable SM particle. The content of an event, i.e., how many of these different objects are found and their relation, e.g., relative angle or mass, make up the “event signature”.

As indicated, in this chapter, the ATLAS experiment (ATLAS Collaboration 2008) will be considered as the example.

The ATLAS detector extends 40 m along the beam line and has a diameter of 25 m in the transverse plane. Each of its components is sensitive to a particular type of particle, offering redundant and rather comprehensive information. The wide coverage of the particles from the interaction allows ATLAS to probe almost all of the LHC physics. In a sense ATLAS combines some 100 distinct experiments of previous times. This allows for optimal use of each event including cross calibrations, an essential point for the validation of the simulation—as will be discussed later.

The large number and variety of particles in each event reflects the statistical properties of quantum mechanics. The main experimental challenge is to infer the underlying parton scattering from the different event signatures. It is at this point that simulation as a tool for data analysis enters the field. Describing the large number of particles, their complicated and statistically distributed structure, as well as the statistical nature of their interactions in the detector cannot be achieved analytically but requires numerical methods.

Before discussing validation, the article will start with a short overview of the principles of data analysis and the two basic ingredients that enter simulation, the modeling of the physics process and that of the detector. It will then formalize the principles of simulation to motivate validation procedures in particle physics.

26.3 Data Analysis and the Use of Simulations

Before describing the validation process in more detail in Sect. 26.6, we will summarize the ingredients of the simulation in particle physics. The simulation is particle based, involving a wide range of different scales and is realized with computer codes applying Monte Carlo (MC) methods.

26.3.1 From Data to Physics

The principal aim of the simulation of LHC events is to understand how the physics process ("signal", S) of interest would look in the detector. Comparing the measurement with this expectation allows one to extract information about the physics process of interest, e.g., a parameter of the SM, dynamic properties, or evidence, respectively, exclusion of BSM effects.

Each S has a certain event signature. However, such a signature is not unique, but can also be due to competing processes, the 'background B'. To reach high sensitivity, analyses aim at a good S/B ratio by applying special selections exploiting different properties of the S and the B processes. Arriving at a physics conclusion from the data requires the measurement to be compared to theory and thus to simulate both S and B.

In this article, basic concepts of validation will often be discussed along the production of a Higgs (h), and its detection via Z^0 bosons and electrons.[3]

$$g + g \rightarrow h \rightarrow Z^0 Z^{0*} \rightarrow (e^+e^-)(e^+e^-) \tag{26.2}$$

i.e., two gluons g produce a Higgs that decays into two Z^0 bosons,[4] each of which then decays into pairs of electrons. This Higgs signal competes with background processes that also lead to four electrons but do not involve the production of a Higgs.

A key observable for the Higgs production and decay, but also of general importance, is the cross-section σ_S, essentially a measure of the production yield. The cross section of a signal process is obtained from the measurement by

$$\sigma_S = \frac{N_{\mathrm{sel}} - N_{\mathrm{B}}}{\mathscr{L} \cdot \varepsilon_S} \tag{26.3}$$

where $\mathscr{L}$ is the luminosity, the total number of all possible proton–proton collisions provided by the LHC, N_{sel} is the number of data events selected, here with four electrons, N_{B}, the number of background events, and ε, the efficiency to detect a selected signal event, i.e.,

$$\varepsilon_S = \frac{N^{\mathrm{S}}_{\mathrm{sel}}}{N^{\mathrm{S}}_{\mathrm{prod}}} \tag{26.4}$$

where $N^{\mathrm{S}}_{\mathrm{sel}}$, $N^{\mathrm{S}}_{\mathrm{prod}}$ are the numbers of signal events that are selected, respectively, produced.

26.3.2 The Role of Simulation for Data Analysis

Both N_B and ε_S are obtained from simulation and should be determined to a high precision. Per year some 10 trillion events are simulated for ATLAS, each simulation taking several minutes. To achieve this, 100,000 CPUs and 100s of petabytes of disk storage are provided by a worldwide computing grid. To guarantee a constant quality, the simulation is continuously checked with benchmark processes.

Simulations in particle physics are used for several purposes.

- In the data analysis, they are instrumental to obtain model predictions and their experimental signatures that can be compared to measurements.

This is the broadest and most challenging application that will be primarily considered in this article. In addition, simulation is used to optimize tools and strategies.

[3] For simplicity particles and anti-particles will just be denoted by the name of the particle.

[4] The notation Z^{0*} means that the boson is “off shell”, i.e., its mass is different from the default 91 GeV due to quantum mechanical uncertainty.

- Both in experimental and theoretical studies, simulation is used to evaluate the feasibility of a technique.
- Simulations are used to optimize the layout of future detectors, or detector conditions to run an experiment.

In all three use cases, simulation is an auxiliary method to experiments, but in no way replaces measurements with real experiments. E.g., even if a detector component is first devised with simulation, it is only after studying an actual prototype, that the component will be integrated into the experiment. Simulations allow complicated manipulations of fairly involved models. Their results are believable only to the extent that the input and structure of simulations is believable, which in turn requires the input to and the structure of simulation to agree with experimental data. I.e., validation of the simulation is essential.

In particle physics, two main parts have to be simulated, the underlying physics processes leading to a distribution of hadrons, photons, and leptons, which can be measured in the detector, and the detector response to each of these particles. Both parts will be introduced in turn in the subsequent sections.

26.4 Modeling the LHC Processes

Consider a LHC collision of two partons $A,\ B$ producing a signal S, which decays into two particles $C,\ D$:

$$A + B \ \rightarrow\ S\ \rightarrow\ C + D \tag{26.5}$$

A characterizing parameter for the parton collisions is its hardness Q^2, which is given generically by $Q^2 = (p_C - p_A)^2$. Here, $p_i\ =\ (E_i, \mathbf{p}_i)$ are the four momenta of the particles, with E the energy and $\mathbf{p}_i$ the momentum components in the three spatial directions. “Soft collisions” have a Q^2 of a few GeV^2, hard collisions at the LHC are typically (100–1000 GeV)^2.

If the outgoing particles are electrons as in Eq. 26.2, they are rather easy to simulate. Reactions in which the final state particles C, D are strongly interacting partons are more complicated. Due to the properties of the strong interactions, these partons cannot be observed directly, but are “dressed” by many additional partons at distances smaller than 10^{-15} m. In the detector, hard partons appear as narrow cones of some 20–40 particles, deemed “jets”. A jet event recorded with the ATLAS detector is shown in Fig. 26.1. Understanding this dressing is important for inferring the initial parton collision from the measurement. It is modeled on the basis of the precisely probed theory of strong interactions (Quantum Chromodynamics, QCD) and simulated starting from the high Q^2 parton collision Eq. 26.5 to low Q^2, where QCD-inspired models of parton emissions are invoked.[5] These models are not

[5] Here the particle physicists’ notion of “model” is used, which refers to a theoretical description of a physics process that is not fully calculable from the well founded and established “theory” of the Standard Model.

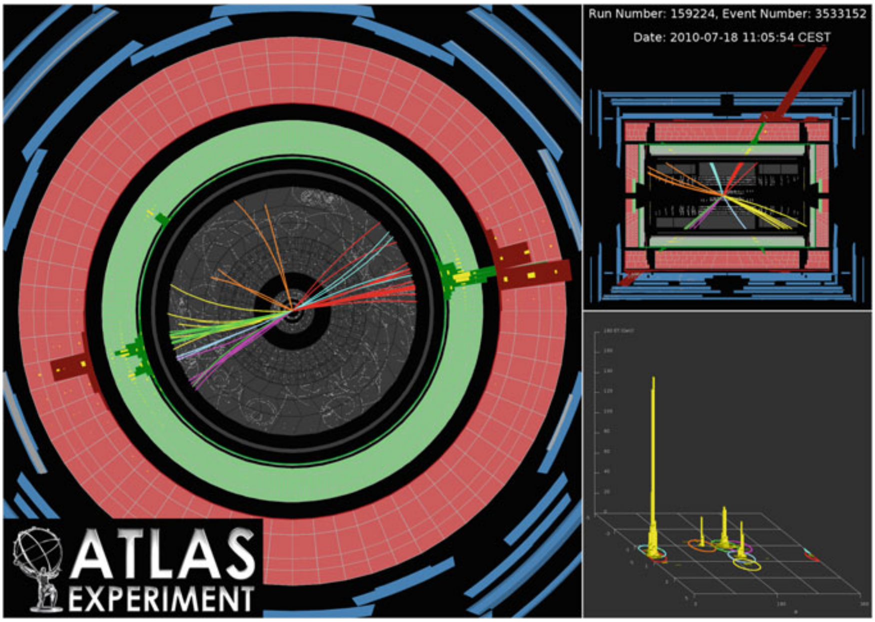

Run Number: 159224, Event Number: 3533152
Date: 2010-07-18 11:05:54 CEST
ATLAS
EXPERIMENT

Fig. 26.1 Two-jets event recorded at the LHC. The different components of the ATLAS detector are shown in the plane perpendicular to the beam direction (left), along the beam (upper right) and in the plane of polar and azimuthal angles (lower right). The tracks in the inner detector and the energy deposition in the calorimeters are indicated and show that energy is collimated in a rather small region, i.e., shows "jetty" behavior. (ATLAS Collaboration 2014b) (ATLAS Experiment © 2018 CERN)

unambiguous and several variants exist, all being cast into computer codes using Monte Carlo methods. We will now provide more details on the individual steps.

26.4.1 The Matrix Element of the Hard Collision

The fundamental differential equation for the process Eq. 26.5 is known, but cannot be solved in a closed form. Instead, it is perturbatively expanded in the strong coupling $\alpha_s(Q^2)$ as given in Eq. 26.1, leading to quantum mechanical matrix elements, which can be calculated exactly. Schematically this expansion can be written as (see e.g. Salam 2010)

$$\sigma(A+B \to S) \; = \; \sigma_0(A+B \to S) \; + \; \alpha_s \sigma_1(A+B \to S) \; + \; \alpha_s^2 \sigma_2(A+B \to S) \; + \; ..... \tag{26.6}$$

The higher the order in α_s, the more complex is the calculation, but since α_s is of the order of 0.1, higher order terms also become smaller. The complete perturbation series cannot be determined and is instead truncated, for LHC processes typically at the α_s^2 term. Consequently, small contributions to the full cross section are missing.

26.4.2 Parton Distribution Functions: Dressing the Initial State

Since the partons A, B in the initial state are subcomponents of protons they carry just a fraction x of the proton energy p.

x follows a probability distribution denoted as parton distribution function (pdf), and enters the prediction of a cross section, but cannot be calculated. The Q^2 dependence of the pdfs is, however, given by theory and has been experimentally well confirmed (see Sect. 26.8.1).

26.4.3 Dressing of the Outgoing Partons

To describe the dressing of the outgoing partons C, D into jets, QCD-based models are invoked, in which each final parton of the matrix element calculation is assumed to branch into further partons, which subsequently branch again.

The models follow each parton individually neglecting quantum mechanical interferences. Each of these branchings happens at some Q^2, which decreases with the order of the branching. Since $Q \sim 1/t$ (t denoting time), this parton showering can be interpreted as a time-ordered (Markov) chain. Modeling particle production in jets can be classified into three steps (for more details see Salam 2010; Seymour 2004).

1. The scattered partons C, D split into more partons, a so-called "parton shower". How the energy of a single parton is split among its daughters follows directly from theory.
2. These partons are finally turned into stable hadrons making up "jets". The details of this "hadronization" are condensed in "hadronization functions" expressing the probability of the energy and the kind of hadron to be produced. They are provided using previous measurements.
3. Whereas in the hard process just one of the many partons inside each proton interacts, the remnant ones also produce a spray of particles. This "underlying event" is measured using special LHC processes.

This means that jet evolution can be simulated with a Markov chain by describing individual steps with special probability distributions. A rather complicated particle structure emerges, as schematically depicted in Fig. 26.2, where the many and diverse parton branchings become apparent, motivating the use of Monte Carlo simulation.

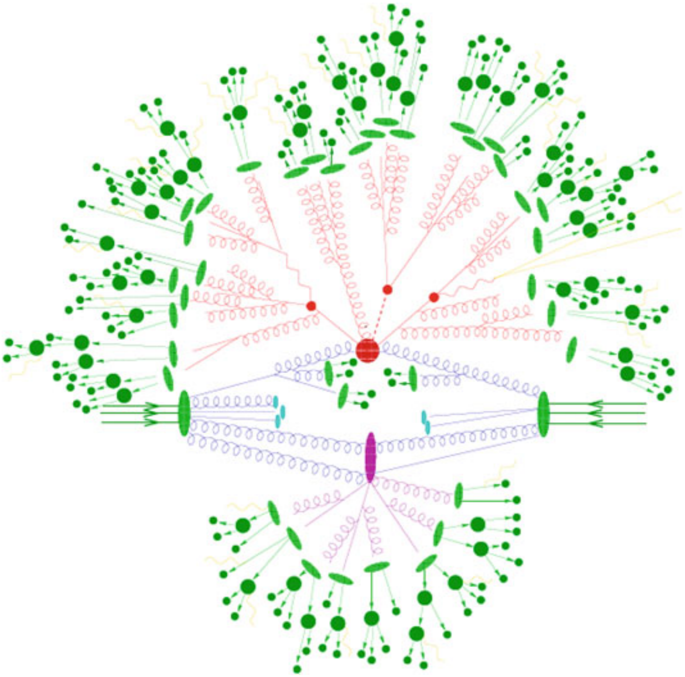


Fig. 26.2 Schematic LHC event: straight lines denote quarks, curly lines gluons. The initial partons make hard collisions (denoted by the large red circle) and split into other partons. Four immediately outgoing partons are shown, one (dashed line) decays only after flying a small distance. These partons branch further with increasingly smaller energy and angle resulting in jet-like structures. Finally, these partons turn into hadrons (shown as green lines/dots). In parallel, the other partons inside the protons also interact (elongated pink area) leading to hadrons. (Gleisberg et al. 2009)

Events at the LHC are further complicated by the occurence of hadrons from additional *pp* collisions in the same LHC bunches. These contribute a "pedestal" of particles to the hard collision of interest and are denoted as "pile-up" events. These can be determined from data (see Sect. 26.8.2).

26.5 Detector Simulation

The modeling of the physics process terminates with a list of stable particles: hadrons, photons, electrons, muons, and neutrinos. These are relevant for what is seen in the detector. Apart from neutrinos, all particles interact with the detector material and generate electronic signals. Roughly speaking, as becomes apparent in Fig. 26.3, each of the components of the ATLAS detector has a special sensitivity to one of these particles. Their reconstruction provides a picture of what has happened at the collision point.

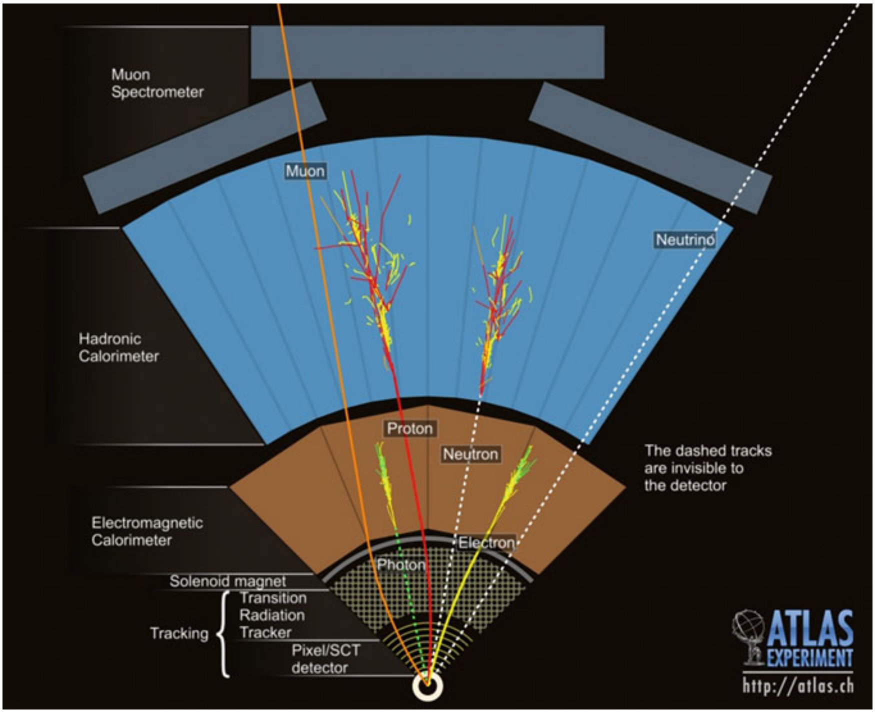


Fig. 26.3 Schematics of the interactions of various types of stable particles in the components of a typical LHC detector. By combining the information, the particle type can be identified (ATLAS Collaboration 2013) (ATLAS Experiment © 2018 CERN)

The principle of detector simulation is that each stable particle is traced inside the detector up to a volume element containing some material. They may interact according to probabilities obtained from models. The products of the interactions are then further traced to the next volume element with material and so on.[6] All materials that are part of the detector affects the passage of a particle and has to be considered, but only part is "active" meaning that it generates electronic signals that are used to reconstruct the event.

These interactions and generated signals are cast into computer codes (GEANT4 Collaboration 2003) that have been developed over the past 30–40 years and applied and validated in various, rather different detector environments. The main ingredients of these simulations are

- the geometry and materials of the detector,
- methods of numerical integration to follow a particle, partly inside a magnetic field,

[6] For a more detailed and also historical account of detector simulation in particle physics see (Daniel Elvira 2017).

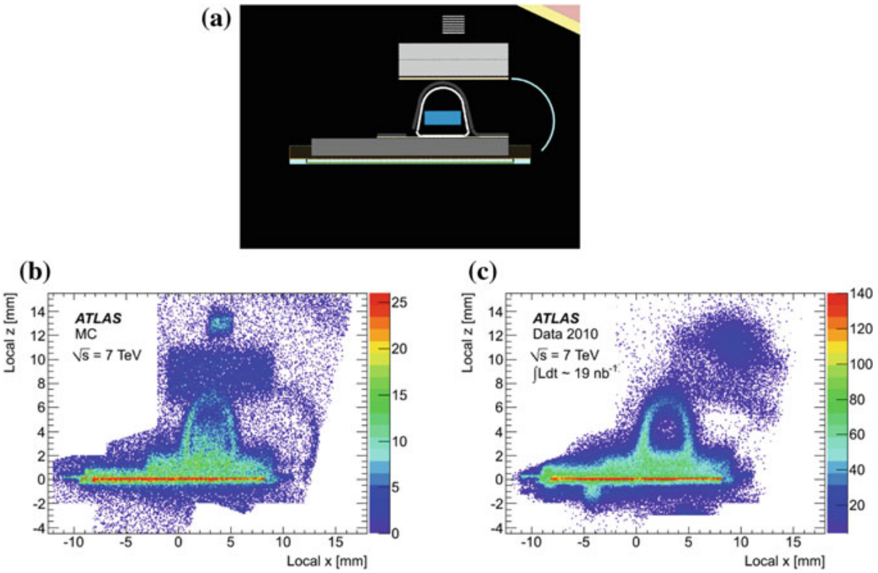


Fig. 26.4 Display of a component of the ATLAS pixel detector. The total length of the object is 2 cm, the widths of individual parts are as small as 0.02 mm. **a** The distribution of the material in the simulation (ATLAS Collaboration 2013), **b** the material in the simulation as seen in interactions of particles, **c** the material as seen in real collision events. Red indicates many interactions, blue relatively few (ATLAS Collaboration 2012a). (ATLAS Experiment © 2018 CERN)

- modeling the particle interactions in the material.

The detector geometry is mapped in a first step according to engineering drawings that have been used for building the detector. When relevant, one includes fine details, in part as small as several μm^3, as can be seen from Fig. 26.4(left) showing schematically one of the 80,000 modules of the pixel detector, the innermost component of ATLAS. The interactions with the material are simulated stochastically using the interaction cross section of the incoming particle h with the material A under consideration. i.e.,

$$\sigma_{\text{int}}(h + A) \; = \; \sigma_1(f_1) \; + \; \sigma_2(f_2) \; + \; \ldots\ldots \tag{26.7}$$

where the f_i are possible final states, which may consist of several particles. The respective momenta are generated according to models cast into computer codes, e.g., (Ford and Nelson 1978; Ferrari et al. 2005). Those for incoming electrons and photons are rather well understood, interactions of hadrons are more uncertain.

The interactions in the active volume will then be digitized, i.e., translated into an electronic signal.

While the simulation should be as precise as possible, it is not meaningful to exceed the measurement uncertainty, and also the precision should be balanced with the required computing time. Therefore, some effects are integrated over and condensed into a single parameter.

The simulation of the pixel detector is an example of the balance between fine details and required precision. There an electronic signal is caused by some 30,000 electron–hole pairs that are produced from the passage of a particle through a 250 μm thick silicon layer. It is well understood how these pairs are produced and move toward the electrodes, thus simulating these would be possible. However, such detailed simulation is only meaningful if supplemented with a simulation of each of the somewhat different 80,000 electronic circuits, requiring an excessive use of computing resources. Instead, the response of the pixel detector to particles is measured and the probability distribution of the signal depending on the particle's properties is used in the simulation without considering the details of its generation. No uncertainty is induced by this discretization.

As a side remark, one should be aware of the hugely different scales in detector simulation. Within a detector of a global diameter of 25 m, structures of the size of 0.00001 m are considered, if important. How finely the volume elements in the simulation are granulated, depends on their impact on the measurement and has been tested.

The simulated electronic signals in the various parts of the detector are then subjected to the identical procedure as the recorded data to reconstruct physics objects, i.e., electrons, muons, jets, etc.

26.6 Principles of Validation and Uncertainties

Formally, simulation in particle physics translates an original, "true" distribution T of partons $C,\ D$ with energy $E_{C,D}$, 3-momenta $\mathbf{p}_{C,D}$ and types $f_{C.D}$ into n reconstructed particles with respective energies, 3-momenta, and types, i.e.,

$$\begin{pmatrix} E_1 \\ \mathbf{p}_1 \\ f_1 \\ E_2 \\ \mathbf{p}_2 \\ f_2 \\ \cdots\cdot \\ \cdots\cdot \\ E_n \\ \mathbf{p}_n \\ f_n \end{pmatrix} = \mathscr{M} \times \begin{pmatrix} E_C \\ \mathbf{p}_C \\ f_C \\ E_D \\ \mathbf{p}_D \\ f_D \end{pmatrix} \tag{26.8}$$

where n is $\mathscr{O}(1000 - 100{,}000)$. The transformation $\mathscr{M}$ expresses what is happening in the simulation. Instead of listing all momenta of individual particles, the main idea can be illustrated in a simpler way by considering some distribution in a variable z of (physics) interest, e.g., the mass M of an event.

Along the discussion of Sect. 26.4.1, the matrix element calculation in some model yields the "true" mass distribution $T(M_{C,D})$. In the simulation, the $C,\ D$ are dressed and affected by experimental resolutions such that $T(M_{C,D})$ is transformed into a prediction $P(M)$ for the mass distribution that is supposed to be measured for the model under consideration. Actually measured in the experiment would be a distribution $D(M)$. To interpret the measurement, e.g., to infer on the validity of the model, $P(M)$ has to be compared to $D(M)$.

26.6.1 Factorization of Migration

In terms of z, simulation means

$$P(z) = \mathscr{M}T(z) \tag{26.9}$$

Discretizing z in intervals, the transformation can be expressed by a square matrix of the migration of a theoretical mass value to an observed one.

$$\mathscr{M} = \mathbf{m_{ij}} = \begin{pmatrix} m_{11}\ m_{12}\ .....\ m_{1n} \\ m_{21}\ m_{22}\ .....\ m_{2n} \\ .....\ .....\ ....\ ..... \\ m_{n1}\ m_{n2}\ .....\ m_{nn} \end{pmatrix} \tag{26.10}$$

where each element in the matrix relates a value of z in the incoming step to a z of the outgoing steps, e.g., a generated mass to the measured mass. E.g., m_{2n} is the probability that an event with a true z in bin n has a reconstructed z in bin 2. In the previous section it was discussed that simulation proceeds in (time-ordered) steps. Using the formalism each step corresponds to a special migration matrix $\mathscr{M}_\alpha$. Here, α denotes the effect under consideration, e.g., α = 1 the effect of the pdfs, α = 3 the one due to hadronization, α = k is the distortion from a detector component etc. The complete simulation $\mathscr{M}$ is then factorized into $\mathscr{M}_\alpha$,

$$\mathscr{M} = [\mathscr{M}_k \times ..... \times \mathscr{M}_2 \times \mathscr{M}_1] \tag{26.11}$$

$$P(z) = [\mathscr{M}_k \times ...... \times \mathscr{M}_2 \times \mathscr{M}_1] \times T(z) \tag{26.12}$$

More precisely, adding pile-up $U(z)$ and background processes $B(z)$ to the signal process $S(z)$, one arrives at modified (primed) results.

$$T(z) \rightarrow T'(z) = (S+U)(z) + (B+U)(z) \tag{26.13}$$

$$P'(z) = \mathscr{M} \cdot [(S+U)(z) + (B+U)(z)] \tag{26.14}$$

The physics question is whether $P'(z) = D(z)$, which means that S is correct, if $\mathscr{M}$, $U(z)$ and $B(z)$ are correct, but can also be fulfilled if, e.g., S and $\mathscr{M}$ are incorrect. Inferring S from D requires one to validate $\mathscr{M}$ and the background distributions.

26.6.2 Is Factorization Correct?

These considerations, which will be crucial for the validation procedure discussed in the next sections, depend on factorization. The latter holds true if the steps are incoherent and follow a time sequence. This is clearly true for the detector simulation, where a particle from the pp interaction first interacts in the tracking detector before entering the calorimeter etc. That is, only those particles have to be considered in the simulation of the calorimeter response that leave the tracking detector (cp. Fig. 26.3).

The time-sequence in the transition from the matrix element to the final hadrons in the physics generator, however, is only approximate since quantum level interferences are left out. However, the factorization assumption is justified from theoretical arguments (e.g., Salam 2010), at least for pdfs, to rather high precision. The basic argument is that the different steps occur at significantly different energy scales, which hardly influence each other. Similar arguments for the validity of factorization can be applied to the showering and hadronization effects. In addition, the assumption of factorization has been tested by comparing processes involving electroweak particles.

26.7 General Procedures of Validation in Particle Physics

At face value, data and simulation agree for SM processes at the LHC to stunning precision. In so far, the requirement in the philosophical literature, "validation .... is said to be the process of determining whether the chosen model is a good enough representation of the real-world system for the purpose of the simulation" (Winsberg 2015) seems to be fulfilled. However, as outlined above, the agreement is a necessary, but not a sufficient condition to claim the correctness of the model of the underlying parton process. Instead, the migration matrix $\mathscr{M}$ and the backgrounds have to be validated. The principle validation methods will be discussed in the next sections. The basic ideas are as follows.

1. For each $\mathscr{M}_\alpha$ the corresponding model is validated by either data from previous experiments, possibly together with a well founded theory, or better even, by an in situ experimental validation. For the validation of each $\mathscr{M}_\alpha$, specific, particularly sensitive processes are used, where this transformation can be isolated and therefore be tested. Examples will be given below. If all steps of the simulation are validated, obviously, the whole simulation is validated and the sufficient condition is fulfilled to infer from the agreement of simulation and data to the underlying

physics process. All steps mean that both the underlying physics model, which is a genuine part of the simulation, and the description of the detector response have to be validated

2. Given that the simulation should help to provide quantitative statements on SM parameters or BSM effects, validation should provide an estimate of how well a process is validated. This is expressed as a "systematic" uncertainty for each validation step, $\delta(\mathscr{M}_\alpha)$. In this sense, the process of validation is synonymous with the process of assigning an uncertainty. By error propagation, this leads to a total uncertainty $\delta(P'(x))$ of the simulation prediction (see Chap. 4 by Roy in this volume about errors and the treatment in general).

It is beyond the scope of this article to cover the complete validation procedure, which is documented in numerous publications and notes (some will be mentioned in the text). Instead, some examples will be presented in more detail, highlighting general methods used in the validation of both physics and detector modeling

- a combination of previous experiments and model assumptions,
- measurements of LHC processes that are complementary to the process of interest,
- in situ measurements of properties of the detector,
- adjusting the simulation to data using previous precision measurements.

As will hopefully become apparent in the following sections, the different "factors" that enter simulation at the LHC are experimentally tested and simulation is applied on a significantly constrained material basis. This discussion, although maybe sometimes technical, seems to be important in view of claims that "in the context of the LHC .... there is a lack of experimental data for comparison [with simulated data]" (Morrison 2015)(290).

26.8 Validation of the Physics Generators

The parton distributions in a space–time region smaller than 10^{-15} m and the model of how they turn into hadrons, have been discussed in Sect. 26.4. These processes are statistically distributed and thus a generic part of the computer simulation. Moreover, since the parton evolution is described by models, these have to be validated.

As examples, validation of pdfs and pile-up events will be discussed in the following sections. But before, a brief comment on the other steps is in order. As discussed, the perturbation series of the matrix element of the hard process are truncated. The magnitude of the remaining terms is estimated by a convention that has been tested in various processes such that uncertainties of typically some 3–5% are assigned. The basic understanding of the major other steps, i.e., showering and hadronization (see Sect. 26.4.3), has been obtained from previous experiments at e^+e^- colliders (Mättig 1989; Knowles and Lafferty 1997) and cross-checked with LHC data (see e.g., ATLAS Collaboration 2012b). The steps are described by various models, and their differing outcomes provide the uncertainty range assigned to showering effects.

26.8.1 pdfs

The pdfs are an essential ingredient of the simulation and instrumental for the interpretation of measurements. The cleanest way to measure pdfs is via the scattering of electrons on nucleons. A compilation of the major experiments is shown in Fig. 26.5 for various values of the parton energy fraction x as a function of Q^2. Whereas the pdfs themselves cannot be calculated from first principles, their Q^2 dependence is precisely known in QCD and excellently confirmed by measurements. Knowing the pdf at one Q^2 allows one to extrapolate it to another Q^2 (for more details see Campbell 2007).

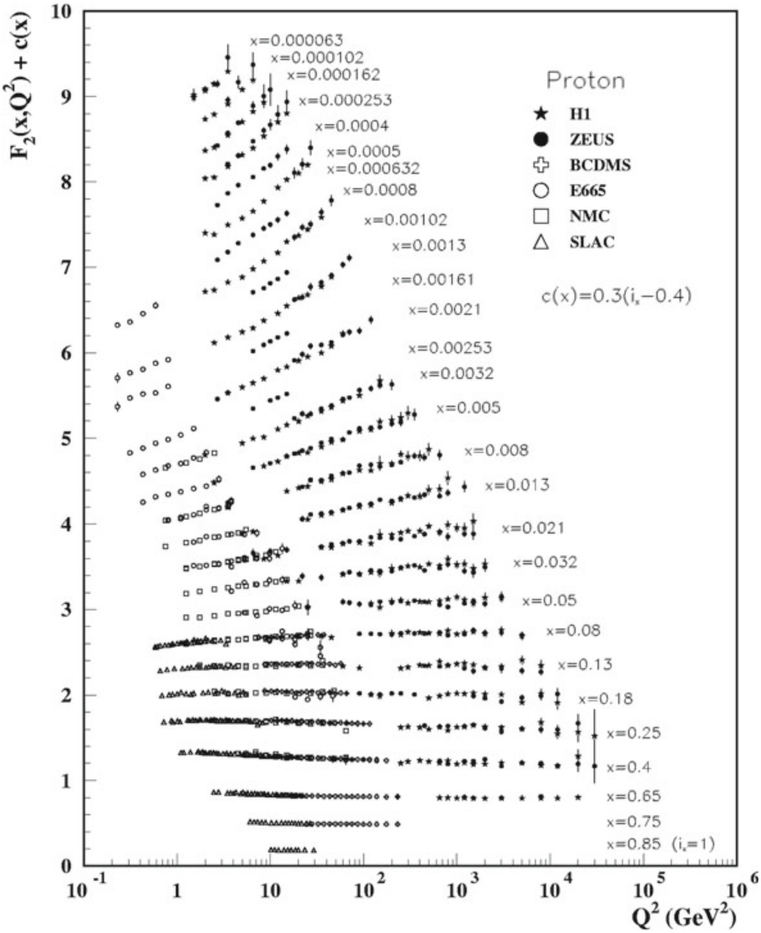


Fig. 26.5 Measurements on the Q^2 dependence of the pdf. Shown are measurements from several experiments and in different bins of x (Patrignani et al. 2016)

Such extrapolations are needed for the LHC, for which much higher Q^2 can be reached. Validation of the LHC has to address the question: what variations of pdf models are allowed by data at low Q^2? Given the knowledge of the Q^2 dependence how large are the variations at LHC conditions? Are there ways to test pdfs directly at the LHC avoiding circularity?

The uncertainties of pdfs at a certain Q^2 are due to both measurement uncertainties and theoretical ones. The latter ones come e.g., from inter- and extrapolating the binned measurements to a continuous distribution. Given the number of input bins there is in general only a small amount of variation allowed. However, at the extremes like $x = 0, 1$, measurements are less constraining and these variations can be important. These ambiguities lead to different pdf models with some variation in the expectations for the LHC, none of which can be a priori excluded.

As an example, four different pdf models for different kinds of parton species at a $Q = 100$ GeV (i.e., close to the relevant scale for the Higgs production) are shown in Fig. 26.6. As can be observed, the difference between different models is in general smaller than the uncertainty assumed for an individual model—an indication that the constraints from data are fairly tight. The variation between these pdf models

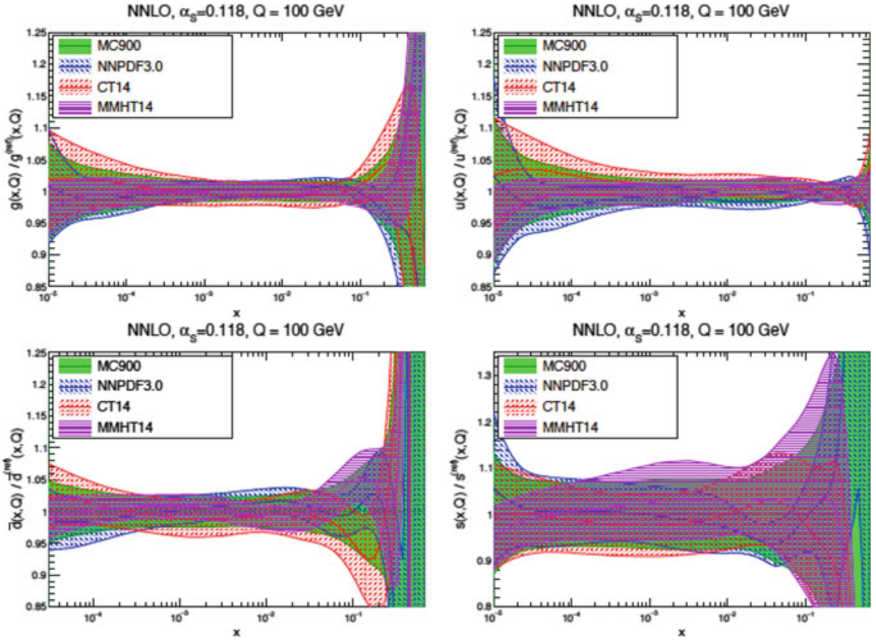


Fig. 26.6 Predictions for the pdfs of gluons (upper left), up quarks (upper right), anti-down quarks (lower left), and strange quarks (lower right) are shown for $Q = 100$ GeV and an x range stretching over 5 orders of magnitude. The different colors correspond to different pdf models, all are shown being normalized to the average of these. The coloured bands correspond to the uncertainties assigned to the corresponding model. (Butterworth et al. 2016)

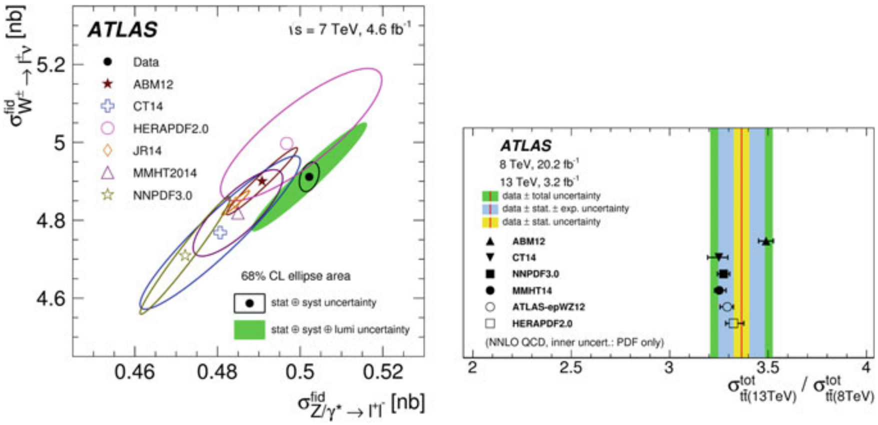


Fig. 26.7 ATLAS Measurement of $W^\pm$ versus Z^0 (left), being sensitive to other quark species. The data are compared to the expectation from various pdf models. The ellipses indicate the corresponding uncertainties, i.e., the 66% certainty range. (ATLAS Collaboration 2017a). The right figure compares production of pairs of top quarks at two different LHC energies (ATLAS Collaboration 2017c). (ATLAS Experiment © 2018 CERN)

together with their individual uncertainties are used as an overall pdf uncertainty of typically below 10%. A notable exception are very high x values.

In addition, these predictions can be validated by using special LHC processes. Particularly those involving W and Z^0 production provide in situ constraints on pdfs (ATLAS Collaboration 2017a). In addition, the energy dependence of the yields of theoretically well understood processes like the production of top quarks ($t\bar{t}$) can be used (ATLAS Collaboration 2017c). Both examples are depicted in Fig. 26.7. These cross checks show that the assumed pdf models derived from non-LHC experiments and their uncertainties are consistent with the LHC measurements.

26.8.2 Pile-up in pp Scattering

As stated in Sect. 26.4.3, every hard interaction of interest is accompanied by some 30 additional and incoherent pp interactions in one bunch crossing, called pile-up. This "pile-up" has to be an integral part of the simulation of a LHC event. Validation of pile-up simulation requires the properties of these events to be understood. This is achieved by direct LHC measurements (e.g., ATLAS Collaboration 2016a).

Since pile-up events have a high cross section at the LHC, they are frequently produced in bunch crossings, but can also be measured individually with short special LHC runs. Based on these measurements, QCD-inspired models are devised to describe pile-up events. These models are then used in the simulation to overlay pile-up to the physics processes of interest.

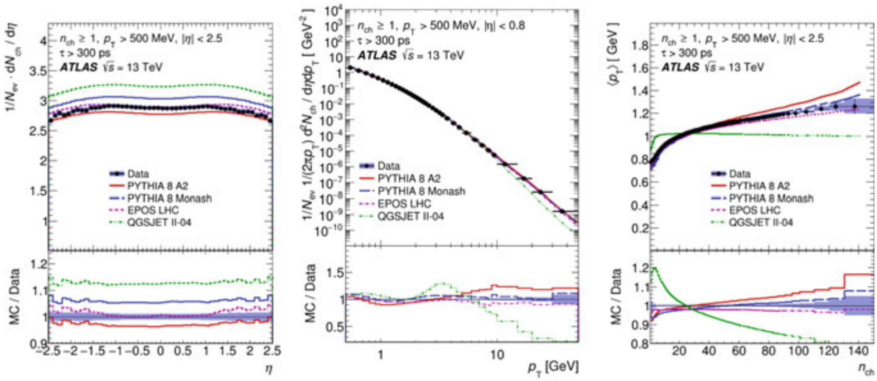


Fig. 26.8 Relevant properties of one pile-up event as measured in low luminosity runs. The production angle wrt. the beam direction (left) and the transverse momentum p_T distribution of charged particles are shown (center) and the average event p_T as a function of the number of charged particles in an event (right) (ATLAS Collaboration 2016a) (ATLAS Experiment © 2018 CERN)

Pile-up events are found to largely produce an isotropic spray of low energy particles as is apparent from the measurements shown in Fig. 26.8. In the Figure, data are compared to several models. As can be seen, some of them describe the measurements rather well and are used in simulation. The remaining deviations between the preferred models and data are small and do not affect the measurement in any relevant way.

Pile-up events can be seen as an example, where the large variety of LHC processes can be used to measure some process i directly as input to simulation for complementary physics processes j.

26.9 Validation of Detector Simulation

The modeling of the detector is crucial to understand how well particles can be reconstructed from electronic signals in the detector components (cp. Sect. 26.5). Here, we focus on detector geometry and simulation of electron properties.

26.9.1 Testing the Detector Geometry

While the starting point for the description of a detector are engineering drawings, once the detector is installed, details may change rendering the simulation inaccurate. As an example, consider a module of the pixel detector whose engineering drawing is shown in Fig. 26.4(left) and serves as a blueprint for the simulation. Data allow one to "see" the material distribution with a tomography-like method.

The method is based on the number and position of particle interactions inside the detector.[7] Their frequency is a measure of the material. For a pixel module the number of interactions in a projected area of $100 \times 100\,\mu\text{m}^3$ is shown in Fig. 26.4(low, left) (ATLAS Collaboration 2012a). The one-to-one correspondence with the input geometry clearly shows the method to work.[8]

The interesting step is to apply this tomography to data. The result is shown in Fig. 26.4(low, right). Whereas the general structure agrees well with the simulation, there are also important differences. E.g., the rectangular component in the simulation around z ~ 10 cm correspond to cables, which are much more spread out in the real detector, the circular shape around $x = 3$ cm is a cooling pipe. In the simulation, it is considered to be mostly filled with a liquid, in the data it becomes apparent that most of it is in gaseous state. Furthermore at $x = -4$ cm an electronic component (capacitor) is visible in the data, which had been omitted in the simulation. After these differences became known, the simulation has been adjusted.

The details of this discussion are not important, but the message is that the geometry of the detector can largely be checked and corrected with the data themselves.

A variety of procedures are used for different components. They are examples of in situ measurements based on well established types of reactions and redundancies in the detector.

26.9.2 Validation of Electron Simulation

A key particle for LHC physics is the electron, which is produced in many SM and perceived BSM processes and relatively easy to identify. Data analysis requires one to know the detection efficiency (cp. Eq. 26.4), the "true' energy E_{true} and the energy resolution σ_{E} as given by

$$E_{\text{meas}} = (1+\alpha)\,E_{\text{true}} \tag{26.15}$$

$$E_{\text{true}} \rightarrow f(E) \propto e^{(E-E_{\text{true}})^2/2\sigma_{\text{E}}^2} \tag{26.16}$$

where $1+\alpha$ is the electron energy scale indicating a possible mismeasurement, and σ_E is a measure of how strongly the measurements fluctuate. Evidently it is crucial for the simulation to describe these appropriately.

Simulations of interactions of electrons with material are based on a well-established theory, and have been studied numerous times and with many detectors, also in test-beams for the calorimeter components later installed in ATLAS. The

[7] These interactions in the detector are completely distinct from those *pp* interactions to test the SM and find BSM signals.

[8] This test of validation tools represents another use of simulation in particle physics, mentioned in Sect. 26.3.2.

electronic response in the calorimeter to the electron energy, however, needs to be calibrated. Given the precise knowledge of the Z^0 mass from previous measurements at the e^+e^- collider LEP (LEP 2006), see Eq. 26.1, and the abundant and clean Z^0 production at the LHC (see Fig. 26.9(left) ATLAS Collaborations 2016b), the electron energy is calibrated to reproduce the Z^0 mass. To take into account distortions of the shape of the Z^0 peak due to secondary interactions of the electron with the material in front of the calorimeter, this calibration is based on the Z^0 shape obtained from simulation.

Before this general procedure is realized, the simulation of the electron inside the calorimeter has to be validated in situ. E.g., inhomogeneous material distributions are obtained from the longitudinal evolution of a shower, and modifications of the calorimeter geometry due to gravitational effects are derived from angular modulations of the ratio of redundant measurements in the tracking detector and the calorimeter. The observed deviations are taken into account by parametric corrections of the simulation (ATLAS Collaboration 2014a).

Whereas the Z^0 mass has been measured and can serve as a reference, the energy resolution is detector specific. Still, the Z^0 serves as a marker for validation in that its observed width is a measure of σ_E. Differences between data and simulation are actually observed and the simulation is smeared to accomodate this discrepancy. The impact of this correction in the simulation can be seen in Fig. 26.9(right), where the ratio of data and simulation of m_{ee} is shown before and after the correction. Similarly, the electron efficiency is determined by in-situ measurements of the Z^0 decay. The simulation is adjusted to agree with the data (ATLAS Collaboration 2017b).

As an upshot, the simulation of electrons can be validated in situ, making use of the redundancy of the detector, the wide range of LHC processes and in particular a highly

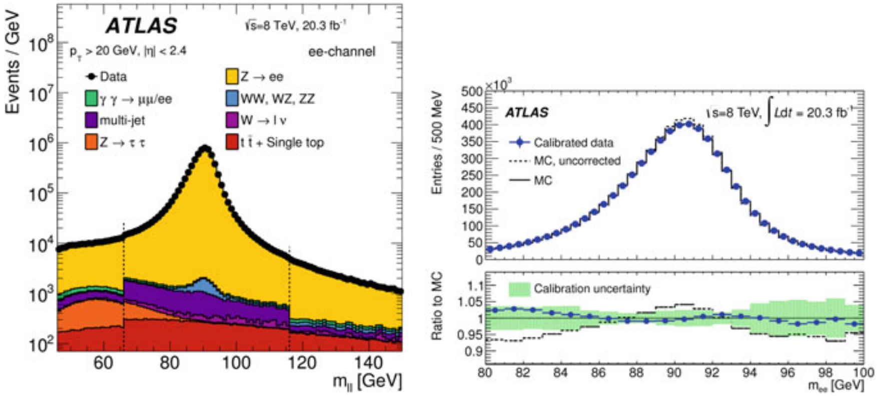


Fig. 26.9 The e^+e^- mass distribution observed with the ATLAS detector (left) (ATLAS Collaborations 2016b). Also shown are the background contributions. Comparison of data and simulation before and after adjusting the simulation to the Z^0 peak as a function of the e^+e^- mass (right) (ATLAS Collaboration 2014a). The ratio between data and simulation is shown in the lower part. (ATLAS Experiment © 2018 CERN)

precisely measured reference process. The high statistics and precise knowledge of the references also imply that the uncertainties assigned to simulation is small. Similar procedures also exist for other particles, e.g., (ATLAS Collaboration 2014c, 2015).

26.10 How Simulation is Applied in Data Analysis

In this section two examples will be discussed of how simulation is applied in data analyses. These will also show methods to validate the background simulation. In the first example properties of a known particle, the Higgs boson, are determined in a kinematic region that is well probed. The second example is a search for a BSM effect which requires extrapolations into as yet unexplored energy regions.

26.10.1 Measurement of the Higgs Cross Section

In July 2012 both the ATLAS and the CMS collaborations announced the observation of a new particle with a mass of 125 GeV (ATLAS Collaboration 2012b; CMS Collaboration 2012) through its decay into pairs of particles, especially into ZZ^* and $\gamma\gamma$. To establish if this is the long sought for Higgs boson, the properties of the particle had to be determined. One key measurement is the cross section (cp. Eq. 26.3) $\sigma_{h\to XX}$ for the Higgs production and its decay into two particles XX. Since the SM and alternative models for mass generation lead to different cross sections, their precise measurements can discriminate them. Here we will discuss $X = Z, Z^*$ along (ATLAS Collaboration 2018), with each Z decaying into a pair of electrons.

In Fig. 26.10 the observed the mass distribution of the four leptons (either electrons or muons) from the ZZ^* decays is shown. An enhancement around 125 GeV can be seen over an almost flat background. How this signal is translated into $\sigma_{h\to ZZ^*}$, will be outlined assuming a pure electron signal, although in the analysis both electrons and muons are used.

In a first step simulation is applied to find selections to reach a good signal to background ratio (S/B, cp. Sect. 26.3.1). For the analysis, events are selected that contain four electrons of which at least one has a minimum momentum p_T in the plane transverse to the beam direction of 20 GeV. All leptons should have an angle larger than 160 mrad wrt the beam axis. This defines the “fiducial volume”, in which the cross section is determined. Inspecting Eq. 26.3, one needs for the cross section measurement the efficiency and the background, i.e., the number of events that originate from other processes but also lead to four electrons.

The efficiency is affected by two major contributions. The first one is how many of the generated electrons are inside the fiducial volume. This is in a first step given by the matrix element, however, detector effects like the energy scale and resolution of electrons (see Sect. 26.9.2) have to be taken into account. The second is the detection

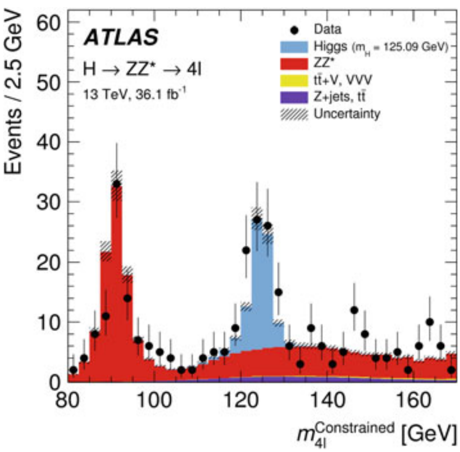


Fig. 26.10 Spectrum of the four lepton mass consistent with a decay ZZ^* as measured by the ATLAS experiment (ATLAS Collaboration 2018). The peak around 91 GeV is the Z^0 peak. (ATLAS Experiment © 2018 CERN)

efficiency of the electrons. Since the matrix element is known and the detector effects have been validated, the detection efficiency for $\sigma_{h\to XX}$ is obtained from Monte Carlo simulation by convoluting the physics model with the adjusted detector performance.

The dominant background source is continuum Z^0Z^* production, i.e., without an intermittent Higgs boson (see red area in Fig. 26.10). Its yield is tightly constrained by the measurement outside the signal region around 125 GeV. Minor backgrounds are, e.g.,due to top quark-pair production. These are estimated by a data driven method similar to the one that will be discussed in the next section.

Relative to the theoretical expectation (LHC Higgs Cross Section 2013) the measured cross section for a SM Higgs is

$$\left(\frac{\sigma(h \to 4l)_{\text{data}}}{\sigma(h \to 4l)_{\text{SM}}}\right)_{\text{fid}} = 1.28 \pm 0.18 \pm 0.07 \pm 0.07 \qquad (26.17)$$

where the uncertainties are statistical, experimental systematic and theoretical, the latter including pdfs. The experimental systematic uncertainty corresponds to the uncertainty of the validation from simulation.

26.10.2 Search for a Stop Quark

A large fraction of the analyses at the LHC tries to find BSM signals. In this section, it will be discussed, how simulation is applied in the search for supersymmetry, the most popular BSM model. The focus will be on methods to validate simulation of background processes.

More specifically, the search for the pair production of the supersymmetric partner of the top quark, denoted as stop ($\tilde{t}$), in the decay mode

$$pp \;\rightarrow\; \tilde{t}\bar{\tilde{t}} \;\rightarrow\; t\bar{t}\chi^0\chi^0 \tag{26.18}$$

will be considered, where χ^0 is a particle that leaves no trace in any detector—and therefore is also a dark matter candidate. The production and decay properties of this process are rather precisely predicted such that the distribution of the outgoing partons can be reliably simulated. Here we assume a stop quark of high mass (of 1 TeV) and a massless χ^0, following (ATLAS Collaboration 2016c).

The experimental signatures of this process are a detectable pair of SM top quarks ($t\bar{t}$), but in association with a very high imbalance of the detectable momentum in the plane transverse to the beam axis, caused by the two undetectable χ^0s. This is denoted as "missing transverse energy", $E_{\mathrm{T,miss}}$. One major background is due to two neutrinos from decays of top quarks,

$$pp \;\rightarrow\; t\bar{t} \;\rightarrow \nu\bar{\nu} + X \tag{26.19}$$

where X represents all other particles in the event. The validation of the modeling of this process will be discussed here. Simulations are used to suggest a selection that leads to an optimal S/B ratio for stop production. This signal region (SR) is defined through six observables, of which the most important are

- $m_T \; > 160$ GeV, the mass of the lepton p_T and $E_{\mathrm{T,miss}}$ system.
- am$_{T2} \; > 175$ GeV, a measure in how far the observed jets and leptons agree with coming from two stop quarks of mass 1 TeV.[9]

The retained events are in kinematic regions, which have not yet been probed and are prone to possible misrepresentations of both Standard Model physics and detector modeling. Using the formalisms of Sect. 26.6.1, one can separate the whole distribution for simplicity into regions that have, respectively have not, been probed, i.e., $z < z_{\mathrm{cut}}$ and $z > z_{\mathrm{cut}} = (z_{\mathrm{cut}} + \Delta)$. The z_{cut} may be identified with the selection requirement used to isolate stop pair—events and the region $(z_{\mathrm{cut}} + \Delta)$ with the SR. In this region simulation of the Standard Model distributions is given by

$$P'(z_{\mathrm{cut}} + \Delta) \;=\; [\mathscr{M} + \Delta\mathscr{M}] \cdot (B + U)(z_{\mathrm{cut}} + \Delta) \tag{26.21}$$

$\Delta\mathscr{M}$ represents the migration that has not been tested before, whereas $B(z_{\mathrm{cut}} + \Delta)$ the Standard Model distribution in the not yet tested region.

It follows from Eq. 26.21 that the contribution to $P(z > z_{\mathrm{cut}})$ is due to two sources: the Standard Model contribution $B(z > z_{\mathrm{cut}})$ and the migration of events out of $B(z < z_{\mathrm{cut}})$ (and of course both together). The simulation of the background yield needs to be validated in the region $(z_{\mathrm{cut}} + \Delta)$. This is done in two steps

[9] The exact definition is

$$\mathrm{am}_{T2} \;=\; \min_{\mathbf{q}_{Ta}+\mathbf{q}_{Tb} \;=\; E_{\mathrm{T,miss}}} [max(m_{Ta}, m_{Tb})] \tag{26.20}$$

I.e., the minimum parent mass consistent with the observed kinematic distributions assuming input masses m_{Ta} and , m_{Tb} and certain mass combinations.

- In a first step, a "control region" (CR) for the process 26.19 is defined, which uses the same observables as the SR, with cut values in general as close as possible to the SR, but inverting the am$_{T2}$ requirement. This enriches $t\bar{t}$ events, makes the SR and CR regions disjunct and the CR void of any signal. The normalized distributions (shapes) are sensitive to detector effects and data and simulation are compared for $z > z_{\text{cut}}$. As an example the m_T distribution is depicted in Fig. 26.11(left) showing a good agreement. It underlines that the physics distributions and the detector effects are well described also for $m_T > 160$ GeV, which is sensitive to a potential stop particle. Once the shape is confirmed, the simulated cross section for $z > z_{\text{cut}}$ is scaled by.[10]

$$r_{t\bar{t}} = \left(\frac{N_{\text{data}}(t\bar{t})}{N_{\text{simulation}}(t\bar{t})} \right)_{\text{CR}} = 1.01 \pm 0.15 \qquad (26.22)$$

  and therefore adjusted to the measurement.
- In a second step, a "validation region" (VR) is defined where the selection is chosen to be in between the CR and the SR. A small, but negligible, fraction of events might come from the signal. Using the adjusted cross section for the simulation of the background, it is tested in how far the observed number and shape of events agrees with the expectation. Fig. 26.11(right) shows that the data can be consistently described.

These studies validate the background distributions in the kinematic vicinity of the SR, but not in the SR itself. After having adjusted simulation to data in the CR and VR, one uses simulation to extrapolate. However, since these extrapolations depend on details of strong interactions (see Sect. 26.4.3), a range of QCD models is used to estimate its additional uncertainty taking into account the constraints from the CRs and VRs.

In conclusion, the background in the new region is estimated with methods which use simulation as guidance, but in the validation process they are adjusted to agree with the data. This procedure implies a significant uncertainty such that the search is sensitive only in regions where the S/B is high. In the stop analysis, the number of Standard Model background events in the SR is expected to be 3.8 ± 1.0, where the uncertainty is mostly due to the modeling uncertainties. The expected signal contribution from a stop would be 6. In the data 8 events are observed, i.e., more than expected from SM sources alone, but consistent with just a statistical fluctuation.

[10] Note that using this method, other backgrounds, like W+jets, show a visible discrepancy between simulation and data.

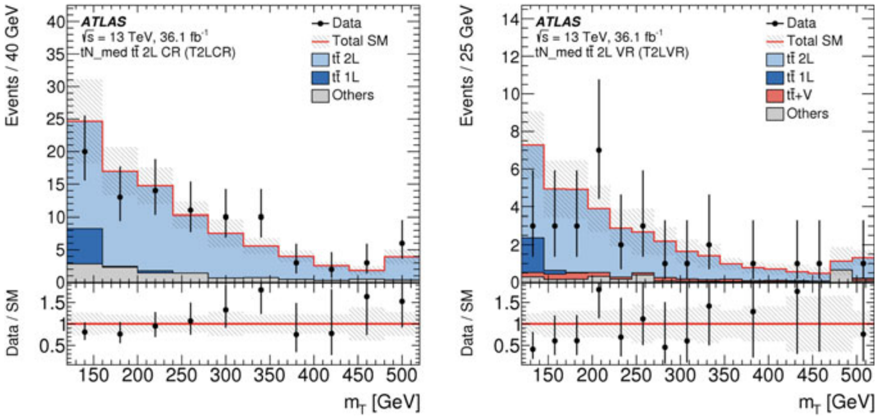


Fig. 26.11 The m_T distribution for the control region (CR, left) and the validation region (VR, right). Shown are the expectations from the dominant Standard Model contributions from $t\bar{t}$ production and the data. The lower part shows the ratio data/simulated events (ATLAS Collaboration 2016c). (ATLAS Experiment © 2018 CERN)

26.11 Discussion

Simulation in science and its validation are debated in the philosophical literature from various perspectives. In this short section a few points are commented from the view of particle physics without being able to address them in detail.

Simulation of LHC events is characterized by the remoteness of the underlying physics and the complexity of the measurement device. As a result it involves hugely different scales. To put it in striking terms, simulation covers the physics from distances of 10^{-18} m to those of several meters in the detector. For the different scales specific models are employed, such that instead of one model a chain of models is used in simulation. This is the basis of factorization addressed in this article.

Simulations allow one to take into account nonlinear effects and stochastic distributions and thus a more detailed modeling than analytical calculations. Simulations are therefore instrumental to improve the precision of the measurements and their interpretations. At least most of the fundamental techniques of data analysis using simulations are similar to those that have been invoked before. E.g., it was a standard method in analytical χ^2 minimizations to vary parameters in a model and see which one fits best. Simulations allow one to account for also subtle effects and many parameter variations with detailed templates. While such parameter variations motivate some philosophical literature to consider simulation as an "experiment" by itself (Winsberg 2015) they are no new type, no new quality of scientific practice (see Chap. 37 by Beisbart in this volume).[11]

[11] New potentials through simulations may have been opened for using machine learning techniques in data analysis.

The higher precision comes at the price of higher complexity which naturally reflects itself in the complexity of validation. Required is the validation of each model in the chain since incorrectness of one of the models implies the whole simulation to be incorrect. Since simulation is factorizable, the complexity of validation is broken down into the validation of many "simple" models. To the extent that the factorization is exhaustive and each factor is validated, the complete simulation is validated.

The models applied used have different confirmation statuses. Most of these are embedded in well established scientific practices that have grown and been confirmed over decades. Once such models have attained a very high level of confirmation, they acquire an almost autonomous status in simulation, i.e., their predictions are largely accepted. Other models are less certain, calling for more detailed scrutiny. E.g., simulation of the detector response for electrons is significantly better known than the emergence of jets from partons. Such models are less trusted and call for specific care in validation.

However, even if models are strongly confirmed, there is a reluctance among particle physicists to rely too strongly on these. For once all models assume certain parametrizations and parameter values and it may depend on special circumstances if these are applicable. For instance, even though the interactions of an electron in a material are well known, the distribution of the material may be not. Therefore, in situ validations or data driven methods are used to a large part. As discussed in this article, the individual model predictions at the LHC can be tested directly with complementary processes without any circular argumentation and are rather tightly constrained by the data. This appears to be in contrast to the claim of Morrison quoted in Sect. 26.7 above. In all these cases simulation is adjusted to describe the data—not vice versa, pointing to very different epistemic statuses of data and simulation.

Validation, however, has to account for the obvious fact that simulations cannot describe phenomena in all fine details. There is no validation "per se" but only a validation with some uncertainty—validation in particle physics has to be quantifiable, leading to a systematic uncertainty of the simulation, which is a systematic uncertainty of the understanding of the measurement. At the LHC validation of the simulation and assigning a systematic uncertainty is almost synonymous and most of the work in LHC's data analysis is devoted to estimating these uncertainties. For some processes at the LHC the systematic uncertainty is reduced to the 0.1% level, for some they are much higher.

26.12 Summary and Conclusion

Simulation in particle physics is performed in factorizable steps that can be interpreted as migrating a "true" distribution of an underlying parton process to the measurable one. This factorization allows one to validate the individual steps using specific measurements in which these steps are isolated and circularity with the target physics process is avoided.

In the validation process, several models of different confirmation status are used. Some have been found to agree with data in many different experiments in the past, others are relatively new. This confirmation status affects how these models are applied, respectively which uncertainties are assigned. In all cases, particle physicists take the pain to validate the models and their application using data, trying to minimize the reliance on simulation and its validation on models. One can distinguish the following methods:

- in situ calibration of detector and physics processes,
- adopting previous precision measurements as references,
- deriving models from previous measurements and applying them to LHC data.

These different methods rely on data and models at different levels, also implying different ways to estimate the uncertainties.

In conclusion, substantial effort in particle physics is devoted to the validation of event simulation. The validation is in all steps significantly constraint by data. Models as autonomous entities are only invoked if they have been strong confirmation from previous measurements. However, even these are checked in an experiment. A quantitative estimate of the validity of the simulation is based on the factorization of individual contributions. Highly efficient methods have been devised to estimate and minimize the uncertainties with a strong effort to constrain those from data—either directly or reducing model choices.

Acknowledgements I am grateful to Martin King and Michael Stöltzner and anonymous referees for valuable comments. I also profited highly from discussions with colleagues from the Research Group "Epistemology of the LHC" funded by the DFG under grant FOR 2063.

References

ALICE Collaboration. (2008). The ALICE experiment at the CERN LHC. *JINST*, *3*, S08002.

ATLAS Collaboration. (2008). The ATLAS experiment at the CERN Large Hadron Collider. *JINST*, *3*, S08003.

ATLAS Collaboration. (2012a). A study of the material in the ATLAS inner detector using secondary hadronic interactions. *JINST*, *7*, P01013. arXiv:1110.6191.

ATLAS Collaboration. (2012b). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, *716*, 1. arXiv:1207.7214.

ATLAS Collaboration. (2013). An computer generated image representing how ATLAS detects particles. https://cds.cern.ch/record/1505342.

ATLAS Collaboration. (2014a). Electron and photon energy calibration with the ATLAS detector using LHC run 1 data. *European Physical Journal C*, *74*, 3071. arXiv:1407.5063.

ATLAS Collaboration. (2014b). http://cds.cern.ch/record/1697048.

ATLAS Collaboration. (2014c). Muon reconstruction efficiency and momentum resolution of the ATLAS experiment in proton-proton collisions at $\sqrt{s} = 7$ TeV in 2010. *European Physical Journal C*, *74*(9), 3034. arXiv:1404.4562.

ATLAS Collaboration. (2015). Jet energy measurement and its systematic uncertainty in proton-proton collisions at $\sqrt{s} = 7$ TeV. *European Physical Journal C*, *75*, 17. arXiv:1406.0076.

ATLAS Collaboration. (2016a). *Charged-particle distributions in* $\sqrt{s} = 13$ TeV *pp interactions measured with the ATLAS detector at the LHC*. j.physletb.2016.04.050, arXiv:1602.01633

ATLAS Collaborations. (2016b). Measurement of the transverse momentum and ϕ^*_η distributions of Drell–Yan lepton pairs in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *European Physical Journal C*, *76*(5), 1–61. arXiv:1512.02192.

ATLAS Collaboration. (2016c). Search for top squarks in final states with one isolated lepton, jets, and missing transverse momentum using 36.1 fb^{-1} of $\sqrt{13}$ TeV pp collision data with the ATLAS detector, JHEP 06 (2018) 108. arXiv:1711.11520.

ATLAS Collaboration. (2017a). A Precision measurement and interpretation of inclusive W^+, W^- and Z/γ production cross sections with the ATLAS detector. *The European Physical Journal C*, *77*, 367. arXiv:1612.03016

ATLAS Collaboration. (2017b). Electron efficiency measurements with the ATLAS detector using 2012 LHC proton-proton collision data. *European Physical Journal C*, *77*, 195. arXiv:1612.01456.

ATLAS Collaboration. (2017c). Measurements of top-quark pair to Z-boson cross-section ratios at $\sqrt{s} = 13, 8, 7$ TeV with the ATLAS detector, *Journal of High Energy Physics, 02*, 117. arXiv:1612.03636

ATLAS Collaboration. (2018). Measurement of the Higgs boson coupling properties in the $H \rightarrow ZZ* \rightarrow 4l$ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector. *JHEP*, *03*, 095. arXiv: 1712.02304.

Butterworth, J., et al. (2016) PDF4LHC recommendations for LHC run II. *Journal of Physics G: Nuclear and Particle Physics*, *43*, 023001. arXiv:1510.03865.

Campbell, A. W., Huston, J. W., & Stirling, W. J. (2007). Hard interactions of Quarks and Gluons: A primer for LHC physics. *Reports on Progress in Physics*, *70*, 89. arXiv:hep-ph/0611148.

CMS Collaboration. (2008). The CMS experiment at the CERN LHC. *JINST*, *3*, S08004.

CMS Collaboration. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, *716*, 30. arXiv:1207.7235.

Daniel Elvira, V. (2017). Impact of detector simulation in particle physics collider experiments. *Physics Reports*, *695*, 1. arXiv:1706.04293.

Evans, L., Bryant, P. (Eds.). (2008). LHC machine. *JINST*, *3*, S08001.

Ferrari, A., Sala, P., Fasso, A., & Ranft, J. (2005). *FLUKA: A multi-particle transport code*, CERN-2005-10, INFN/TC 05/11, SLAC-R-773.

Ford, R., & Nelson, W. (1978). *The EGS Code System - Version 3*. Stanford Linear Accelerator Center Report SLAC-210.

GEANT4 Collaboration, Agostinelli, S., et al. (2003). Geant4 - a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *506*, 250–303.

Gelmini, G. (2015). TASI 2014 lectures: The hunt for dark matter. arXiv:1502.01320v1.

Gleisberg, T., et al. (2009) *Event generation with SHERPA 1.1. JHEP*, *02*, 007. arXiv:0811.4622.

Knowles, I. G., & Lafferty, G. D. (1997). Hadronization in Z^0 decay. *Journal of Physics G: Nuclear and Particle Physics*, *23*, 731. hep-ph/9705217.

LEP collaborations ALEPH and DELPHI and L3 and OPAL and SLD Collaborations and LEP Electroweak Working Group and SLD Electroweak Group and SLD Heavy Flavour Group. Schael, S., et al. (2006) Precision electroweak measurements on the Z resonance. *Physics Reports*, *427*, 257. hep-ex/0509008.

LHC Higgs Cross Section Working Group, Heinemeyer, S., Mariotti, C., Passarino, G., & Tanaka, R. (Eds.). (2013). Handbook of LHC higgs cross sections: 3. Higgs properties, CERN-2013-004 (CERN, Geneva). arXiv: 1307.1347 [hep-ph].

LHCb Collaboration. (2008). The LHCb experiment at the CERN LHC. *JINST*, *3*, S08005.

Mättig, P. (1989). The structure of jets in $^+e^-$ collisions. *Physical Report*, *177*, 141.

Morrison, M. (2015). *Reconstructing reality*. Oxford University Press.

Patrignani, C., et al. (Particle Data Group), Chin. Phys. C, 40, 100001 (2016) and 2017 update.

Pich, A. (2012). *The standard model of electroweak interactions*. CERN Yellow Report CERN-2012-001. arXiv:1201.0537.

Planck Collaboration (P.A.R. Ade et al.). (2016). Planck 2015 results. XIV. Dark energy and modified gravity. *Astronomy & Astrophysics*, *594*, A14. arXiv:1502.01590 (2015).

Quigg, C. (2013). *Gauge theories of the strong, weak, and electromagnetic interactions*. Princeton University Press.

Salam, G. P. (2010). *Elements of QCD for hadron colliders*. CERN Yellow Report CERN-2010-002 (pp. 45–100). arXiv:1011.5131.

Seymour, M. H. (2004) *Quantum ChromoDynamics*. Lectures given at the 2004 European School of High-Energy Physics, St. Feliu de Guixols, Barcelona, Spain, 30 May–12 June 2004 and at the 2009 Latin American School of High-Energy Physics, Recinto Quirama, Antioquia, Colombia, 15–28 March 2009 (pp. 97–143). CERN-2006-003 and CERN-2010-001 arXiv:hep-ph/0505192.

Winsberg, E. (2015). Computer Simulations in Science. *Stanford Encyclopedia of Philosophy*.

Chapter 27
Validation in Fluid Dynamics and Related Fields

Patrick J. Roache

Abstract A brief description of fluid dynamics is presented for a general audience. Within the context of fluid dynamics practice, the distinction is made between general (weak) models and specific (strong) models. Three historical options concerning a working definition of validation are briefly considered. Current validation practice in fluid dynamics is described based mostly on *ASME V&V 20-2009*, an ANSI Standard document. Model form uncertainty and other epistemic uncertainties, while sometimes important in model applications, are argued not to be important issues in validation. The weakest link in validation practice is claimed to be the reluctance, by both experimentalists and computationalists, to go beyond use of nominal set point data. This is clarified by the paradigm of experiments designed specifically for model validation. Coding features that facilitate model validation are described. Counter arguments are given to claims, based on extrapolation of the philosophy of falsificationism, that validation is impossible even in principle.

Keywords Validation · Verification · Model · Epistemic uncertainty · Falsificationism

27.1 Fluid Dynamics and Related Fields

Fluid dynamics is the subset of classical physics that describes flows of liquids and gases. It has a huge range of applications, including aerodynamics (calculating forces on airplanes), lubrication in bearings, flow rate through petroleum pipelines, weather forecasting, heat transfer, groundwater flow in radioactive waste disposal sites, plasma dynamics, combustion, ocean currents, blood flow in artificial hearts, microflagellate propulsion, etc. All of these applications require only *continuum* flow descriptions because there exists a well-founded separation of scales between molecular motion and the continuum flows for which aggregated quantities like velocity,

P. J. Roache (✉)
Consultant, 1215 Apache Dr., Socorro, NM 87801, USA
e-mail: hermosa@sdc.org

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_27

pressure, density, temperature are defined. The interaction effects of molecular-scale motion are aggregated into macroscopic semiempirical equations and their parameters, which are often remarkably simple. For example, viscous drag for low-speed aerodynamic flows is caused by the complex interactions of atmospheric gases (at least two principal ones) but the continuum result is a single number, the empirical viscosity coefficient, which simply multiplies the velocity gradients obtained from continuum solutions.

Classical fluid dynamics is based on conservation laws for mass, momentum (Newton's second law), and energy, expressed in integral or differential form. These are the Navier–Stokes equations (expanded), which famously do not have closed-form general (nondegenerate) solutions. Classical fluid dynamics has been expanded to include relativistic effects, multiphase phase flows (e.g., foams), chemically reacting flows, and others. The continuum equations considered herein are nonlinear partial differential equations (PDEs). These fluid dynamics equation are included within much more complicated models such as those of the next three chapters: astrophysics, weather forecasting, climate science.

The validation concepts described herein are applicable to all these disciplines that incorporate fluid dynamics into their formulations, and indeed to a much broader range of related non-fluid fields: any that utilize PDEs in their model formulation. (Note: *validation* in this chapter refers to *model* validation; this will be amplified in later sections.) These include physics problems such as heat transfer (via conduction and radiation as well as fluid dynamic convection), electric field calculations, optical scattering, solid mechanics, chemical reaction rate calculations, radioactive decay, plus non-physics disciplines such as biological systems (e.g., predator–prey models, disease vectors), and economics models. (This fluid dynamicist has consulted on a PDE model for forecasting bond prices.)

27.1.1 Weak Models, Strong Models, and RANS Turbulence Models

The Navier–Stokes equations provide the fluid dynamics continuum "model" of a general mathematical formulation, often termed a "weak" model. "Model" in a specific sense, often termed a "strong" model, includes all the parameter values, geometry, domain size, boundary and initial conditions needed to define a complete problem; only a strong model can be used in a computation. (Roache 1998a, b, 2009, Sect. 9.18.)

The classical Navier–Stokes equations for air at common conditions are based on unchallenged conservation laws and semiempirical relations; thus the weak model sense of Navier–Stokes is already validated. In principle, only particular "strong" models of fluid dynamics are yet to be validated. However, fluid dynamics exhibits chaotic solutions (and historically provided the impetus for modern chaos theory). For large values of the parameter called Reynolds number (see any fluids text) the flow pattern can change abruptly from laminar (smooth) flows to turbulence. In principle, the Navier–Stokes equations still apply, but the time and length scales of the solution structure change so radically that the discretization resolution requirements become almost insurmountable. Brute force Direct Numerical Simulation (DNS) of even simple turbulence problems are tremendously expensive. The great majority of practical turbulence simulations are performed not by using Navier–Stokes equations but Reynolds-Averaged Navier–Stokes (RANS) equations as the weak-sense PDE model. (See Wilcox 2006.) There are dozens of different RANS models, each with disappointingly limited range of applicability. For example, the k-ε model works passably for attached flows on airfoils as long as the streamwise pressure gradient is favorable (negative) but quickly fails when the gradient becomes positive. These RANS models need constant validations for new problems (geometries, etc.).

27.2 Separation of Verification and Validation

Broadly speaking, code verification involves verifying that the code faithfully executes the discretized PDE model and converges toward mathematically correct answers. It implies nothing regarding the model accuracy, i.e., the agreement between the model results and reality. Solution verification involves demonstrating that a computational solution to a particular problem has achieved approximately correct mathematical results. Again, this implies nothing regarding the model accuracy, which is assessed by validation when the model output is compared with reality.

The strong consensus opinion of specialists in verification and validation (V&V) for Computational Fluid Dynamics (CFD) and many related fields is that code verification and solution verification, though necessary forerunners of validation, are best kept separate from validation. (See Chaps. 3, 4, 11, and 12 in this volume.) All three topics are separate conceptually since both verifications involve only mathematics whereas validation involves science, via comparison of the mathematical results of the PDE model with observational science (reality). Furthermore, these three spheres of activity can be, and in practice often are, performed by different people in relation to the same computer code which embodies the model.

In this chapter, we assume that code verification and solution verification already have been performed at the nominal *set point* of the validation experiment (the specified conditions and parameter values such as fluid properties, geometry, flow velocity, boundary conditions, etc.)

27.3 Errors and Uncertainties

The solution verification at the nominal set point of the experiment has produced estimates of numerical *errors* and *uncertainties*. (See Chaps. 4 and 5.) These are related but distinct concepts. The distinction is not specific to computational models, but applies to measurement in any field, e.g., carpentry. If a piece of lumber is represented as having nominal length $L_{\text{nom}} = 2$ m and a precise measurement shows that the length L of one piece is actually $L = 1.997$ m, then the *error* δ (true value—nominal) is $\delta = (L - L_{\text{nom})} = -3$ mm. For a population of pieces, the 95% *uncertainty* of the true lengths might have been estimated to be $U_{95\%} = 8$ mm, which means that random samples taken from the population are expected to be within the range

$$L_{\text{nom}} - U_{95\%} \leq L \leq L_{\text{nom}} + U_{95\%}$$

or

$$1.992\text{ m} \leq \text{L} \leq 2.008\text{ m}$$

in roughly 95% of cases. It is also possible to have asymmetric uncertainties like

$$L_{\text{nom}} - 8\text{ mm} \leq L \leq L_{\text{nom}} + 5\text{ mm}$$

which might include one-sided (or *signed*) uncertainties like

$$L_{\text{nom}} - 8\text{ mm} \leq L \leq L_{\text{nom}}.$$

However, these are less common and are difficult to work with.

In spite of the clear distinction between error estimates and uncertainty estimates, it is not unusual for the concepts to be confused. This is understandable because they are related; a signed error estimate corresponds to a signed uncertainty estimate at the 50% coverage level or $U_{50\%}$ (Roache 2009, 2016).

27.4 Validation—What Does It Mean?

Issues involved in the meaning of the term *validation* in the fluid dynamics community were discussed in Roache (2004, 2008, 2009, Appendix B). We can *describe* (rather than *define*) validation broadly (legitimate, minimal validation) as the comparison of model results *and their associated uncertainties* with experimental (more inclusively, observational) results *and their associated uncertainties*. In this view, the term *validation* refers to the continuum model (PDE and associated boundary conditions, initial conditions, and parameter values). Strictly speaking, a model is validated and a code is verified, but loosely one speaks of a validated code meaning that the model embodied in the code has been validated. Of course, in order to validate a model its output must be calculated in a code, but the same validated model can then be embodied in different codes without a requirement for revalidation provided that the new codes are verified.

The descriptive versus legalistic definition approach has the advantage that it does not give a false impression of precision and universality. When legalistic definitions have been stated, at least four issues have arisen regarding (1) acceptability criteria (pass/fail), (2) necessity for experimental data, (3) intended use, and (4) prediction. Since each choice is binary, this gives $2^4 = 16$ possible interpretations of the legalistic definition, without even getting into arguments about what is meant by *model*, i.e., computational, conceptual, mathematical, strong, weak (ASME 2009; Roache 2009).

27.4.1 Issue #1. Acceptability (Pass/Fail) Criteria

The issue here is not what variable or acceptance level to use, but whether or not to include any pass/fail decision under the heading of *validation*. The actual practice is so disparate that it is necessary to examine any claim of validation in order to have confidence in the meaning intended. There are good arguments on each side. Initially people generally assume that validation indicates that the model has passed an accuracy requirement due to the correct recognition that pass/fail decisions must be made in any project. However, people quickly see the value of the alternative view. Although project-specific pass/fail criteria are certainly project requirements, those requirements do not necessarily need to be included in the term validation; the preferred terms are *accreditation* or *certification*. There are no universal criteria for acceptability even for a specific problem and metric; agreement with experiment within 10% may be adequate for one project while 1% is not good enough for another. Validation is best thought of as essentially *assessment* of model accuracy (or its obverse, model error) for a specific class of problems but probably applicable to multiple scientific and engineering problems; validation is a scientific activity. Certification or accreditation is best thought of as the next step, combining the accuracy assessment resulting from validation with project-specific acceptability (pass/fail) requirements and reaching a decision; it is not a scientific activity so much as an engineering management activity. The methodologies employed in each process have little in common.

However, there are dangers involved in this position. In some usage, a model whose results have been compared to experiments has been labeled *validated* regardless of the agreement achieved. In this loosest use of the term, *validated* then is not a quality of the code/model per se, but just refers to the QA (Quality Assurance) process. Carried to an extreme, this viewpoint gives the designation *validated* even to very poor models. It would be misleading to assign the inevitably value-laden term *validated* to a model that produces unarguably poor results just because it has gone through the QA process. A more moderate usage is to call the model *validated* but to state explicitly that the model is validated to a specified level and within the validation uncertainties. Also, some minimal accuracy should be required. The necessarily vague level of minimal accuracy must be determined by common practice, or state-of-the-art standards, in the discipline involved. Certainly, if a model cannot produce even qualitative trends, it is useless and does not deserve the dignity of the term *validated*.

Many discourage the use of the term *validated code* (or better, *validated model* within a code) because it might be misleading or even deliberately misused. However, it does not seem realistic to try to outlaw the past participle, and a code/model that has gone through a validation exercise will inevitably be referred to as a "validated code". Nevertheless, as Tsang (1991) noted, "almost by definition, one can never have a validated computer model without further qualifying phrases." The qualifications include knowledge of the experimental validation set points, the specific validation variables or metrics, whether or not pass/fail criteria are included, what is included in *model*, and of course the level of validation achieved, which requires stated uncertainties of both computations and experiments. Especially, the concept of a totally validated code or model, ascertained to be so independent of the metric, is a myth. (Roache 1998a, 2009, Sect. 9.19)

27.4.2 Issue #2. Necessity for Experimental Data

The resolution of this issue is clear. Many have said unequivocally that experimental (observational) data are the *sine qua non* of validation. *Experimental* is used here in a broad sense of *reality*, rather than limited to controlled laboratory measurements; it includes field measurements such as familiar weather parameters, ocean currents, astrophysical data.

$$No\,experimental\,data \;=>\; No\,validation$$

Many other factors remain, of course, as discussed in Roache (2008) and references therein. There have been some dissenters, whose apparent motivation is to try to gain the approval implicit in *validation* without the onerous requirement for obtaining real data. There are difficult problems, e.g., nuclear stockpile certification, for which further testing is outlawed. It is not always clear what these dissenters would substitute. Some look for agreement between different models. It is true that if one model has been previously validated, it can be regarded as a repository of experimental information, a set of second-hand experimental data plus smoothing and interpolation/extrapolation to parameter values other than experimental set points (Roache 2009). But in general, code-to-code comparison is not validation. The recommended view is uncompromising; no experimental data means no validation.

27.4.3 Issue #3. Intended Use

Whether or not validation requires a statement of intended use for the model might seem to be intimately related to Issue #1 (pass/fail criteria) but it can also be independent. Even though we may agree that pass/fail criteria are not necessary, it has been argued that validation requires a statement of intended use because the use will determine the relevant validation metrics. Indeed, intended use might dictate unusual validation metrics. But, for example, when Wilcox first evaluated (~validated) the k-ω RANS turbulence models for adverse pressure gradient flows in 1972 (Wilcox 2006) he did not need to have a specific intended use in mind. And even though he did have some uses in mind, a modeler need not have the same uses in mind decades later.

27.4.4 Issue #4. The Prediction Issue

A fourth issue only briefly treated in Roache (2008) is the overly literal interpretation of *prediction*. Validation as described here involves a comparison of the model with experiment. Of course, this means comparison of the *outcome* or *predictions* of the model with the results of the experiment. However, *prediction* has sometimes been taken in a literal temporal sense, i.e., the model results for validation should (or even must) be obtained before the experimental results. Although such temporal predictions are more persuasive, and avoid problems of post-experiment tuning of model parameters to improve agreement, we cannot take this issue seriously. If outright fraud is not an issue, there is no logical advantage to temporal precedence. The logical issue is simply comparison of model outcome with experimental outcome; this applies equally to theory as to computational models. See further discussions in Roache (1998a, b, 2004, 2009, Sect. 9.2.4).

27.5 Validation Methodology Based on *ASME V&V 20-2009*

A validation methodology for assessing model accuracy that includes numerical, experimental, and parametric uncertainties is given in ASME (2009), V&V 20-2009, which is an American National Standards Institute Standard document. As an ANSI Standard, it rather unequivocally qualifies as representative of accepted practice and arguably as “best practice”. The focus of V&V 20-2009 is on “unit problems” (Oberkampf and Roy 2010) which isolate one simple physical system (e.g. steady low-speed laminar flow over a sphere) rather than complex systems (e.g., unsteady turbulent two-phase flows in a maze of piping inside a nuclear reactor plant experiencing pipe ruptures).

27.5.1 ASME V&V 20-2009 *Background, Motivation, and Philosophy*

The objective of any validation exercise is to estimate the model error δ_{model} by comparison of a computational solution S (and its associated uncertainties) with experimental data D (and its associated uncertainties) for a specified metric at a specified validation point (or experimental *set point*) for cases in which the conditions of the actual experiment are simulated, expressed as the *experiment "as run."* For example, in computational aerodynamics, computational results are to be validated against wind tunnel results, without consideration of the applicability of the wind tunnel tests to free flight, which is a separate subject. The experiment "as run" is accepted as the reality of interest, so the conditions of the actual experiment are the *validation point* that is simulated. Usually a validation effort will cover a range of conditions within a domain of interest, which is highly recommended.

Implicit to the V&V 20-2009 development is a definition of *model* as a *continuum* conceptual model (e.g., the Navier–Stokes equations) evaluated at the experimental set points. Note that this is not a universally used concept; weather, climate, and ocean modelers typically use *model* to include the mesh, e.g., speaking of a "1/10 degree model of the Gulf of Mexico." Others may include discretization schemes in *model*.

For the brief description in this chapter, we use two convenient simplifications. V&V 20-2009 begins its development using standard uncertainty *u*. This leads to some complications. The methodology is more straightforward to describe and apply using "expanded" uncertainty estimates $U = U_{95\%.}$ This familiar concept of uncertainty is more useful and intuitive, indicating roughly 95% coverage, or roughly 20:1 odds of a sample from the parent population being inside the interval [estimate $\pm$ U]. Also, it is convenient herein to assume independence of experimental, parametric, and numerical uncertainties. This independence is often practical, notably if the validation metric is a directly measured in the experiment. Otherwise, the procedures require additional calculations that are straightforward conceptually (either chain rule numerical differentiation or Monte Carlo calculations) but are tedious to describe and carry out; see the detailed presentation in V&V 20-2009.

27.5.2 Validation Metrics

There are many possible metrics by which we might compare a model solution with physical data. The construction of validation metrics based on various functionals of the solution, often including weighted quadratures of all field points and variables, is an interesting and important area (Oberkampf and Roy 2010). Here we consider only the simplest type of metrics or "Quantities of Interest", e.g., maximum temperature, wing lift coefficient, net heat transfer rate, etc. Also, only independent single-point comparisons are considered; a forthcoming V&V 20-2009 supplement will address multi-point validations.

27.5.3 Defining Validation Uncertainty U_{val}

Consider a multiple-realization validation experiment for aerodynamic lift of an airfoil run at a single set point. We denote by T the True but unknown single value of lift that would be obtained from an ideal single experiment conducted at the same set point as the actual multiple-realization experiment. The values of lift produced by the simulation and the experiment are not single values. Rather, each covers an interval of lift values defined by a nominal center value and a ± uncertainty band; we denote the single nominal values by S and D. The continuum model errors δ_{model} are the lift values that would result from an exact solution of the continuum equations, minus the true (physical) value T.

The *nominal* validation comparison error E is defined as

$$\mathrm{E} = \mathrm{S} - \mathrm{D} \tag{27.1}$$

(This is what the naive analyst would take as the model error δ_{model} and, in the absence of other errors, this would be correct.) The error in the solution value S is the difference between S and the true value T.

$$\delta_{\mathrm{S}} = \mathrm{S} - \mathrm{T} \tag{27.2}$$

The error in the experimental value D is

$$\delta_{\mathrm{D}} = \mathrm{D} - \mathrm{T} \tag{27.3}$$

From these three equations, E is expressed as

$$E = S - D = (\mathrm{T} + \delta_{\mathrm{S}}) - (\mathrm{T} + \delta_{\mathrm{D}}) = \delta_{\mathrm{S}} - \delta_{\mathrm{D}} \tag{27.4}$$

E is thus the combination of all errors in the simulation result and the experimental result, and its sign and magnitude are known once the validation comparison is made.

All errors in the simulation solution S can be assigned to one of three categories (Coleman and Stern 1997).

- error δ_{MODEL} due to (continuum) modeling assumptions and approximations
- error δ_{NUM} due to the numerical solution of the equations (discretization error)
- error δ_{INPUT} due to errors in the simulation input parameters.

Thus

$$\delta_{\mathrm{S}} = \delta_{\mathrm{MODEL}} + \delta_{\mathrm{NUM}} + \delta_{\mathrm{INPUT}} \tag{27.5}$$

The objective of a validation exercise is to estimate δ_{MODEL} to within an uncertainty range. Combining the previous two equations gives

$$\delta_{\mathrm{MODEL}} = \mathrm{E} - (\delta_{\mathrm{NUM}} + \delta_{\mathrm{INPUT}} - \delta_{\mathrm{D}}) \tag{27.6}$$

Once the simulation and the experiment are run, S, D, and E are known. The errors δ_{NUM}, δ_{INPUT}, and δ_{D} are unknown but the corresponding uncertainties U_{NUM}, U_{INPUT}, and U_{D} would have been estimated. A key step of the V&V 20-2009 methodology is defining the *validation uncertainty* U_{VAL} as "an estimate of the uncertainty of the parent population of the combination of errors (δ_{NUM} + δ_{INPUT} - δ_{D})." (ASME 2009, p. 4.)

The estimation of U_{VAL} is thus at the core of the V&V 20-2009 methodology, and the methodology to estimate it will be described in the following Sect. 27.5.4. Once it is determined, the estimated U_{VAL} will be used (like any uncertainty estimate) as follows.

(E ± U_{VAL}) will characterize an interval within which δ_{MODEL} probably falls.

The working premise of the V&V 20-2009 project is that the net result of the model validation exercise would be a hand-off from people performing validation to the analysts who would be using the model, and that hand-off would essentially consist of E and $U_{\mathrm{VAL.}}$ In the simplest situations, E and U_{VAL} would consist of two real numbers, which would be either representative of, or upper bounds on, the range of E and U_{VAL} over all set points defining the *domain of validation* (flow parameters, geometries, etc.).

27.5.4 Estimating Validation Uncertainty

Total validation uncertainty U_{VAL} is a combination of U_{NUM}, U_{INPUT}, and U_{D}. The numerical uncertainty U_{NUM} is determined by various means, e.g. the classical Grid Convergence Index (GCI) or Least Squares GCI (Chaps. 4 and 11 in this volume; Roache 2009; Oberkampf and Roy 2010). The parameter uncertainty U_{INPUT} is determined from propagation through the code of parameter effects on S by either a sensitivity coefficient method or a Monte Carlo method; details and extensive examples are presented in Sect. 27.3 of V&V 20-2009. The experimental uncertainty U_{D} is determined using well-accepted techniques and is discussed in Sect. 27.4 of V&V 20-2009. A comprehensive and highly recommended end-to-end example of the application of the V&V 20-2009 methodology is presented and discussed in Sect. 27.7 of V&V 20-2009. For a recommended overview of sensitivity analysis and uncertainty propagation, see Blackwell and Dowding (2006).

If $\delta_{\mathrm{NUM,}}$ $\delta_{\mathrm{INPUT,}}$ and δ_{D} are effectively independent and their probability distribution functions (PDFs) are roughly Gaussian, then the corresponding uncertainties

$U_{\mathrm{NUM},}$ U_{INPUT} and U_{D} can be easily combined by the usual statistical assumption of root-sum-square (RSS) summation to calculate U_{VAL} as

$$U_{\mathrm{VAL}} = \surd\left[U_{\mathrm{NUM}}^2 + U_{\mathrm{INPUT}}^2 + U_{\mathrm{D}}^2\right] \tag{27.7}$$

For good discretization convergence work, the PDF of U_{NUM} is more like a shifted Gaussian. A more conservative summation is obtained by breaking out U_{NUM} from the RSS, as recommended in Roache (2016).

$$U_{\mathrm{VAL}} = U_{\mathrm{NUM}} + \surd\left[U_{\mathrm{INPUT}}^2 + U_{\mathrm{D}}^2\right] \tag{27.8}$$

There will be little difference between Eqs. (27.7) and (27.8) in case of either $U_{\mathrm{NUM}} >>$ or $<<$ the other two uncertainties. For all three terms equal, the ratio of Eqs. (27.8) to (27.7) is $(1 + \surd 2)/\surd 3$ giving a difference of 39% (corrected from Roache 2016). This minor modification is probably unnecessary if larger values of factors of safety are used in calculating U_{NUM} by the GCI. Also, note that U_{NUM} is not used in any part of the method until this final aggregation of uncertainties to calculate U_{VAL}, even if the component errors are not independent.

Fortunately, the condition of effective independence is often practical. In the important case in which the validation variable is directly measured, the assumption of effectively independent errors is generally reasonable. However, in the also common case in which the validation variable is determined using a data reduction equation (e.g., see ASME 2009), the experimental and computational values can be functions of shared variables, and δ_{iNPUT} and δ_{D} (and to a much lesser extent, δ_{NUM}) are not independent. Much of V&V 20-2009 (all of Sect. 27.5) is devoted to detailed examples of estimating validation uncertainty when the errors are not independent. The methods are conceptually simple but tedious, and the simpler use of Eqs. (27.7, 27.8) is expected to be more common. Also, the error of this simpler approach often will be conservative, since the dependent effects result in some double-counting which increases the estimate of validation uncertainty.

27.5.5 Interpretation of Validation Results and Caveats

The advantage of the V&V 20-2009 approach is evident when the interpretation of validation results is considered. ($\mathrm{E} \pm U_{\mathrm{VAL}}$) characterizes an interval within which δ_{MODEL} probably falls.

$$\delta_{\mathrm{MODEL}}\varepsilon\,[\mathrm{E} \pm U_{\mathrm{VAL}}] \tag{27.9a}$$

$$\mathrm{E} - U_{\mathrm{VAL}} \leq \delta_{\mathrm{MODEL}} \leq E + U_{\mathrm{VAL}} \tag{27.9b}$$

Case 1. If

$$|\mathrm{E}| >> U_{\mathrm{VAL}} \tag{27.10a}$$

then the analyst confidently estimates

$$\delta_{\mathrm{MODEL}} = \mathrm{E}. \tag{27.10b}$$

This is the most important case. $|\mathrm{E}| > > U_{\mathrm{VAL}}$ indicates a successful validation exercise since Eq. (27.10b) gives a reliable estimate of δ_{MODEL} and the object of the validation exercise was to evaluate $\delta_{\mathrm{MODEL.}}$ However, it does not indicate a good or acceptable model; no pass/fail criterion is used here, only assessment. (See Sect. 27.4.1.) E could be too large for any practical application, and a good validation exercise has resulted in a well-founded rejection of a poor model. (See Sect. 27.5.8.)

Fuzzy statements like Eq. (27.10a) are common in science an engineering, but for practical use in a testing protocol we require an algorithmic statement. A common interpretation of < < or > > is an order of magnitude, which usually morphs without protests into a specific cut-off as a factor of 10 ratio. Obviously, judgments will vary. A specific demarcation was proposed in Roache (2017). A factor of 10 for Eq. (27.10a) would be unnecessarily conservative for many problems. (Note there are some inherent conservative aspects of the methodology, such as the double-counting when $\delta_{\mathrm{NUM,}}$ δ_{INPUT} and δ_{D} are assumed to be independent, and the GCI use of ± symmetric uncertainty even though the error estimate from which it is calculated is one-sided.) The fuzzy cut-off Eq. (27.10a) is replaced by the following, which is judged to be adequate for most projects (perhaps with the exception high-risk decisions.)

Case 1A. If

$$|\mathrm{E}| \geq 7U_{\mathrm{VAL}} \text{then the analyst estimates } \delta_{\mathrm{MODEL}} = \mathrm{E} \text{ as in Eq. (10b).} \tag{27.11}$$

Case 2. If the validation exercise does not result in $|\mathrm{E}| > > U_{\mathrm{VAL}}$ then the validation exercise must be judged of poor quality. Equations (27.9a, 27.9b) still hold but does not provide a reliable estimate of δ_{MODEL}. However, it may still provide some information, although interpretation is prone to mistake and misunderstanding.

Consider a *very* poor validation exercise that results not in $|\mathrm{E}| > > U_{\mathrm{VAL}}$ but rather $U_{\mathrm{VAL}} > > |\mathrm{E}|$. The intended evaluation of δ_{MODEL} has been swamped by the numerical, parametric, and experimental uncertainties. For $\mathrm{E} = 0$, Eqs. (27.9a, 27.9b) would reduce to

$$|\delta_{\mathrm{MODEL}}| \leq U_{\mathrm{VAL}} \tag{27.12}$$

Although correct, this equation has multiple problems. It does not give any hint of the true sign of δ_{MODEL}. Further, as noted by Eça (2018), it suggests a functional relation between δ_{MODEL} and U_{VAL} when none exists. Nevertheless, it does provide information that we did not possess before we conducted the validation exercise: a bound on the magnitude of δ_{MODEL}. In this sense, the ITTC(2002) validation methodology for ship hydrodynamics claims validation at the U_{VAL} level for any validation exercise with $U_{\text{VAL}} > |\text{E}|$ (See also Stern et al. 2001 and Stern 2007). This would be literally true only for $U_{\text{VAL}} >> |\text{E}|$ but still misleading. "Validation at the U_{VAL} level" would seem to be an endorsement of the model and a desiderata of the validation exercise, yet it is easily achieved for any model by doing sloppy experimental and/or numerical work! See also Eça and Hoekstra (2009).

Case 3. Even more difficult to interpret generally is the vague intermediate case of $O(|\text{E}|) \approx O(U_{\text{VAL}})$. Equations (27.9a, 27.9b) still hold but provides only asymmetrical inequalities with little information. For example (Eça 2018), with $U_{\text{VAL}} = 30$ and E $= 20$, Eq. (27.9b) gives

$$-10 \leq \delta_{\text{MODEL}} \leq +50$$

which is close to useless. Without considering all possible cases of combinations of sign of E and relative magnitudes of |E| and U_{VAL} we can state the symmetric general bound on magnitude only is (Roache 2017)

$$|\delta_{\text{MODEL}}| \leq U_{\text{VAL}} + |\text{E}|. \tag{27.13}$$

27.5.6 Observations

As noted, Case 1 provides sharp estimates of both sign and magnitude of δ_{MODEL} whereas Case 2 provides only an unsigned estimated bound on $|\delta_{\text{MODEL}}|$, and far from a sharp bound. This cannot fairly be construed as a criticism of the V&V 20-2009 methodology per se. "Please do not shoot the messenger." This situation is no different from that of a physical measurement in which the accuracy or precision of the instrument is inadequate.

The importance of distinguishing these cases is evident when one considers not just the evaluation of a computational model but the possibility of improving the model. In Case 1 we have information that can possibly be used to improve the model, i.e., reduce the modeling error. In Case 2, however, the modeling error is within the "noise level" imposed by the numerical, input, and experimental uncertainties, so that formulating model improvements is more problematic. An analyst could hardly justify changing the model form, or even tuning parameters, without first reducing the "noise level" U_{VAL} by repeating the validation exercise to reduce uncertainties. [Also note that such tuning of parameters constitutes model *calibration*, which is not validation (Roache 2009; Oberkampf and Roy 2010).]

This interpretation of Case 2 is more evident with the methodology of V&V 20-2009 than with older approaches, and provides major advantages. In particular, it avoids a false-negative evaluation of the model when |E| is larger than some certification requirement for δ_{MODEL} but $U_{\text{VAL}} > |E|$. It alerts the analyst that a problem may exist with the experiment rather than with the model. The Third Lisbon V&V Workshop in 2009 provided a good example of such interpretation. The problem was the classic CFD computation of 2-D turbulent flow over a backstep using the Spalart–Allmaras model, for which there were six submissions at the Workshop. See Eça et al. (2009) for a complete evaluation, or Roache (2009, Sect. 11.10) for an excerpt.

27.5.7 Importance of Case 2

Case 2 is not the preferred outcome for a validation exercise. Preferably, $|E|$ would be small in an absolute sense, say 2%D, and U_{VAL} would be even smaller, say $U_{\text{VAL}} = 0.25\%$. This would allow a confident claim that $\delta_{\text{MODEL}} \sim E$. But the unfortunate fact is that many if not most "practical" problems in CFD are simulated using parameter ranges, computer resources, and personnel resources that lead to large U_{NUM} contributors to U_{VAL}. Likewise, economic constraints on validation experimentation lead to large U_{D} contributions to U_{VAL}. Even if the preferred Case 1 exists, it may lead to Case 2 later after model calibration. So recognition of Case 2 is important.

27.5.8 Model Quality Versus Validation Quality

It is easy to lose sight of a fundamental fact, related to the easy confusion of error and uncertainty. If U_{VAL} is unacceptably large, this says *nothing* about poor quality of the model. (To avoid semantic confusion it is essential that *model* here refers to the continuum model, not including the mesh.)

$$The\ magnitude\ of\ U_{\text{VAL}} does\ not\ reflect\ upon\ the\ quality\ of\ the\ model.$$

The magnitude of U_{VAL} increases because of poor computational work, poor parameter estimation, and poor experiments, not from a poor model. It does not depend on δ_{MODEL}. The model quality and the validation quality are different issues. The development of a model creates δ_{MODEL} while the performance of a validation exercise (including the execution of the experiment and the *use* of the model in the simulations) creates U_{VAL}.

A poor quality model combined with a high-quality validation exercise leads to $|E| >> U_{\text{VAL}}$ and therefore to trustworthy $\delta_{\text{MODEL}} \approx E$. If δ_{MODEL} is excessively large for any reasonable application, the result (certainly for certification, and arguably for validation) is a well-founded rejection of the *poor quality model* enabled by a *high-quality validation exercise*.

In the reverse situation, we could have a high-quality model, with δ_{MODEL} smaller than any foreseen application needs (or even a perfect model with $\delta_{\mathrm{MODEL}} = 0$), yet the validation exercise could fail because of excessive U_{VAL} (due to poor computational work, parameter estimation, and/or experiments). Fortunately, in the V&V 20-2009 methodology, this does not lead to a false-negative evaluation of model accuracy, but only to very useful information that well-founded conclusions about model quality cannot be made unless improvements are made, not in the model, but in the validation exercise itself. Thus, the only link between model quality and validation quality is that high-quality models can only be assessed with high-quality validations.

27.5.9 Forthcoming Addenda to V&V 20-2009

The ASME V&V 20 committee is presently working on addenda to V&V 20-2009. One topic is extending the domain of validation, or regression to an application point. Consider a new simulation S_A at *application point A*, at set-point parameters within a domain of validation other than validation points themselves. Estimation of accuracy for S_A involves interpolation of model error and uncertainties from validation points to application points. It also adds new U_{num} associated with $S_{A.}$For initial considerations, see Roache (2009, Sect. 11.12.)

Another topic involves extension to multiple set points. The original validation approach applies to a single validation set point. A planned supplement will utilize validation results from multiple set points (space, time, parameters) using a multivariate metric. Importantly, it accounts for correlations of errors at different set points. The multivariate metric will facilitate quantitative comparison of performances of different models.

Other supplement topics will include expanded discussion of interpretations and caveats (see Sects. 27.5.5, 27.5.6, 27.5.7 and 27.5.8).

27.6 Model Form Errors Versus Parameter Errors

Since the uncertainty contributions to U_{VAL} take into account all the error sources in δ_{NUM}, δ_{INPUT}, and δ_{D}, then δ_{MODEL} includes only errors arising from modeling assumptions and approximations; these are the *model form* errors. For example, in a simple heat conduction problem, deviation from the correct value of constant conductivity K would be part of the input parameter error δ_{INPUT}, while deviation from the assumption of constant K (i.e., neglect of dependence of $K(x,y,z,T,\ldots)$ and neglect of tensor versus scalar conductivity would be part of model form error. In practice, numerous gradations can exist in the choices of which error sources are accounted for in δ_{input} and which are defined as an inherent part of the model form error δ_{MODEL}. It includes errors in the governing *continuum* equations of the model and errors due to any other non-ordered approximations such as inflow and outflow boundary conditions (for strong form models); these errors do *not* → 0 as $\Delta \to 0$ (where Δ is a representative measure for the grid cell size). Errors resulting from finite distance to a far-field boundary are not included in δ_{NUM} because they are not ordered in Δ. Rather, they are included in δ_{MODEL}. Actually, there is nothing inherently "exact" about using free-stream conditions "at ∞" to model free flight when an exponential atmosphere model would be a better approximation to nature (Roache1998b).

The code used will often have more adjustable parameters or data inputs than the analyst may decide to use, especially for a commercial code. The decision of which parameters to include in the definition of the computational *simulation model* (conceptually separate from the *code*) is somewhat arbitrary. Some (even all) of the parameters available may be considered fixed for the simulation. For example, an analyst may decide to treat parameters in a chemistry package as fixed ("hard-wired") and therefore not to be considered in estimating U_{INPUT}, even though these parameters could have been accessed and had associated uncertainties. The point here is that the computational simulation being assessed consists of the code and a selected number of simulation inputs which are considered part of the simulation, while other simulation inputs have uncertainties that are accounted for in U_{INPUT} and thus do not contribute to δ_{MODEL}. If all parameter values are considered fixed in the model, this is the limit of what has been termed a strong-model approach. [See Sect. 2.2 of Roache (2009); see also Appendix C of V&V 20-2009 for related discussions.]

This distinction is required to explain the following paradox. As the analyst improves the thoroughness of a validation study by investigating parametric uncertainty more extensively, the total validation uncertainty will become larger, not smaller. Every additional parameter variation considered will add to $U_{\mathrm{INPUT.}}$ The resolution of the paradox lies in recognizing that, with every addition of another parameter uncertainty (e.g., considering variable conductivity instead of fixed K) one is changing the "model' under evaluation. In the limit of a strong model approach, with all parameter values hard-wired, there simply is no parametric uncertainty; $\delta_{\mathrm{INPUT}} = 0$ and $U_{\mathrm{INPUT}} = 0$.

27.7 Model Form Uncertainty and Probability Distribution Functions

The V&V 20-2009 methodology is implicitly based on the traditional concept of Probability Distribution Functions (PDFs) though not limited to Gaussian PDFs. In an unusual case where an important input parameter uncertainty cannot be characterized by some PDF (not even by the usually conservative uniform distribution PDF) then rigorous analysis may require interval-valued uncertainty. This can be treated by Probabilistic Bounds Analysis (PBA) or other methods; see e.g., Oberkampf and Roy (2010). Compared to PDF, PBA and others are expensive, difficult to understand,

and result in much larger net uncertainties. Interval-valued uncertainties are closely linked conceptually with model form uncertainties. They arise more commonly in *applications* of CFD, and are more likely for complex systems. However, recall that the focus of V&V 20-2009 is on *validation* (rather than application) of *unit problems* (rather than complex systems).

The point is, we know what model we are trying to validate. Parameter uncertainties in the model are accounted in the input uncertainty term. Essentially, model form uncertainty, epistemic uncertainty, PBA and related methods are negligible concerns for validation of unit problems (Roache 2016).

For a paper addressing CFD validation methodology for complex multilevel systems applied to industrial flare chemistry and emissions with multirange parameters and including disparate experimental databases, all using a PDF approach, see Jatale et al. (2017).

27.8 Weakest Link in Validation Practice

My perception of the weakest link in validation practice involves the reluctance, by both experimentalists and computationalists, to go beyond use of nominal set point data.

For fluid dynamics experimentalists, this covers the widespread disregard of measuring and documenting initial conditions for unsteady flows and complete boundary conditions, particularly inflow boundary conditions in wind tunnels. Most accusations against modelers of post-experiment tuning of model parameters to improve agreement are actually the fault of incomplete experimentation (Oberkampf and Trucano 2002). Although the importance should be obvious, the meticulous measurements and documentation require high-quality facilities and huge amount of work. This issue is what distinguishes a true validation experiment from old-fashioned engineering wind tunnel tests (see next section). The pioneering work of Aeschliman and Oberkampf (1997) remains an exemplar of such scrupulous work.

On the computational side, CFD analysts are often not willing to go beyond use of nominal parameters rather than doing the tedious work of using such data when it is available. Code builders often believe it should be sufficient to allow for nominal inflow properties, e.g., inflow velocity specified by a single number for axial flow component assumed constant across the inflow boundary. The needed computational practice requires code functionality that allows user input of boundary conditions that are functions of (x,y,z,t). This is the same feature that is needed for code verification by the Method of Manufactured Solutions; see Roache (2004) and Chap. 12 in this volume. Fortunately, this feature has become more common in widely used commercial codes.

The third component of blame in an industrial context is the failure of engineering management to demand and fund this level of work.

27.9 New Paradigm of Experiments Designed Specifically for Validation

The recognition of different requirements for CFD validation experiments, particularly in aerodynamics, can be said without exaggeration to constitute new paradigm in experimentation (e.g., Roache 2004; Oberkampf and Roy 2010, Sect. 10.1). In my opinion (Roache 2004), the most revolutionary concept in computational physics during my half-century career, other than simulation itself, has been the new paradigm of experiments designed specifically for validation. The new paradigm recognizes that requirements for validation are distinct and that validation experiments are much easier than traditional experiments in some respects but more demanding in others.

In aerodynamics, for example, the emphasis in precomputational days was on wind tunnel experiments, which attempted to replicate free-flight conditions. Great effort was expended on achieving near-uniform inflow, model fidelity, and minimizing wall and blockage effects. The latter requires small models, which sacrifice parameter fidelity (Reynolds number) and aggravate geometric fidelity.

The new paradigm approaches the problem differently, sacrificing some fidelity between the wind tunnel flow and free flight, but requiring that more nearly complete details of the experimental conditions and field data be obtained. No longer it is so important to achieve uniform inflow, but it is critically important to report in detail what those spatially varying inflow conditions are, so that they may be input to the simulation. The idea is that if the computational model is accurate for a flow perturbed from the free-flight conditions, it will probably be accurate for the free-flight condition. Thus blockage effects are not such major issues (and the tunnel wall itself may be modeled), so physical test models can be larger, thereby improving fidelity of Reynolds number and test model geometry. Alternately, wind tunnels can be smaller and therefore cheaper. Analogous situations occur in other experimental fields.

27.10 Unrealistic Expectations Placed on Experimentalists

The responsibility for matching boundary conditions clearly rests with the modeler. It is unrealistic, even arrogant, for a code builder or user to require an experimentalist to match idealized boundary conditions. Simple constant-value boundary conditions that are a mere convenience for the code builder can require major effort, cost, and time for an experimentalist. They often compromise other more desirable qualities of the experiment, and in fact may be literally impossible to achieve. A major contribution by the code builder to the synergistic cooperation between computationalists and experimentalists (which is also part of the new paradigm) is achieved by this relatively simple work of building the code with general boundary conditions.

27.11 Can Models be Validated? A Discussion of Falsificationism Versus Validation

It has sometimes been asserted, notably by Oreskes et al. (1994), that validation of computational models is impossible, even in principle, based upon Popper's widely known philosophy of *falsificationism* (Popper 1959). In *A Defense of Computational Physics* (Roache 2012) I presented counter arguments. (See also Roache 1998a, Appendix C). Various interpretations of Popper were paraphrased in a shorthand statement.

Popular Popper Précis

A scientific theory, and by extension a computational model, cannot be validated, i.e. proven to be true. It can only be invalidated or falsified, i.e. proven to be false. It must be capable (in principle) of being invalidated, i.e. be falsifiable, otherwise it is not a scientific theory but only a pseudo-scientific theory (or perhaps metaphysics).

To avoid the impression of setting up straw men, five examples were given of the use of Popper's assertions in modern and important works in *computational physics* (very broadly interpreted).

Popper was quoted as an authoritative witness to the fundamental impossibility of validation of computational physics models by a Blue Ribbon Panel on Simulation-Based Engineering Science (NSF 2006, pg. 34). Popper's falsificationism was foundational to the widely cited papers by Oreskes et al (1994) and Konikow and Bredehoeft (1992), the latter titled "Groundwater Models Cannot be Validated." In my own professional experience, these two papers have been taken seriously and have caused real problems. Oden et al. (2010) stated that "in line with Popper's principle, a model can never actually be validated." Hazelrigg (2003) invoked Popper to discredit validation in general engineering deign.

This importance in the computational physics modeling community is remarkable, considering that the applicable philosophical arguments appear in the first edition of Popper's most cited book, *The Logic of Scientific Discovery* (1959), the first German edition of which was published in 1934, well before the advent of modern computers and computational modeling. Whatever Popper's contributions or claims were, they were not directed specifically toward validation of computational physics models, but to scientific theories in general. (In fact, the word *validation* only appears in two footnotes.)

The approach in (Roache 2012) was to critique falsificationism at three levels: (a) philosophy of science, (b) empirical data on how science is actually conducted in the twenty-first century, and (c) applicability to computational physics modeling and the question of validation. A summary follows.

Categorical claims of impossibility of validation of computational models based on Popper's *falsificationism* are not justified and can themselves lead to ethical difficulties. Popper's demarcation criterion of *falsifiability* is a valuable concept (though not original), but is not without philosophical problems even when applied to scientific theories, as he intended. It is not adequate for an "if and only if" demarcation of science versus pseudo-science, as he had finally claimed. And in *Falsificationism*

Falsified, Hansson (2006) empirically demonstrated that actual scientific practice in 69 of 70 cases examined in 2000 did not follow the Popper prescription of how science is done.

When applied to validation of computational models, *falsificationism* as usually understood is inappropriate. However, Popper himself recognized a distinction that makes sense—that of "numerically exact" (finite in number) comparisons of theory and experiment which he did in fact recognize as "verifiable." For computational models, we would now say "validatable." The inappropriate application to modeling of the dictum that a scientific theory is never verifiable but only falsifiable is not only incorrect but can cause ethical violations, e.g., being used to categorically reject what possibly might be the best solution for nuclear waste disposal, or to categorically reject the possibility of usefully accurate climate modeling, etc. And the restriction (Oreskes et al. 1994) of this dictum to "natural systems" while accepting validation for manufactured systems is not such a sharp distinction, e.g., airplanes fly in a "natural" variable environment, etc.

Proponents of the impossibility of validation of computational models often have a rarefied view of validation that (a) has nothing to do with practical science and engineering and (b) is contradictory to widely accepted and pragmatic concepts of validation as used by most computational physics modelers. These include three graduate-level reference monographs, two ANSI Standards, and three publication policy statements of scientific journals (see Roache 2012). These have a more authoritative claim to defining semantic distinctions and setting normative practice for computational physics modeling than citations of Popper. When there are genuine fundamental difficulties with computational physics simulations, as in groundwater flow modeling, the difficulties will be related to obvious technical problems such as coarse mesh resolution, or lack of knowledge of physical properties, or inadequate accuracy of "laws" like Darcy flow, or uncertain model input such as rainfall, but nothing at all to do with Popper's *falsificationism*.

27.11.1 Truth *Versus* Accuracy

A most important distinction is easily made.

> Popper and his *falsificationism* are concerned with *Truth*, whereas validation of models is concerned simply with *accuracy*.

Does *validation* of a computational model imply *truth* of the model? The subject is discussed extensively in Sect. 2.6.4 of Roache (2012). According to Glanzberg (2006), *truth* is highly problematical to philosophers.

> Truth is one of the central subjects in philosophy. It is also one of the largest. Truth has been a topic of discussion in its own right for thousands of years. Moreover, a huge variety of issues in philosophy relate to truth, either by relying on theses about truth, or implying theses about truth. It would be impossible to survey all there is to say about truth in any coherent way.

Table 27.1 Characteristics of Popper's concept of *falsificationism* of scientific theories contrasted with the accepted and normative concept of *validation* of computational physics models

Popper's *falsificationism*	Computational *validation*
Involves science theories	Involves computational models
Concerned with *Truth*	Concerned with accuracy
Ambitious	Modest
Surprising predictions are valued	Surprising predictions are suspect
Goal of universality	Limited domain
Kuhn's "crisis science" or "revolutionary science"	Kuhn's "normal science“

The *truth* of a scientific theory was Popper's concern. However, in modern use of *validation* in conjunction with computational physics models, we are not concerned with some tortuously defined concept of *truth* but rather with the simple, well-defined concept of *accuracy*. The question for validation is this. Is the computational physics model accurate? This is an easy question, often with an unambiguous answer after a good validation exercise. Perhaps a dozen fluid dynamics RANS turbulent models (Wilcox 2006) have been validated to useful accuracy level for limited ranges of flow variables; none are, nor claim to be, true.

For a more provocative example of truth versus accuracy, consider the question of validation of Ptolemaic theory of geocentric motion of planets. This theory was used and trusted and validated, even in the pass/fail sense, for well over a millennium. Historically, its accuracy has been demonstrated arguably longer than any other scientific theory or model. Developed by Ptolemy around 140 C.E., it required 80 distinct circles; not what we would call an elegant model, but accurate. As a scientific theory, the Ptolemaic system was supplanted by the Copernican view of sun-centered rotations, demonstrated by Galileo in 1609 (published 1610). The Ptolemaic model is not *true*. However, the model users (with practical applications to local clocks, calendars, and navigation aids as well as mystical ones) continued to use it for another hundred years, and even past Newton, because it was accurate and easier to use for computations than the elliptical heliocentric orbits of Newton. (Chaisson and McMillan 2008). As a computational model, it was *accurate*, i.e., in our terminology, validated even in the pass/fail sense.

27.11.2 Summary of Falsificationism Versus Validation

Of course, *falsifiability* (as opposed to falsificationism, which allows "*falsifiability only*") is a tremendously important concept to science. However, Popper's philosophy of *falsificationism* (a) is not defensible philosophically, (b) is not normative of modern science practice, and (c) is neither applicable to modern computational physics modeling, nor endorsed by most of its practitioners. Table 27.1 summarizes some characteristics of *falsificationism* versus the concepts of computational model validation. Many of these characteristics involve widely known insights attributed to Kuhn (1962).

To avoid disputation and agonizing over what Popper or we may mean by *truth*, we might grant his statement that "every scientific statement must remain *tentative forever*" in some rarefied and hopefully harmless sense, but also note that validation of computational models is thereby positioned in the same category as Newton's laws of motion and gravity, Einstein's theories, entropy, Darwinian evolution, conservation of mass, Fourier heat conduction, etc. We computational modelers are in good, respectable company.

Acknowledgements I gratefully acknowledge help from C. Beisbart, L. Eça, and N. Saam.

References

Aeschliman, D. P., & Oberkampf, W. L. (1997). *Experimental methodology for computational fluid dynamics code validation, SAND95-1189* (p. 1997). Albuquerque, New Mexico: Sandia National Laboratories.

ASME. (2009). *ASME V&V 20-2009. Standard for verification and validation in computational fluid dynamics and heat transfer*, 2009.

Blackwell, B. F. & Dowding, K. J. (2006). Sensitivity analysis and uncertainty propagation of computational models. In W. J. Minkowycz, E. M. Sparrow & J. Y. Murthy (Eds.), *Handbook of numerical heat transfer* (2nd ed., pp. 443–469). New York: Wiley.

Coleman, H. W. & Stern, F. (1997). Uncertainties in CFD code validation. *ASME Journal of Fluids Engineering*, *119*, 795–803.

Chaisson, E., & McMillan, St. (2008). *Astronomy today* (6th ed.). San Francisco: Pearson Addison Wesley.

Eça, L., & Hoekstra, M. (2009). On the numerical accuracy of the prediction of resistance coefficients in ship stern flow calculations. *Journal of Maritime Science and Technology*. https://doi.org/10.1007/s00773-008-0003-8.

Eça, L., Hoekstra, M., Roache, P. J. & Coleman, H. (2009). *Code verification, solution verification and validation: An overview of the 3rd Lisbon* (Workshop, AIAA Paper No. 2009-3647). 19th AIAA Computational Fluid Dynamics, San Antonio, Texas, June 2009.

Eça. L. (2018). Personal communication 5/25/2018.

Glanzberg, M. (2006). Truth. In *Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/truth/ (June 13, 2006).

Hazelrigg, G. A. (2003). Thoughts on model validation for engineering design, DETC2003/DTM-48632. In *Proceedings of ASME 2003 Design Engineering Technical Conferences and Computers and Information in Engineering Conferences*, Chicago, IL, U.S.A., 2–6 Sept 2003.

Hansson, S. O. (2006). Falsificationism falsified. In *Foundations of science* (Vol. 11, pp. 275–286). Springer. https://doi.org/10.1007/s10699-004-5922-1.

ITTC. (2002). CFD general uncertainty analysis in CFD verification and validation methodology and procedures. In Quality *Manual, International Towing Tank Conference, Effective Date 2002*, Revision 01.

Jatale, A., Smith, P. J., Thornock, J. N., Smith, S. T., Spinti, J. P., & Hradisky, M. (2017). Multiscale validation and uncertainty quantification for problems with sparse data. *Journal of Verification, Validation and Uncertainty, 2*(1), Paper No: VVUQ-16-1022; https://doi.org/10.1115/1.4035864.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*, (2nd ed., enlarged, 1970; 3rd ed., 1996). University of Chicago Press, Chicago, Illinois.

Konikow, L. F., & Bredehoeft, J. D. (1992). Groundwater models cannot be validated. *Advances in Water Resources, 15*(1992), 75–83.

NSF. (2006). *Simulation-based engineering science: revolutionizing engineering science through simulation*. Report of the NSF Blue Ribbon Panel on Simulation-Based Engineering Science.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*(4), 641–646.

Oberkampf, W. L. & Trucano, T. G. (2002). Verification and validation in computational fluid dynamics. *AIAA Progress in Aerospace Sciences*.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge, UK: Cambridge University Press.

Oden, T., Moser, R., & Ghattas, O. (2010). Computer predictions with quantified uncertainty, Part I. In *SIAM News* (Vol. 43, Number 9, November 2010). (Part II Number 10, December 2010).

Popper, K. (1959). *The logic of scientific discovery*, (translation of *Logik der Forschung, first publication 1934*), Hutchinson, London, UK. (last copyright 1980. Routledge version 2006).

Roache, P. J. (1998a). *Verification and validation in computational science and engineering, Appendix C*. Albuquerque, New Mexico: Hermosa Publishers.

Roache, P. J. (1998b). *Fundamentals of computational fluid dynamics*. Albuquerque, New Mexico: Hermosa Publishers.

Roache, P. J. (2004). Building PDE Codes to be verifiable and validatable. In *Computing in science and engineering* (pp. 30–38). (Special Issue on Verification and Validation, September/October 2004).

Roache, P. J. (2008). Perspective: Validation-What does it Mean? *ASME Journal of Fluids Engineering*, *131*(3), CID 034503. (Also, *ASME Journal of Fluids Engineering* March 2009).

Roache, P. J. (2009). *Fundamentals of verification and validation*, Hermosa Publishers, Albuquerque, New Mexico. (Chapter 3 and Appendix C).

Roache, P. J. (2012). *A defense of computational physics*. Albuquerque, New Mexico: Hermosa Publishers.

Roache, P. J. (2016). Verification and validation in fluids engineering: some current issues. *ASME Journal of Fluids Engineering,* FE-16-1206. https://doi.org/10.1115/1.4033979.

Roache, P. J. (2017, June). Interpretation of validation results following ASME V&V 20-2009. *ASME Journal of Verification, Validation and Uncertainty, 2*, 024501-1–4.

Stern, F., Wilson, R. V., Coleman, H. W., & Paterson, E. G. (2001). Comprehensive approach to verification and validation of CFD simulations-Part 1: methodology and procedures. *ASME Journal of Fluids Engineering, 123,* 793–802.

Stern, F. (2007). Quantitative V&V of CFD solutions and certification of CFD Codes with examples for ship hydrodynamics. In *Symposium on Computational Uncertainty, AVT-147*, December 2007, Athens, Greece.

Tsang, C.-F. (1991). The modeling process and model validation. *Ground Water, 29*(6), 825–831.

Wilcox, D. C. (2006). *Turbulence modeling for CFD*. La Canada, California: DCW Industries.

Chapter 28
Astrophysical Validation

Alan C. Calder and Dean M. Townsley

Abstract We present examples of validating components of an astrophysical simulation code. Problems of stellar astrophysics are multidimensional and involve physics acting on large ranges of length and time scales that are impossible to include in macroscopic models on present computational resources. Simulating these events thus necessitates the development of sub-grid-scale models and the capability to postprocess simulations with higher fidelity methods. We present an overview of the problem of validating astrophysical models and simulations illustrated with two examples. First, we present a study aimed at validating hydrodynamics with high energy density laboratory experiments probing shocks and fluid instabilities. Second, we present an effort at validating code modules for use in both macroscopic simulations of astrophysical events and for postprocessing Lagrangian tracer particles to calculate detailed abundances from thermonuclear reactions occurring during an event.

Keywords Astrophysics · Supernovae · Nucleosynthesis · Fluid instabilities

28.1 Introduction

Verification and validation (V&V) of models and simulations of astrophysical phenomena present challenges because the problem of studying these phenomena is largely one of indirectly observing multi-scale, multi-physics events. Other aspects of astrophysics also contribute to challenges. The enormous length scales of astrophysical objects and vast distances to most astrophysical events preclude ready experimental access, limiting the availability of validation data. As with a great

A. C. Calder (✉)
Stony Brook University, Stony Brook, NY 11794-3800, USA
e-mail: alan.calder@stonybrook.edu

D. M. Townsley
University of Alabama, Tuscaloosa, AL 35487-0324, USA
e-mail: Dean.M.Townsley@ua.edu

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_28

many applications, models suffer from epistemic uncertainty in the underlying basic physics (e.g., turbulence, fluid instabilities, and nuclear reaction rates), which is difficult to control and assess in simulations incorporating multiple interacting physical processes. The large range of length and time scales in many astrophysical problems frequently necessitates capturing sub-grid-scale physics within simulations, relevant examples being thermonuclear flames and turbulent combustion. The requirement of the development of sub-grid-scale models for these physical processes obviously introduces an additional level of complexity to V&V. Finally, the magnitude of the requisite computations for astrophysical events means that even with sub-grid-scale models, simulations may only capture the bulk effect of the underlying physics and some properties such as detailed compositions must be obtained by postprocessing the simulation results with augmenting, higher fidelity routines.

Even with these issues, V&V are vital parts of computational astrophysics as with any research domain. We present two studies aimed at validating components of Flash, a freely available, parallel, adaptive mesh simulation code used for modeling astrophysical phenomena and other applications. We first present a study of validating the hydrodynamics routines in Flash with experiments designed to replicate the high energy density environments of astrophysics and probe the underlying physics. The investigation formally addresses the issues of concern in validating hydrodynamics and serves as a well-controlled case study. The second study we present addresses physics that is difficult to include in whole-star simulations, due to limits in computing power, but that can be incorporated with approximate models and also calculated by postprocessing simulation results. The problem is thermonuclear combustion and describing the overall reactions while including minimal nuclear species, and this work addresses the issue of comparing prohibitively expensive detailed models and simpler models that allow three-dimensional simulations.

As we will describe below, the challenges to astrophysical validation made parts of our study incomplete. The effort, however, was rewarding and very much worth the investment. Verification tests quantified the accuracy of code modules for problems with an analytic or accepted result, and the regular application of these tests serves for regression testing as the code is developed. Validation tests, though incomplete, demonstrated reasonable agreement between experiment and simulation for the case of the hydrodynamics study. Comparison between models of increasing sophistication allowed us to quantify the trade-off between fidelity of the method and expense. These studies all led to a deeper understanding of the underlying physics, and while we cannot say the modules and code were completely "validated," the process greatly increased our confidence in the results.

28.2 Approach to Verification and Validation

Our methods for V&V largely follow accepted practices from the fluid dynamics community (AIAA 1998; Roache 1998a, b; Oberkampf and Roy 2010, see Chap. 27 by Roache in this volume). We adopt the following definitions (based on definitions from the American Institute of Aeronautics and Astronautics AIAA 1998).

> Model: A representation of a physical system or process intended to enhance our ability to understand, predict, or control its behavior.
>
> Simulation: The exercise or use of a model. (That is, a model is used in a simulation).
>
> Verification: The process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution of the model.
>
> Validation: The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.
>
> Uncertainty: A potential deficiency in any phase or activity of the modeling process that is due either to a lack of knowledge (epistemic uncertainty or incertitude) or due to variability or inherent randomness (aleatory uncertainty).
>
> Error: A recognizable deficiency in any phase or activity of modeling that is not due to lack of knowledge.
>
> Prediction: Use of a model to foretell the state of a physical system under conditions for which the model has not been validated.
>
> Calibration: The process of adjusting numerical or physical modeling parameters in the computational model for the purpose of improving agreement with experimental data.

Our definition of uncertainty differs from the original definition of the AIAA in that we expand the definition of uncertainty to also include aleatory uncertainty (see Calder et al. 2018; Hoffman et al. 2019; and references therein).

Another perspective comes from Roache (1998b), who offers a concise, albeit informal, summary of V&V terminology:

> First and foremost, we must repeat the essential distinction between Code Verification and Validation. Following Boehm (1981) and Blottner (1990), we adopt the succinct description of "Verification" as"solving the equations right", and "Validation" as"solving the right equations". The code author defines precisely what partial differential equations are being solved, and convincingly demonstrates that they are solved correctly, i.e., usually with some order of accuracy, and always consistently, so that as some measure of discretization (e.g., the mesh increments) $\Delta \rightarrow 0$, the code produces a solution to the continuum equations; this is Verification. Whether or not those equations and that solution bear any relation to a physical problem of interest to the code user is the subject of Validation.

Roache also notes that a "code" cannot be validated but only a calculation or range of calculations can be validated. Roache also makes a distinction between verifying a code and verifying a calculation, noting that "use of a verified code is not enough". We also adhere to this explication of V&V terminology and note that following Roache, validation can be described as probing the range of validity of a code or model Calder et al. (2002).

Our approach to verification consists of testing simulation results against analytic or benchmarked solutions and quantifying the error. The comparisons typically consist of simulations performed at increasing spatial and/or temporal resolutions to confirm convergence of the simulation to the correct answer. Details of these tests have appeared in the literature, and many of the tests are incorporated into automated regression testing of Flash (Calder et al. 2002; Weirs et al. 2005a, b; Dwarkadas et al. 2005; Hearn et al. 2007; Dubey et al. 2009, 2015).

We validate by performing similar tests against data from experiments designed to replicate astrophysical environments. We take a hierarchical approach to validation, beginning by isolating the basic underlying physics and testing how well simulations capture it. We then devise tests of aggregate problems that capture the expected behavior of the astrophysical events. In the case of sub-grid models or postprocessed results, we simulate simple problems with these models and compare against either actual validation data or direct numerical simulations. As with verification, we perform convergence tests, though as we describe below the process of demonstrating convergence is difficult for some fluid dynamics problems.

Another aspect of our testing concerns quantifying error on the adaptive simulation mesh (described below). Our approach is to test solutions on the finest simulation mesh against data or a solution, but the methodology for quantitatively comparing the solution at the different resolutions of an adaptive mesh is incomplete (Li 2010; van der Holst et al. 2011; Shu et al. 2017; Li and Wood 2017). We typically check for consistency between simulations on an adaptive mesh and simulations of the same problem on a fully refined mesh while quantifying the accuracy of the solution on the fully refined mesh (Calder et al. 2002). Also, in addition to problems characterizing solutions on an adaptive mesh, just simulating fluids at the extreme Reynolds numbers of astrophysics on adaptive meshes presents challenges (Kritsuk et al. 2006; Mitran 2009). We describe the difficulties of simulating extreme Reynolds number flow in the discussion of our hydrodynamics method below.

We close discussion of our approach to V&V with a general note on the role of validation in astrophysics. Because of the literally astronomical distances to astrophysical events and extreme conditions involved, experimentally accessing astrophysical phenomena or even just replicating the environments of astrophysics is difficult. Thus, one cannot readily perform validation experiments, which typically leads to an incomplete process of validation. Simulations of astrophysical events are therefore generally in the realm of prediction, that is, foretelling the state of a physical system under conditions for which the model has not been validated. Despite this, the process of V&V in astrophysics serves to build confidence in these predictions even if one cannot conclude that simulations or codes are "validated".

28.3 Simulation Instruments

Our principal simulation instrument is the Flash code, which we use for simulating astrophysical events. Fundamentally, Flash simulates problems of fluid dynamics and consists of solvers for hydrodynamics and the additional physics of astrophysical events (described below). With Flash, we construct the numerical implementation of our conceptual model of the astrophysical event, and the act of simulating is the exercise of the model. We note that the exercise of a model is far more than just solving a set of differential equations. Multi-physics applications like astrophysics combine multiple solvers, each of which may rely on possibly uncontrolled assumptions (See Winsberg 2010, for a thorough discussion). For this reason, we take the hierarchical approach to validation of modules in Flash mentioned above.

Our second instrument is a nucleosynthetic postprocessing toolkit used in tandem with Flash. In the case of supernovae, comparison to observations requires the calculation of light curves (the intensity of light from the object as a function of time) and spectra. However, the yield of a particular element, titanium for example, may be critical for accurate spectra, but mostly unimportant to the energy release. Many elements fall into this category, so that the computation of the explosion is much less expensive when split into two stages. The energy release and explosion is computed with a small number of species in Flash, and is followed by postprocessing to obtain all important species. The postprocessing tools we present below apply state-of-the-art nuclear reaction networks to Lagrangian thermodynamic histories sampled from the Flash simulation. The resulting abundances are used to calculate light curves and spectra (e.g., Miles et al. 2016).

28.3.1 The Flash Code

The simulation instrument we use for modeling astrophysics events is the Flash code, developed at the University of Chicago (Fryxell et al. 2000; Calder et al. 2000; Dubey et al. 2009, 2013, 2014). Flash is a parallel, adaptive mesh, hydrodynamics plus additional physics code originally designed for the compressible fluid flows associated with astrophysics. Flash incorporates multiple hydrodynamics methods (Fryxell et al. 2000; Lee and Deane 2009; Lee 2013; Lee et al. 2017a, b) coupled with modules for the requisite additional physics of the applications. In particular, Flash has undergone considerable development for high energy density physics applications (Tzeferacos et al. 2015).

The hydrodynamics modules solve the Euler equations of compressible hydrodynamics, shown here with gravitational sources as would apply to a self-gravitating problem such as a star.

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0$$

$$\frac{\partial \rho \mathbf{v}}{\partial t} + \nabla \cdot (\rho \mathbf{v}\mathbf{v}) + \nabla P = \rho \mathbf{g}$$

$$\frac{\partial \rho E}{\partial t} + \nabla \cdot (\rho E + P)\,\mathbf{v} \;=\; \rho \mathbf{v} \cdot \mathbf{g} + S\;.$$

Here, ρ is the mass density, $\mathbf{v}$ is the velocity, P is the pressure, E is the internal energy of the gas, $\mathbf{g}$ is the gravitational acceleration, and S represents any additional source. The system is closed by an equation of state of the form

$$P = P\,(\rho, E)$$

and Flash offers choices for particular applications. Flash calculates the acceleration due to gravity from the gravitational potential,

$$\mathbf{g} = -\nabla\Phi \ .$$

which is calculated by solving the Poisson equation for self-gravity

$$\nabla^2\Phi\left(\mathbf{r}\right) = 4\pi G\rho\left(\mathbf{r}\right) \ .$$

Here, ϕ is the gravitational potential and G is Newton's gravitational constant. Flash also tracks different material species by advecting mass scalars,

$$\frac{\partial X\rho}{\partial t} + \nabla\cdot\left(X\rho\mathbf{v}\right) = 0 \ ,$$

where X is the mass fraction of a given species (Fryxell et al. 2000).

Our first validation example addressed the Flash hydrodynamics module (without gravity) for the case of experiments involving fluid instabilities thought to occur during one class of stellar explosions known as a core collapse supernova (Fryxell et al. 1991). The particular hydrodynamic module in Flash used for this study is based on the PROMETHEUS code (Fryxell et al. 1989) and evolves the Euler equations in one, two, or three dimensions using a modified version of the Piecewise-Parabolic Method (PPM) (Colella and Woodward 1984). The implementation allows use of general equations of state as is required for simulating stellar material (Colella and Glaz 1985), but this capability was not used in the validation example.

PPM is a higher order version of the method developed by Godunov (1959), Godunov et al. (1962), a finite-volume conservation scheme that solves the Riemann problem at the interfaces of the control volumes to compute fluxes into each volume. The conserved fluid quantities are treated as cell averages that are updated by the fluxes at the interfaces. This treatment has the effect of introducing explicit nonlinearity into the difference equations and permits the calculation of sharp shock fronts and contact discontinuities without introducing significant nonphysical oscillations into the flow. In addition, PPM utilizes a dissipative shock capturing scheme to further stabilize shocks and contact discontinuities, and is thus not directly solving the Euler equations (Majda 1984; Winsberg 2010).

The adaptive mesh of Flash is block structured and is supported primarily through the Paramesh Library (MacNeice et al. 1999, 2000), though it is under the process of migrating to the AMReX library (AMReX 2018). The view of AMReX from other units in the Flash code will remain similar to that of Paramesh, and in the near future, the two packages will be available as alternative implementations of the Grid unit. Later, the support for Paramesh may be dropped if it becomes too inefficient on newer platforms.

28.3.2 The Postprocessing Toolkit

The nucleosynthetic postprocessing toolkit uses the recorded Lagrangian history of fluid elements to compute the yield of nuclides (elements and their isotopes) produced in a stellar explosion (Travaglio et al. 2004; Townsley et al. 2016). The Lagrangian thermodynamic history is determined by integrating the position of a conceptual microscopic fluid element by

$$\mathbf{r}(t) = \mathbf{r}_0 + \int_0^t \mathbf{v}(\mathbf{r}, t')\, dt' \ ,$$

where $\mathbf{r}_0$ is the initial position and $\mathbf{v}(\mathbf{r}, t)$ is the velocity field as computed by the hydrodynamic simulation. This conceptual fluid element is often called a particle because it moves as a fluid-embedded particle would. From the resulting $\mathbf{r}(t)$, it is possible to also record the thermodynamic state, namely $T(t) = T(\mathbf{r}(t))$ and $\rho(t) = \rho(\mathbf{r}(t))$, the temperature and density, respectively. Such recorded histories are often called tracks or trajectories because they represent how the fluid element evolves in location and thermodynamic state space as a function of time.

Nucleosynthetic postprocessing is performed in order to obtain the composition of material after it is processed by combustion and ejected. Composition is parameterized by abundances of various species quantified as the fraction of a unit of mass that is in the form of a particular species. For example, the fraction, by mass, that is in the form of ^{12}C, may be written $X_{^{12}\mathrm{C}}$, and must be between 0 and 1. The abundances are found in postprocessing by integrating

$$X_i(t) = X_{i,0} + \int_0^t \dot{X}_i(\rho(t'), T(t'))\, dt' \ ,$$

where $\dot{X}_i(\rho, T)$ are determined by the density and temperature-dependent reaction rates for processes which involve species i. Any given specie is typically involved in multiple reactions, forming a network that is used to evaluate each rate. The end of the necessary integrations is typically well defined. As the star expands, T and ρ fall until most reactions will become very slow compared to the time being simulated, effectively freezing out. Consideration of further evolution, typically radioactive decay, may be necessary depending on the usage of the resulting abundances. These integrations are typically performed for a large number of tracks which sample the ejected material by a suitably distributed choice of their initial positions $\mathbf{r}_0$.

28.3.3 Simulating Reactive Flow

With both Flash and the postprocessing toolkit, the goal of simulations is to capture the evolution of stellar material during the course of an astrophysical event. Because stars are essentially self-gravitating gas, the interiors of stars are well described by the equations of fluid flow. During an astrophysical event, thermonuclear reactions drive the evolution by changing the composition and by releasing energy, which changes thermodynamic conditions like pressure and density. This combustion typically occurs in a relatively small region of space, e.g., a thin flame, that may be difficult to resolve in simulations of the event. The two validation examples we present address the two principal parts- fluid flow, including shocks and fluid instabilities, and the evolution of the composition.

For fluid dynamics problems, there are two fundamental classes of simulation distinguished by whether or not the scales of the numerical grid can resolve viscous effects (Calder et al. 2002; Winsberg 2010; and references therein). Simulations that can resolve viscous effects are said to be "Direct Numerical Simulations," while those that cannot and rely on a (possibly uncontrolled) sub-grid-scale model for viscous effects are referred to as "Large Eddy Simulations". An eddy is a fluid current whose flow direction differs from that of the general flow, and the motion of the fluid is the net result of the movements of the eddies that compose it (Encyclopaedia Britannica 2006). Large eddy simulations do not resolve either the explicit viscosity of the fluid or the contribution to the viscosity from eddies on unresolved scales (Fureby 1996; Zhiyin 2015, and references therein).

The issue of convergence of a solution for fluid flows is not as simple as it might seem. The enormous size of objects means that astrophysical regimes typically have Reynolds numbers far in excess of the Reynolds numbers of terrestrial flows, which are themselves higher than can be readily captured in hydrodynamics simulations. Even when run on contemporary supercomputers, simulations cannot capture the possibly $\geq 10^8$ Reynolds numbers of astrophysical flows making direct numerical simulations impossible. Thus simulations of astrophysical events are large eddy simulations that can either rely on sub-grid-scale models for turbulent flow or just allow the intrinsic numerical diffusion of the hydrodynamics method to set the limiting Reynolds number. This latter case, known as Implicit Large Eddy Simulation (ILES), is frequently applied and is the approach taken in the studies presented here. In ILES, changing the resolution changes the effective viscosity and hence the Reynolds number, which changes the problem itself and leads to the question of convergence of results with resolution. Considerable study has gone into determining the validity of this approach (Margolin and Rider 2002; Grinstein et al. 2007; Margolin and Shashkov 2008; Margolin 2014). As our results show, large eddy simulations may not demonstrate convergence of a solution with resolution.

28.4 Validation Examples

As of this writing, Flash has had 20 years of development by generations of scientists. Much of this effort has been subjected to rigorous V&V (Calder et al. 2002; Timmes et al. 2004; Weirs et al. 2005a, b; Dwarkadas et al. 2005; Hearn et al. 2007; Dubey et al. 2009, 2015; Townsley et al. 2016). In this contribution, we present two examples of validating the Flash code and postprocessing toolkit for astrophysical applications. The first example is from early work comparing simulations to laboratory experiments addressing fluid instabilities in high energy density environments similar to the interiors of stars. The second example is ongoing work on computing reaction products in three-dimensional simulations of type Ia supernovae. This study includes comparison between methods for use in the simulations of the events and for calculating detailed abundances from the density and temperature histories of Lagrangian tracers.

While this contribution describes two examples of the V&V efforts for the Flash code, we note that V&V efforts continue as the capabilities and applications of Flash evolve. A recent survey of software engineering practice in scientific computing includes Flash as a case study and offers an independent perspective on the development of Flash (Storer 2017).

28.4.1 Overview of Flash Problems

The Flash code was originally designed to investigate astrophysical thermonuclear flashes, explosive events powered by thermonuclear fusion. These events all involve a close binary star system with matter being transferred (accreted) onto a compact star (either a white dwarf or a neutron star) from a companion (Rosner et al. 2000). The three flash problems originally addressed by Flash were type I x-ray bursts (Zingale et al. 2001), classical novae (Alexakis et al. 2004), and type Ia supernovae (Plewa et al. 2004; Townsley et al. 2007).

X-ray bursts occur when a thermonuclear runaway occurs in a thin $\sim$10–100 m layer of H- or He-rich fuel accreted onto the surface of a neutron star. The radius of the underlying neutron star my be inferred from observations and thereby allow constraints on the properties of dense matter. Classical novae occur when a thermonuclear $\sim 10^4$ m thick layer of H-rich material similarly explodes. In this case, material from the explosion is unbound and these events are thought to synthesize some intermediate-mass elements found in the galaxy. Type Ia supernovae are thought to occur when a pair of white dwarf star merge and/or when a white dwarf accretes enough mass to ignite fusion in the core. In this case, enough energy is added to overcome the gravitational binding and the star is completely disrupted, producing a bright explosion that may be used as an indicator for cosmological distances. (See references in above works for literature on each astrophysical topic, and Calder et al. 2013 for an overview of ongoing investigation of Type Ia Supernovae.)

As mentioned above, these problems involve reactive flow, and in all cases there is a vast difference between the length scale of the combustion front and the astrophysical object. Hence the need for sub-grid-scale models. Fluid instabilities that may influence the burning rate are also of particular importance (Calder et al. 2007; Zhang et al. 2007; Townsley et al. 2016). Accordingly, the validation examples we present address problems of combustion and fluid instabilities.

28.4.2 Shocks and Fluid Instabilities

The high energy density environments of intense lasers interacting with matter are similar to the interiors of stars, and experiments offer opportunities for a quantitative comparison between data and simulation not possible with observations of astrophysical phenomena. The validation study we present was performed by a collaboration between Flash developers and experimentalists working at the Omega laser at the University of Rochester (Soures et al. 1996; Boehly et al. 1995; Bradley et al. 1998). The experiment chosen for the study involved a shock propagating through a multilayer target with layers of decreasing density and was designed to produce hydrodynamic instabilities thought to arise during an astrophysical event known as a core collapse supernova explosion (Arnett et al. 1989; Fryxell et al. 1991). While this type of supernova is not a thermonuclear flash problem, much of the constituent physics is the same, allowing this experiment to serve for validation. The decreasing density configuration is subject to the Richtmyer–Meshkov instability that occurs when a shock propagates though a material interface with decreasing density (Richtmyer 1960; Meshkov 1969). The configuration is also subject to the Rayleigh–Taylor instability (Taylor 1950; Chandrasekhar 1981), which develops after the passage of the shock and subsequently dominates instability growth.

Interest in the problem of fluid instabilities during the process of a core collapse supernova followed from the early observation of radioactive elements from deep in the core of the star in SN 1987A (Muller et al. 1989). Stars with a mass of greater than 8–10 times that of the Sun end their lives in a spectacular explosion known as a core collapse supernova. These events are among the most powerful explosions in the cosmos, releasing energy of order 10^{53} erg at a rate of 10^{45-46} watts, and are important for galactic chemical evolution because they produce and disseminate heavy elements. Core collapses supernovae also signal the birth of neutron stars and black holes, which are the basic building blocks of other astrophysical systems such as pulsars and x-ray binaries.

During their lifetimes, stars are powered by the thermonuclear fusion of elements beginning with hydrogen fusing into helium. In a massive star, fusion continues all the way to iron-group elements. A core collapse supernova occurs when the inert iron core can no longer support the weight of the material above it and the core collapses, which releases gravitational binding energy that is in part converted to the energy of an expanding shock that ejects the outer layers of the star. Just prior to the explosion, the interior of the star has an onion-like structure, with iron-group elements in the core, then layers of silicon, magnesium, neon, oxygen, carbon, helium, and finally the outermost layer may be hydrogen. When the supernova explosion occurs, the shock passes through these layers of decreasing density. The early observation of a core element suggests some sort mixing must have occurred during the explosion, and, accordingly, motivated investigation into the effects of fluid instabilities. The laboratory experiment was designed to probe this scenario.

The experimental configuration consists of a strong shock driven through a target with three layers of decreasing density. The interface between the first two layers is perturbed while the second interface is flat. An initially planar shock created by the deposition of energy from the laser is perturbed as it crosses the first interface, which excites a Richtmyer–Meshkov instability. As the perturbed shock propagates through the second interface, the perturbation is imprinted on the interface. The material begins to flow, leading to the growth of Rayleigh–Taylor instabilities. The three layers of the target are in a cylindrical shock tube composed of Be, with the density decreasing in the direction of shock propagation. The materials were Cu, polyimide plastic, and carbonized resorcinol formaldehyde (CRF) foam, with thicknesses of 85, 150, and 1500 μm and densities 8.93, 1.41, and 0.1 g cm^{-3}, respectively. The shock tube delays the lateral decompression of the target, keeping the shock planar. The surface of the Cu layer was machined with a sinusoidal ripple of wavelength 200 μm and amplitude 15 μm to perturb the shock as it passes this interface.

The experiment was driven by 10 beams of the laser with the target configured so that the laser beams impinge a thin section of CH ablator to prevent direct illumination and preheating of the target. The experimental diagnostics were X-ray radiographs taken at different times during a "shot". The Be shock tube, polymide plastic, and CRF foam are transparent to X-rays, while the Cu layer is opaque to X-rays. Embedded within the polyimide layer was a tracer strip of brominated CH that is also opaque to X-rays. This tracer layer provided the diagnostic for polymide-foam interface, allowing visualization of the shock-imprinted structure.

Figure 28.1 shows X-ray radiographs of the experiment at two times, one relatively early at 39.9 ns (left) and one relatively late at 66.0 ns (right). These images were from two different shots. The long, dark "fingers" are spikes of expanding Cu, and

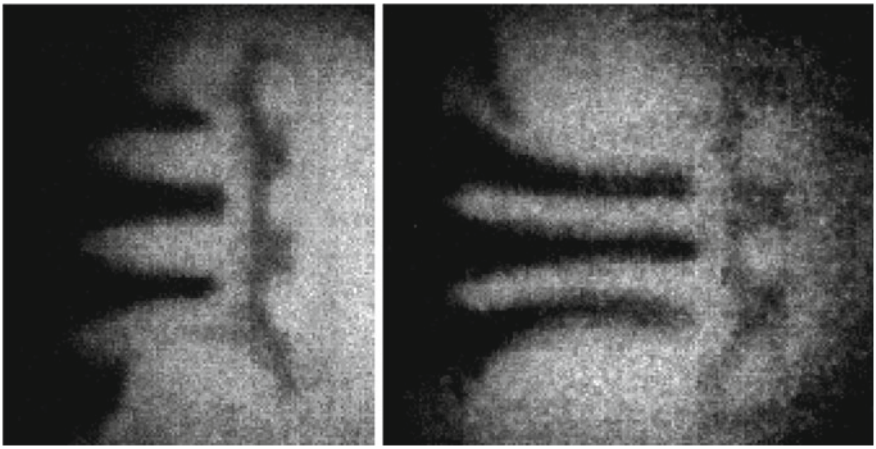


Fig. 28.1 Results of the three-layer target experiment. Shown are side-on X-ray radiographs at 39.9 ns (left) and 66.0 ns (right). The long, dark "fingers" are spikes of expanding Cu, and the horizontal band of opaque material to the right of the spikes of Cu is the brominated plastic tracer showing the imprinted instability growth at the plastic-foam interface. From Calder et al. (2002) © AAS. Reproduced with permission

the vertical band of opaque material to the right of the spikes of Cu is the brominated plastic tracer, showing the imprinted instability growth at the plastic-foam interface. The radiographs illustrate the configuration at early and late times in the evolution of the shocked target. The outer regions of the Cu and brominated strip show the effects of the shock tube, but the central part is largely immune to these effects.

Making a quantitative comparison between the simulations and the experiments and determining the uncertainty in the study required close collaboration between experimentalists and theorists. This is an important point worth stressing. Without the contribution of both to interpreting and quantifying the experiments and simulations, there would have been little chance for a meaningful quantitative comparison. The data from the experiments are the radiographs, and finding a meaningful measurement for comparison to the simulation results required understanding the accuracy of the diagnostics and sources of uncertainty in the experiment. The metric for comparison between simulation and experiment was chosen as the length of the copper spikes, which are fairly obviously seen in the radiograph, but which required a deep understanding of the experiments to quantify. The paragraphs below summarize the sources of error and uncertainty in the experiments and the reader is referred to the original paper for complete details (Calder et al. 2002). A cautionary note concerning these details is warranted, however. The intervening years between these experiments and this writing have seen enormous progress in diagnosing high energy density experiments and the experiments described here are not the current state of the art (Gamboa et al. 2012, 2014; Stoeckl et al. 2012).

The lengths of the Cu spikes in the experimental radiographs were determined by three methods. The first was a straightforward visual inspection of the images using a spatial reference grid located just below the images of Fig. 28.1. The second used a contour routine to better quantify the uncertainty in the location of the edges of the spikes. The third method was consistent with the analysis of the simulations. A section in the center of the images was vertically averaged to produce a single spatial lineout of optical depth through the region occupied by the Cu and CH. The same 5 and 90 threshold values were used to quantitatively determine the extent of the Cu spikes. Taking the average of all three methods, values of 330 ± 25 μm and 554 ± 25 μm are obtained at 39.9 and 66.0 ns, respectively.

Sources contributing to uncertainty in these experimental measurements include the spatial resolution of the diagnostic, the photon statistics of the image, target alignment and parallax, and the specific contrast level chosen to define the length of the Cu spikes. These considerations allowed calculation of the experimental error bars presented in the figure (described below) that compares the experimental results to the simulation results. In addition to the spatial uncertainty, there were also several sources of uncertainty in the temporal accuracy. These arise from target-to-target dimensional variations, shot-to-shot drive intensity variations, and the intrinsic timing accuracy of the diagnostics. The experimental uncertainty in the timing is, however, relatively small, and is approximately indicated by the width of the symbols used in the comparison figure.

The Flash simulations were two-dimensional with a three-layer arrangement of Cu, polyimide CH, and C having the same densities as those of the experimental

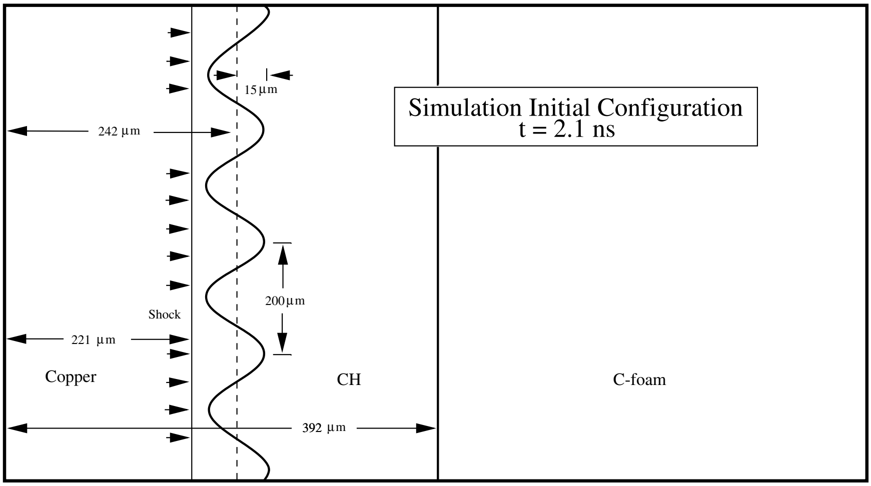


Fig. 28.2 Schematic of the three-layer target simulation initial conditions. Shown are the locations of the three materials, Cu, CH, and C, the shock, and the details of the sinusoidal perturbation of the Cu–CH interface. The schematic is not to scale. From Calder et al. (2002) © AAS. Reproduced with permission

target to model the experiment. The initial conditions for the Flash simulations represent the configuration 2.1 ns after the laser shot. At this point, the laser has deposited its energy and the shock is approaching the Cu–CH interface and the evolution is purely hydrodynamic. The initial conditions (thermodynamic profiles) for the Flash simulation were obtained from simulations of the laser–material interaction performed with a one-dimensional radiation hydrodynamics code (Larsen and Lane 1994) that was able to describe the process of energy deposition occurring in the initial 2.1 ns. These one-dimensional profiles were mapped onto the two-dimensional grid with a sinusoidal perturbation added to the Cu–CH interface. Figure 28.2 illustrates the initial configuration of the Flash simulations. For convenience, the simulations used periodic boundary conditions on the transverse boundaries and zero-gradient outflow boundary conditions on the boundaries in the direction of the shock propagation. The materials were treated as gamma-law gases, with $\gamma = 2.0$, 2.0, and 1.3 for the Cu, CH, and C, respectively. These values for gamma were chosen to give similar shock speeds to the shock speeds observed in the experiments.

From these initial conditions, the simulations were evolved to approximately 66 ns. Figure 28.3 shows simulated radiographs from a simulation at an intermediate resolution, allowing visual comparison to the experimental results. The figure presents simulated radiographs at approximately the two times corresponding to the images from the experiment, 39.9 ns (left panel) and 66.0 ns (right panel). The simulation in Fig. 28.3 had 6 levels of mesh refinement corresponding to an effective resolution of 1024×512 grid zones. The simulated radiographs were created from the abundances of the three materials assigning an artificial opacity to each abun-

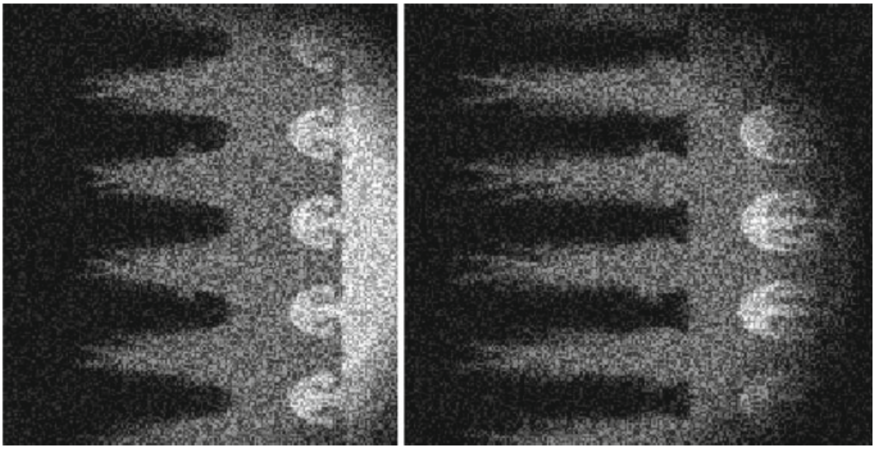


Fig. 28.3 Simulated radiographs from the six levels of refinement (effective resolution of 512 × 256) simulation of the three-layer target experiment. The simulated radiographs were created from the fluid abundances at times corresponding approximately to those of the images from the experiment, 39.9 ns (left) and 66.0 ns (right). Shown are the parts of the simulation domain that match the regions in the experimental results. From Calder et al. (2002) © AAS. Reproduced with permission

dance and applying the opacity to an artificial "beam". In addition, the abundances were de-resolved to match the resolution of the pixels in the experimental images and random Poisson-distributed "noise" was added to the intensity.

An obvious qualitative difference between the simulated and experimental radiographs is readily observed in the curvature of the experimental instabilities that is not present in the simulations instabilities. The use of periodic boundary conditions in the transverse directions in the simulation was not consistent with the boundary conditions of the experiment, which was performed with the three materials of the target inside a cylindrical Be shock tube. The experiment results show the influence of the shock tube walls as a curving or pinching of the outer Cu spikes, while the simulations did not consider these boundary effects.

Comparison of the simulated radiographs to the radiographs from the experiment show that the simulations captured the bulk behavior of the materials, particularly the growth of Cu spikes and the development of C bubbles. We can conclude from this comparison that the simulations resemble the experimental results. Assessment of the comparison as"good" or"bad" is difficult, however, with only a visual comparison, especially one that indicates a difference due to a boundary condition effect. What is needed is a quantitative comparison, and for that we apply the same techniques as we apply to verification, a convergence study to show the simulations converge with resolution and a quantitative comparison to the experimental results.

To test convergence of the solutions, a suite of simulations was performed at increasing resolution. The effective resolutions of the simulations were 128 × 64, 256 × 128, 512 × 256, 1024 × 512, 2048 × 1024, and 4096 × 2048, corresponding to 4, 5, 6, 7, 8, and 9 levels of adaptive mesh refinement. As noted above, the lengths of the Cu spikes were chosen as the metric for quantitative comparison to the experiments. Flash solves an advection equation for each abundance, which allowed tracking the flow of each material with time. The spike lengths in the simulations were measured by averaging the CH abundance in the y-direction across the simulation domain then smoothing the resulting one-dimensional array slightly to minimize differences that would occur owing to very small-scale structure. The length of the Cu spikes was then determined by the average distance spanned by minimum locations of average abundances 0.05 and 0.9. The results were reasonably robust to the amount of smoothing and threshold values.

The results of testing the convergence of the Cu spike length measurements are shown in Fig. 28.4, which depicts percent differences from the highest resolution simulation, 9 levels of adaptive mesh refinement, as functions of time. The trend is that the difference decreases with increasing mesh resolution, with the 7 and 8 level of adaptive mesh refinement simulations always demonstrating agreement to within 5%. The trend of decreasing difference with increasing mesh resolution demonstrates a convergence of the flow, but it is subject to caveats. We note that the trend does not describe the behavior at all points in time (that is, the percent difference curves sometimes cross each other), and this average measurement is an integral property of the flow and in no way quantifies the differences in small-scale structure observed in the abundances. In particular, we note that the difference curve for the simulation with 8 levels of adaptive mesh refinement crosses the curves of both the 7 and 6 level simulations, suggesting that higher resolution simulations may deviate further from these results and produce degraded agreement with the experiment. This result is in keeping with the abovementioned concerns with ILES.

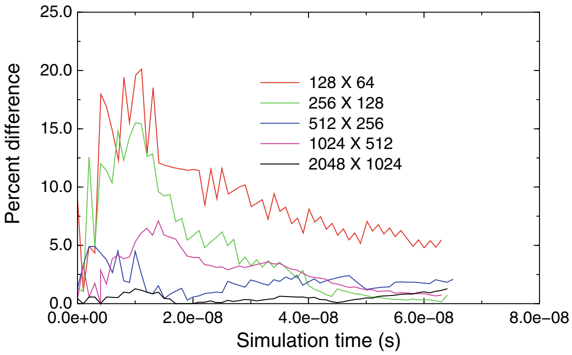


Fig. 28.4 Percent difference of the Cu spike lengths from those of the highest resolution (9 levels of adaptive mesh refinement) simulation versus time. The percent differences are from the lower resolution simulations of 4, 5, 6, 7, and 8 levels of adaptive mesh refinement, with the corresponding effective resolutions in the legend. We note that the convergence is not perfect. The curve from the 8 levels of refinement simulation crosses those of the 6 and 7 levels of refinement simulations, indicating a higher percent difference. Adapted from Calder et al. (2002)

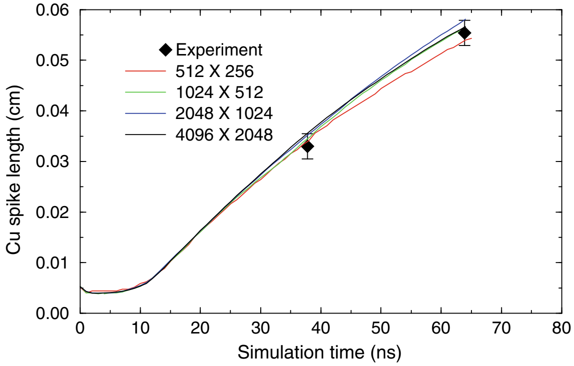


Fig. 28.5 Results from a validation test consisting of a laser-driven shock propagating through a multilayer target. The lengths of the Cu spikes is plotted versus time from 4 simulations at 6, 7, 8, and 9 levels of adaptive mesh refinement in a convergence study. The effective resolutions are given in the legend. Also shown are the experimental results at two times with spatial error bars of ($\pm 25 \mu$m). The timing error is about the width of the diamonds marking the experimental result. The differences between the simulations at different resolutions are less than the uncertainty of the experimental results. Adapted from Calder et al. (2002)

Figure 28.5 shows the Cu spike length versus time for 4 simulations at increasing resolution. Also shown are the abovementioned experimental results. The experimental error bars correspond to ±25 μm, the spatial error of the experiment. The width of the symbols marking the experimental results indicates approximately the timing error. The figure shows that the simulations quantitatively agree with the experimental results at the early and late times to within the experimental uncertainty.

As noted above, this study has previously appeared in the literature. Complete details of the validation study may be found in Calder et al. (2002), Calder (2005), Calder et al. (2006) and additional details of the experiments may be found in Kane et al. (2001), Robey et al. (2001).

28.4.3 Computation of Reaction Products in Large Eddy Simulations Of Supernovae

When a laboratory experiment is available, the distinction between verification and validation is fairly clear, as discussed earlier. However, when creating predictive simulations of astrophysical processes that cannot be reproduced directly in the laboratory, even using appropriate scaling laws, the distinction can become less clear because the task becomes one of confirmation of simulation results without laboratory results. In many situations, notably in stellar combustion, it is possible to have a model that is demonstrably more physically valid but is too expensive or constrained to be used for the desired predictive simulations. Simpler models must be applied to simulate observed phenomena, hence the need for comparison of different methods.

Nuclear reaction networks and multidimensional simulations present a good example of this confluence of verification and validation. In astrophysical detonations, it is possible to compute the steady-state structure of the propagating reaction front with a large reaction network with hundreds of species and thousands of reactions using error-controlled numerical methods (e.g., Sharpe 1999; Moore et al. 2013) Consider the following question: How many species are necessary to accurately predict the characteristics of the flow such as peak temperature and reaction front width? This is not a verification question. We can use verification techniques to demonstrate that the equations governing the time integration of the reactions are being solved to a desired accuracy. Such a test, however, does not demonstrate whether or not a particular selection of species is sufficient for the stated purpose. So, we proceed to compare our model with say three or a dozen "effective" reactions or species to another model which we believe to be more physically valid because it has more complete reaction physics. This situation is neither verification that our model is being solved correctly (that is already done) nor is it validation against a specific physical experiment. It is, however, validation under the definition introduced in Sect. 28.2 above, in that it addresses whether the model is physically correct. Some terminology refers to this as confirmation of one model with a physically more valid model. Since the label depends finely on definitions of terminology, it is useful in discussion to term this type of comparison as something that combines elements of verification and validation (see Chap. 42 by Beisbart in this volume). It is a model-to-model comparison, as verification often is, but addresses the physical applicability of the model, as validation does.

If integration of thousands of reactions were the only issue, this validation of simplified models might not be worthwhile; instead one would just use the better model directly. There are areas of prediction, however, where direct use of the better model can be infeasible. In explosive astrophysical combustion (which powers type Ia supernova explosions), it is typically desirable to predict the overall products and the speeds at which they are ejected. Unfortunately, a simulation that can predict that information must include the entire star, which may be around 10^9 cm in size. The reaction front through which the combustion takes place is one cm or less in thickness (Townsley et al. 2016). Also, the propagation of this front through the star will generally occur in a way that obeys no particularly symmetry, making it necessary to simulate this combustion and ejection of material in three dimensions.

The necessity of simulating the whole star in three dimensions presents several challenges from the standpoint of V&V. First, since the combustion phenomena occur far below the best possible grid scale ($\sim 10^5$ cm), the typical method of verification by convergence study is not valid. Claiming convergence for a numerical solution of differential equations presupposes that the relevant gradients are numerically resolved and become better resolved at higher resolution. This is the very meaning of resolution. However, in the full-scale astrophysical case, an example of the abovementioned large eddy simulation situation, the composition gradients representing the physical reaction front (the length scale over which the fuel is consumed and converted to products) are never actually resolved. Second, while error-controlled methods for ODE integration are well-understood, similar automated control of accuracy is not available in current widely used methods for solution of PDEs, such as in hydrodynamics. Because this control is not built into the method, performing predictive simulations involves a constant process of verification to ensure that solutions obtained do not depend on resolution. That process can be both expensive and time-consuming. Third, it may be computationally infeasible to include hundreds of species and thousands of reactions in the full-scale hydrodynamic simulation, thus even if we were able to verify the methods for reactive hydrodynamics, we would need to use a model for the reactions that we know to have specific deficiencies and would therefore need some form of validation against more physically complete models. Finally, as discussed earlier, because some physical processes such as fluid dissipation due to viscosity is left implicit, a higher resolution simulation may not only be more numerically accurate but also more physically valid. As a result of these issues, verification and validation of the simulation of a stellar explosion can be mixed in a way that is not always cleanly separable.

Here, we will present a discussion of ongoing efforts at verification and validation of methods for computing the products of thermonuclear supernova explosions. The full-star simulations use a simplified model of the reactions for computational efficiency, and are necessarily under-resolved. The overall goal is to compare the results from this computational model to computational models of much higher physical and numerical fidelity. In the case of combustion, those are computations with large, complete nuclear reaction networks computed using resolved, error-controlled numerical techniques. The limitation is that the latter methods can only be used under certain flow conditions, specifically, a steady state. We therefore proceed by treating the methods used in the full-star simulation as the model to be validated by comparison to more physical calculations. This is similar to verification by comparison to a benchmark, except that the two models are known to be different by construction.

Table 28.1 shows a matrix comparing the capabilities of compressible hydrodynamics simulations in various dimensions as well as the fully resolved method, which can only be used in one dimension and for reaction fronts propagating in a steady state through a uniform medium. As shown, a resolved calculation with the full network at all densities relevant to the supernova can only be performed with the steady-state method. However, this method cannot be used to treat transients (e.g., ignition or nonspatially uniform density) or general geometries including the full star. Of the hydrodynamical methods in various spatial dimensions, represented in the other three columns of the table, only one-dimensional calculations can use a full reaction network effectively and resolve the reaction front, though not at all densities. The possible importance of transient effects necessitates a multistep strategy utilizing cross-comparisons of calculations of reaction front structure among several different methods. For example, we can verify one-dimensional dynamical calculations at uniform densities using comparison to steady-state calculations, and then use one-dimensional calculations with nonuniform density to characterize transient

Table 28.1 Capabilities of simulations in various dimensions and assumptions. Comparison of results among simulations is performed in order to validate that full-star three-dimensional simulations reproduce the results of more physically valid one-dimensional calculations of steady-state properties of detonations

Capability	3-d	2-d	1-d	1-d steady
full reaction network	×	×	✓	✓
resolved at low density	×	×	✓	✓
resolved at high density	×	×	×	✓
transients (dynamical)	✓	✓	✓	×
general geometries	✓	×	×	×
full star	✓	✓	✓	×

effects. Even for a transient, it is informative to compare to steady-state solutions in order to provide physical insight to the importance of nonuniformities in density.

Figure 28.6 shows an example of a comparison of the compositional structure of a propagating detonation reaction front computed with the one-dimensional dynamical method and the one-dimensional steady-state method. The hydrodynamical simulation (dashed lines) was performed at a physical resolution of 10^5 cm, which corresponds to a hydrodynamical time step of about 10^{-4} s. The fuel here is mostly ^{12}C and ^{16}O, which is reacted to eventually become ^{56}Ni. The consumption of ^{12}C is not shown, but is even faster than that of ^{16}O. The structure for a detonation propagating in steady state (solid lines) is computed with an error-controlled method using adaptive

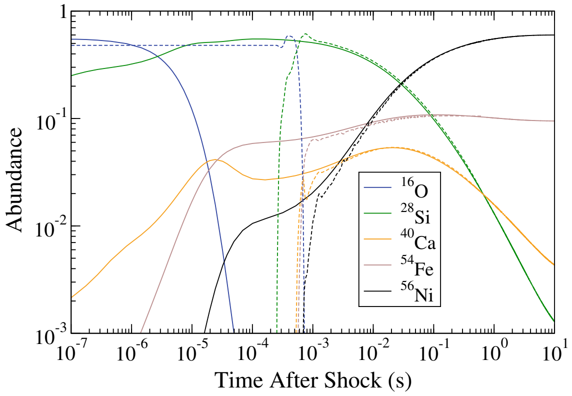


Fig. 28.6 Comparison of planar steady-state detonation structure simulated hydrodynamically at 10^5 cm resolution using postprocessing of Lagrangian tracers (dashed) with the steady-state structure computed directly using error-controlled integration (solid). Abundances here are given as mass fractions. Similar to comparisons made in Townsley et al. (2016). The oxygen consumption structure will remain unresolvable even with more than an order of magnitude higher resolution in the hydrodynamic simulation

time stepping and an error tolerance of order 10^{-6}, and is therefore suitably resolved by construction. The abundance histories from the hydrodynamical model shown here are the result of using a simplified reaction model in the hydrodynamics and then postprocessing the resulting density and temperature histories of fluid elements with a larger reaction network (Travaglio et al. 2004; Townsley et al. 2016). The goal of this comparison is to validate that away from the unresolved portion of the reaction front (timescales $\gtrsim 10^{-3}$ s), the composition history is accurately predicted by the under-resolved calculation with the simplified burning model. This comparison shows that the results are in good agreement for steady-state, planar detonations. For an example of a comparison for nonplanar (curved) detonations, see Moore et al. (2013).

The validation of methods for computing astrophysical combustion in large eddy simulations is ongoing. The various possible calculations represented in Table 28.1 must be compared for geometries and conditions for which there is overlap in capability. This process also entails ongoing improvement of both the simplified reaction model utilized in the large eddy simulations (Townsley et al. 2009; Willcox et al. 2016) as well as improving techniques for computing the final yields (Townsley et al. 2016).

28.5 Discussion

The simulational results for the hydrodynamics validation example fell within the temporal and spatial error bars of the experimental results thus showing quantitative agreement between simulation and experiment for the metric of the lengths of the copper spikes. This agreement demonstrates that the hydrodynamics module in Flash captured the bulk properties of the flow and observable morphology, which builds confidence in astrophysical simulations. We cannot, however, declare the code "validated" for a host of reasons:

- The experimental configuration produced essentially a two-dimensional result, hence our modeling it with two-dimensional simulations. The experiment was three-dimensional, so correctly describing the fluid instabilities, particularly the amount of small-scale structure in the flow may require three-dimensional simulations.
- The models were incomplete. The three materials were modeled as ideal gases, a questionable assumption. Also, for convenience, the simulations neglected the presence of the shock tube surrounding the target and assumed periodic boundary conditions. Thus, the simulations did not include effects due to the shock tube.
- The experimental diagnostics, radiographs, are really shadows that cannot adequately capture small-scale structure. Even if three-dimensional simulations that better described the fluid instabilities had been performed, comparison to the experimental results is limited by the experimental diagnostics.

- The observed degraded agreement between simulations at the highest resolutions indicates the results are not converged. We attribute this result to the fact that the Euler equations allow a changing numerical viscosity with resolution, which changes the Reynolds number and thus the nature of any turbulence. Additional commentary on this issue may be found in Calder et al. (2002).

Even with limitations, the demonstrated ability of the simulations to capture the expected bulk properties of the flow builds confidence in the results of astrophysical simulations, allowing us to conclude that the shocks and fluid instabilities study was a success. The principal differences observed between the results from simulations and the experimental results were in the amount of small-scale structure observed in the flow, with the amount of small-scale structure in the simulations increasing with resolution. This behavior is expected because the effective Reynolds number increases with resolution as described above, and we believe this effect is the source of the observed imperfect convergence. Because the experimental data are radiographs and cannot capture the actual amount of small-scale structure in the flow, the correct amount of small-scale structure remains undetermined and even if the convergence of the simulations had been perfect, we could not conclude the solution converged to the correct result.

In addition to increasing confidence in the results, the hydrodynamics validation study was well worth the investment because of the lessons learned in comparing the experimental and simulational results. The collaborative process of determining the metric for comparison and extracting the results from the experimental and simulational data resulted in a better understanding of the issues, which also increases confidence in the astrophysical results. The experimentalists also benefited from the process of validation because the process of comparison suggested metrics for future comparisons, provided useful diagnostics, and supplied a virtual model that aided in the design of future experiments. A point worth stressing again in conclusion is the importance of close collaboration between the experimentalists and theorists needed to make a meaningful quantitative comparison. Raw experimental data such as a radiograph alone does not allow for a quantitative comparison to simulational results. Finally, we note that the success of this collaboration seeded interest in high energy density physics among the developers of Flash, which subsequently resulted in an extended course of collaborative research into high energy density physics (see Tzeferacos et al. (2015) and references therein).

The product of reactive hydrodynamics study gave a look at the process of comparing models of differing fidelity to ensure that macroscopic (three-dimensional) simulations capture the physics of thermonuclear reactions while also allowing the calculation of detailed abundances. Our approach is to test simplified models against higher fidelity models for a given physical process, here thermonuclear combustion. Simplified models then facilitate three-dimensional simulations that would be intractable otherwise. The results of these studies are also applicable to the problem of determining detailed abundances from the density and temperature histories of Lagrangian tracers. We illustrated this process with a comparison between results from postprocessed tracers from a hydrodynamics simulation and a detailed calculation of steady-state burning structure. This study confirmed that our simulations capture the essence of the reactions in whole-star models, and thereby increased confidence in our predictions of the astrophysical events.

28.6 Conclusions

The cases we present here are but one part of the continuing effort at verifying and validating Flash and associated infrastructure (e.g., the postprocessing method presented here). The first study of validating the hydrodynamics was performed early in the development of Flash. Though very informative, it could have been continued further with additional quantification of the effect of missing physics as a good next step. Also, further modifications to the code would allow it to capture high energy density phenomena better. Such activities, however, were not critical to the astrophysical problems. Still, the case was very informative and served to increase confidence in the results. The second case, the computation of reaction products in large eddy simulations of supernovae, is very much a work in progress and represents our contemporary effort.

Our conclusion from both of these studies is that like any discipline in computational science, V&V are an essential part of the process of modeling astrophysical phenomena. V&V in astrophysics can be particularly challenging due to the inaccessibility of the physical conditions attained and limited ancillary measurements available for distant events. As shown here by these examples, however, positive steps that build confidence in models can be taken based on comparisons using related laboratory experiments and more complete physical models where available.

Acknowledgements This contribution was supported in part by the Department of Energy through grant DE-FG02-87ER40317 and the research described here was supported in part by earlier grant B341495 to the Center for Astrophysical Flashes at the University of Chicago. The software used in this work was in part developed by the DOE-supported ASC/Alliances Center for Astrophysical Thermonuclear Flashes at the University of Chicago. This research has made use of NASA's Astrophysics Data System. The authors thank Eric Winsberg and Bruce Fryxell for helpful discussions about this manuscript.

References

AIAA. (1998). Guide for the verification and validation of computational fluid dynamics simulations. AIAA Report G-077-1998. Reston, VA: American Institute of Aeronautics and Astronautics.

Alexakis, A., Calder, A. C., Heger, A., et al. (2004). *Astrophysical Journal*, *602*, 931.

AMREX. (2018). https://amrex-codes.github.io/, freely available.

Arnett, D., Fryxell, B., & Mueller, E. (1989). *Astrophysical Journal Letters*, *341*, L63.

Blottner, F. G. (1990). *Journal of Spacecraft and Rockets*, *27*, 113. (Also AIAA Paper 89–0269, January 1989).

Boehly, T. R., Craxton, R. S., Hinterman, T. H., et al. (1995). *Review of Scientific Instruments*, *66*, 508.

Boehm, B. W. (1981). *Software engineering economics* (1st ed.). Upper Saddle River: Prentice Hall PTR.

Bradley, D. K., Delettrez, J. A., Epstein, R., et al. (1998). *Physics of Plasmas*, *5*, 1870.

Calder, A. C. (2005). *Astrophysics and Space Science*, *298*, 25.

Calder, A. C., Hoffman, M. M., Willcox, D. E., et al. (2018). *Journal of Physics: Conference Series*, *1031*, 012016.

Calder, A. C., Krueger, B. K., Jackson, A. P., & Townsley, D. M. (2013). *Frontiers of Physics*, *8*, 168.

Calder, A. C., Taylor, N. T., Antypas, K., Sheeler, D., & Dubey, A. (2006). Numerical modeling of space plasma flows. In G. P. Zank & N. V. Pogorelov (Eds.), *Astronomical Society of the Pacific Conference Series* (Vol. 359, p. 119).

Calder, A. C., et al. (2000). In *Proceedings of Supercomputing 2000*. http://sc2000.org.

Calder, A. C., Fryxell, B., Plewa, T., et al. (2002). *Astrophysical Journal Supplement Series*, *143*, 201.

Calder, A. C., Townsley, D. M., Seitenzahl, I. R., et al. (2007). *Astrophysical Journal*, *656*, 313.

Chandrasekhar, S. (1981). *Hydrodynamic and hydromagnetic stability*. New York: Dover.

Colella, P., & Glaz, H. M. (1985). *Journal of Computational Physics*, *59*, 264.

Colella, P., & Woodward, P. R. (1984). *Journal of Computational Physics*, *54*, 174.

Dubey, A., Antypas, K., Ganapathy, M. K., et al. (2009). *Parallel Computing*, *35*, 512.

Dubey, A., Calder, A. C., Daley, C., et al. (2013). *The International Journal of High Performance Computing Applications*, *27*, 360.

Dubey, A., Antypas, K., Calder, A. C., et al. (2014). *The International Journal of High Performance Computing Applications*, *28*, 225.

Dubey, A., Weide, K., Lee, D., et al. (2015). *Software: Practice and Experience*, *45*, 233.

Dwarkadas, V., Plewa, T., Weirs, G., Tomkins, C., & Marr-Lyon, M. (2005). Simulation of vortex-dominated flows using the FLASH code. In T. Plewa, T. Linde, & V. Gregory Weirs (Eds.) (pp. 527–537). Heidelberg: Springer.

Encyclopaedia Britannica. (2006). https://www.britannica.com/science/eddy-fluid-mechanics.

Fryxell, B., Arnett, D., & Mueller, E. (1991). *Astrophysical Journal*, *367*, 619.

Fryxell, B., Olson, K., Ricker, P., et al. (2000). *Astrophysical Journal Supplement Series*, *131*, 273.

Fryxell, B. A., Müller, E., & Arnett, D. (1989). MPIA Technical Report.

Fureby, C. (1996). *Physics of Fluids*, *8*, 1301.

Gamboa, E. J., Drake, R. P., Falk, K., et al. (2014). *Physics of Plasmas*, *21*, 042701.

Gamboa, E. J., Huntington, C. M., Trantham, M. R., et al. (2012). *Review of Scientific Instruments*, *83*, 10E108.

Godunov, S., Zabrodin, A., & Prokopov, G. (1962). *USSR Computational Mathematics and Mathematical Physics*, *1*, 1187.

Godunov, S. K. (1959). *Mathematics Sbornik*, *47*, 271.

Grinstein, F., Margolin, L., & Rider, W. (2007). *Implicit large eddy simulation: Computing turbulent fluid dynamics*. Cambridge University Press.

Hearn, N. C., Plewa, T., Drake, R. P., & Kuranz, C. (2007). *Astrophysics and Space Science*, *307*, 227.

Hoffman, M. M., Willcox, D. E., Katz, M. P., et al. (2019). *Astrophysical Journal*. (in preparation).

Kane, J. O., Robey, H. F., Remington, B. A., et al. (2001). *Physical Review E*, *63*, 055401.

Kritsuk, A. G., Norman, M. L., & Padoan, P. (2006). *Astrophysical Journal Letters*, *638*, L25.

Larsen, J. T., & Lane, S. M. (1994). *Journal of Quantitative Spectroscopy and Radiative Transfer*, *51*, 179.

Lee, D. (2013). *Journal of Computational Physics*, *243*, 269.

Lee, D., & Deane, A. E. (2009). *Journal of Computational Physics*, *228*, 952.

Lee, D., Faller, H., & Reyes, A. (2017a). *Journal of Computational Physics*, *341*, 230.

Lee, D., Tzeferacos, P., Couch, S., et al. (2017b). *Astrophysical Journal*. (in preparation).

Li, S. (2010). *Journal of Computational and Applied Mathematics*, *233*, 3139. Finite Element Methods in Engineering and Science (FEMTEC 2009).

Li, Z., & Wood, R. (2017). *Journal of Computational and Applied Mathematics*, *318*, 259.

MacNeice, P., Olson, K. M., Mobarry, C., de Fainchtein, R., & Packer, C. (1999). *NASA Technical Reports*, CR-1999-209483.

MacNeice, P., Olson, K. M., Mobarry, C., de Fainchtein, R., & Packer, C. (2000). *Computer Physics Communications*, *126*, 330.

Majda, A. (1984). *Compressible fluid flow and systems of conservation laws in several space variables* (p. 172).

Margolin, L. (2014). *Mechanics Research Communications*, *57*, 10.

Margolin, L. G., & Rider, W. J. (2002). *International Journal for Numerical Methods in Fluids*, *39*, 821.

Margolin, L. G., & Shashkov, M. (2008). *International Journal for Numerical Methods in Fluids*, *56*, 991.

Meshkov, E. E. (1969). *Izvestiya Academic of Science USSR Fluid Dynamics*, *4*, 101.

Miles, B. J., van Rossum, D. R., Townsley, D. M., et al. (2016). *Astrophysical Journal*, *824*, 59.

Mitran, S. M. (2009). Numerical modeling of space plasma flows: ASTRONUM-2008. In N. V. Pogorelov, E. Audit, P. Colella & G. P. Zank (Eds.), *Astronomical Society of the Pacific Conference Series* (Vol. 406, p. 249).

Moore, K., Townsley, D. M., & Bildsten, L. (2013). *Astrophysical Journal*, *776*, 97.

Muller, E., Hillebrandt, W., Orio, M., et al. (1989). *Astronomy and Astrophysics*, *220*, 167.

Oberkampf, W., & Roy, C. (2010). *Verification and validation in scientific computing*. Cambridge University Press.

Plewa, T., Calder, A. C., & Lamb, D. Q. (2004). *Astrophysical Journal Letters*, *612*, L37.

Richtmyer, R. D. (1960). *Communications on Pure and Applied Mathematics*, *13*, 297.

Roache, P. (1998a). *Verification and validation in computational science and engineering*. Hermosa.

Roache, P. J. (1998b). *Fundamentals of computational fluid dynamics*. Albuquerque, USA: Hermosa.

Robey, H. F., Kane, J. O., Remington, B. A., et al. (2001). *Physics of Plasmas*, *8*, 2446.

Rosner, R., Calder, A. C., Dursi, L. J., et al. (2000). *Computing in Science and Engineering*, *2*, 33.

Sharpe, G. J. (1999). *Monthly Notices of the Royal Astronomical Society*, *310*, 1039.

Shu, Q., Ateljevich, E., Schwartz, P. O., & Colella, P. (2017). Verification of an adaptive mesh, embedded boundary model for flood modeling applications.

Soures, J. M., McCrory, R. L., Verdon, C. P., et al. (1996). *Physics of Plasmas*, *3*, 2108.

Stoeckl, C., Fiksel, G., Guy, D., et al. (2012). *Review of Scientific Instruments*, *83*, 033107.

Storer, T. (2017). *ACM Computing Surveys*, *50*(47), 1.

Taylor, G. (1950). *Royal Society of London Proceedings Series A*, *201*, 192.

Timmes, F., Dimonte, G., Kane, J., et al. (2004). *Computing in Science and Engineering*, *6*, 10.

Townsley, D. M., Calder, A. C., Asida, S. M., et al. (2007). *Astrophysical Journal*, *668*, 1118.

Townsley, D. M., Jackson, A. P., Calder, A. C., et al. (2009). *Astrophysical Journal*, *701*, 1582.

Townsley, D. M., Miles, B. J., Timmes, F. X., Calder, A. C., & Brown, E. F. (2016). *Astrophysical Journal Supplement Series*, *225*, 3.

Travaglio, C., Hillebrandt, W., Reinecke, M., & Thielemann, F. -K. (2004). *Astronomy and Astrophysics*, *425*, 1029.

Tzeferacos, P., Fatenejad, M., Flocke, N., et al. (2015). High energy density physics. In *10th International Conference on High Energy Density Laboratory Astrophysics*, (Vol. 17, Part A, p. 24).

van der Holst, B., Tóth, G., Sokolov, I. V., et al. (2011). *Astrophysical Journal Supplement Series*, *194*, 23.

Weirs, G., Dwarkadas, V., Plewa, T., Tomkins, C., & Marr-Lyon, M. (2005a). *Astrophysics and Space Science*, *298*, 341.

Weirs, G., Dwarkadas, V., Plewa, T., Tomkins, C., & Marr-Lyon, M. (2005b). *Validating the flash code: Vortex-dominated flows* (pp. 341–346). Dordrecht: Springer. (G. Kyrala (ed.))

Willcox, D. E., Townsley, D. M., Calder, A. C., Denissenkov, P. A., & Herwig, F. (2016). *Astrophysical Journal*, *832*, 13.

Winsberg, E. (2010). *Science in the age of computer simulation*. University of Chicago Press.

Zhang, J., Messer, O. E. B., Khokhlov, A. M., & Plewa, T. (2007). *Astrophysical Journal*, *656*, 347.

Zhiyin, Y. (2015). *Chinese Journal of Aeronautics*, *28*, 11.

Zingale, M., Timmes, F. X., Fryxell, B., et al. (2001). *Astrophysical Journal Supplement Series*, *133*, 195.

Chapter 29
Validation in Weather Forecasting

Susanne Theis and Michael Baldauf

Abstract Numerical simulations are the core technique in forecasting the weather. The simulation calculates a weather forecast by use of an atmospheric model, which is implemented on a computer. The model itself can be partitioned into various complexity levels, and these can be associated with respective validation concepts. The proper design and implementation of the 'dynamical core' (i.e., partial differential equations and their numerical solver) is tested via comparison to idealized test cases. In a subsequent development step, 'parameterizations' are added, and then the simulation is considered a serious attempt to forecast the weather. The quality of the forecast is estimated by the retrospective comparison between simulation output and observed weather. In addition, a day-specific estimate of forecast uncertainty is derived via 'ensemble forecasting' on a routine basis.

Keywords Weather forecasting · Dynamical core · Parameterizations · Ensemble · Uncertainty

29.1 Introduction

Already in 1950, an early attempt at numerical weather simulation was carried out (Lynch 2008). Today, numerical weather simulation is the core technique in forecasting the weather and is applied on a routine basis. The simulation calculates the time evolution of the atmospheric state (e.g., wind, temperature, humidity, etc.) by use of a model which is implemented on a computer. Due to the high complexity of the model and the time pressure in producing a simulation output, numerical weather

S. Theis (✉) · M. Baldauf
Deutscher Wetterdienst, Offenbach am Main, Germany
e-mail: susanne.theis@dwd.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_29

prediction still challenges the technical limits of supercomputing facilities (Bauer et al. 2015).

Forecasting the weather is an ‘initial value problem’, so the simulation is started from an initial state which has to be specified (Daley 1994). In other words, knowing the weather of today is the starting point of forecasting the weather of tomorrow. The computer simulation heavily depends on timely and high-quality input data, collected by national and international observation systems such as ground stations, radiosondes, radar networks, or satellites. These observations can also be used to validate the outcome of the simulations in a retrospective manner. This is standard practice at very many national and international forecast centers (Casati et al. 2008). Thereby, it can be objectively demonstrated that weather prediction has improved substantially during the past decades (Bauer et al. 2015).

However, the optimization of weather forecasts is not just a matter of statistical fitting to observed weather, for example, by performing numerous forecast trials and minimizing the gap between observed and simulated weather. This becomes clear by looking at the architecture of atmospheric models. They are based on physical conservation laws for mass, momentum, and energy (see the Euler equations (29.1)) combined with physical parameterizations that describe subgrid-scale processes. In other words, the model can be partitioned into the ‘dynamical core’ and the ‘parameterizations’ (e.g., see Bauer et al. 2015; Gramelsberger 2010 and references within). Different validation concepts exist which act on various complexity levels of the model. First, the ‘dynamical core’ is tested in idealized test cases, then ‘parameterizations’ are added. The combined version is used in the attempt to simulate the weather as it occurs in reality. Its output is compared to observations as mentioned above.

Operating the source code on a computer is another integral part of the simulation, because it is essential for deriving results from the source code. For example, regarding the COSMO-DE weather forecasting model, the international COSMO Consortium maintains currently about 350,000 lines of source code, which are used for operational weather forecasting in about 35 different countries. In practice, the source code of the model is constantly changed due to progress in research, technology, and user requirements. From a management perspective, these development and maintenance processes can be validated as well. Even if this is not done in an entirely systematic way, traditions of best practice have emerged. For the sake of brevity, this aspect of quality is not described in this chapter.

In many operational weather forecasting centers, uncertainties of the forecast are addressed. They are estimated along with the prediction and are issued as part of the forecast. To some extent, this can be achieved outside the computer simulation by learning from forecast errors observed in the past and applying this knowledge to the current forecast. However, the nonlinear complexity of the system limits the success of these methods. During the past 25 years, a method called ‘ensemble forecasting’ has become standard practice (Bauer et al. 2015; Leutbecher and Palmer 2008; Parker 2010) (also see Chap. 34 by Knutti et al. in this volume). Sources of uncertainties are addressed within the simulation, multiple simulations are carried out and each simulation output is supposed to represent a reasonable possibility of the atmospheric evolution.

For the sake of clarity, we briefly comment on the differences between weather and climate simulations. Naturally, the simulation of weather is also linked to the basic principles of climate simulations, because the underlying laws of physics apply to both prediction ranges. Climate prediction benefits from a realistic simulation of weather phenomena and their statistics. However, weather and climate predictions do not share the same concepts of predictability, validation, and uncertainty. Whereas climate simulations aim at the statistics of meteorological variables on the basis of many years (see Chap. 30 by Rood in this volume), weather prediction aims at the prediction of specific weather events within the next hours and days. In the context of climate, 'weather' is interpreted as a fast and unpredictable variation around the climatological state. Climate models incorporate additional components, which represent slowly varying processes within the earth system. Compared to weather forecasting, the initial state of 'fast' weather components plays a minor role in climate simulations. In this chapter, we focus on the typical prediction ranges of a weather forecast, i.e., a few hours up to 1–2 weeks.

Section 29.2 provides relevant information about weather forecasting. The architecture of atmospheric models is introduced, and the intended use of the simulation output is explained. Then Sect. 29.3 describes in detail validation concepts, which are targeted at the different complexity levels of the model. In addition, Sect. 29.4 outlines the method of ensemble forecasting. Section 29.5 contains a discussion and a summary.

29.2 Setting the Scene

First, we outline the scientific state of the art of atmospheric models, and we show why and how a model can be partitioned into parts (Sect. 29.2.1). Then, we illustrate how the computer simulation is embedded in the entire process chain of weather forecasting and give an idea about the intended use of the simulation output (Sect. 29.2.2).

29.2.1 The Atmospheric Model: State of the Art

The main ingredients of the atmospheric model are the prognostic equations with their discretization in time and space, both for the 'dynamical core' and the 'parameterizations' (Fig. 29.1). The most relevant input data for a numerical weather prediction is the 'initial state' of the atmosphere. This section provides a brief overview of the model components.

Dynamical Core

The 'dynamical core' consists of a numerical solver for a set of nonlinear partial differential equations that are derived from physical conservation laws for mass,

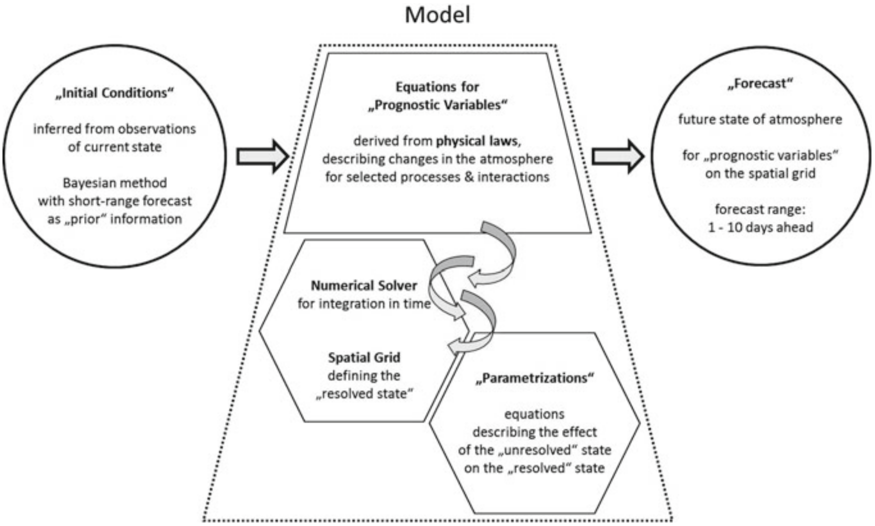

Model
„Initial Conditions“
inferred from observations of current state
Bayesian method with short-range forecast as „prior“ information
Equations for „Prognostic Variables“
derived from physical laws, describing changes in the atmosphere for selected processes & interactions
„Forecast“
future state of atmosphere
for „prognostic variables“ on the spatial grid
forecast range: 1 - 10 days ahead
Numerical Solver for integration in time
Spatial Grid defining the „resolved state“
„Parametrizations“
equations describing the effect of the „unresolved“ state on the „resolved“ state

Fig. 29.1 Key ingredients of the model are placed within the dotted box: prognostic equations with discretization in time and space, and the parameterizations. The curved arrows indicate that these ingredients are interrelated to some extent. The arrows in bold type from left to right illustrate the data flow: the initial state is the starting point, the simulation resembles the time evolution of these conditions in the atmosphere, and thereby produces a forecast

momentum, and energy. These laws lead to the compressible, nonhydrostatic Euler equations, which may be formulated for the prognostic variables density ρ, velocity vector $\mathbf{v}$, and temperature T (alternative formulations are possible):

$$\begin{aligned}\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \mathbf{v} &= 0,\\ \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} &= -\frac{1}{\rho}\nabla p - g\mathbf{e}_z - 2\Omega \times \mathbf{v},\\ \rho c_v \left(\frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T\right) &= -p\,\nabla \cdot \mathbf{v} + Q,\\ p &= R\rho T.\end{aligned} \tag{29.1}$$

Other variables are pressure p and a diabatic heat source term Q. Furthermore, g denotes gravitational acceleration, $\mathbf{e}_z$ is the vertical unit vector, Ω is the angular velocity vector of the earth rotation, c_v is the heat capacity, and R the individual gas constant of dry air.

These Euler equations may be extended by further prognostic equations for water vapor, concentrations of cloud and ice water content, rain and snow, turbulent kinetic energy, or others, depending on the type of physical parameterizations. The parameterizations themselves are expressed as additional terms on the right-hand sides of these equations.

Due to the complexity of these equations, a simple analytical solution is not available. This is why a computer simulation is needed, which derives an approximate solution by employing numerical methods.

The numerical method involves a discretization in time. In other words, the future state of the atmosphere can be estimated from a previous state, by partitioning the entire forecast horizon (e.g., 3 days) into incremental forecasting steps. For example, a weather forecast for the next 3 days involves several thousand of small time steps. These time steps are related first to the numerical stability of the solver and second to the expected accuracy of the numerical solution. Together with the time criticality of the simulation (e.g., at the Deutscher Wetterdienst (DWD) a 24-hour forecast must be ready in about 20 minutes) and the available computing resources, this leads to a trade-off between these various requirements.

Additionally, the numerical method involves a discretization in space. A common technique is the discretization on a three-dimensional spatial grid, so the prognostic variables are defined on prespecified grid points. The density of the grid results from the horizontal distance between neighboring points and the selected thickness of the 'vertical layers'. Today, high-resolution weather forecasting operates with a horizontal grid spacing of a few kilometers. The vertical thickness ranges between a few meters near the surface and several hundred meters at higher levels (e.g., at the 5 km height level). The prognostic equations are numerically solved on the spatial grid. So the computer simulation explicitly resembles atmospheric processes as far as they are represented by the equations and resolved by the grid. The 'resolved state' is defined by the prognostic variables at the grid points.

Parameterizations

The 'resolved state' is accompanied by an 'unresolved state', pertaining to spatial information and processes unresolved by the grid spacing. The discretized set of prognostic equations does not explicitly account for them. However, many physical processes in the atmosphere or at the surface take place on these small spatial scales, for example, the formation of clouds, the interaction between solar radiation and cloud droplets, and the interaction with complex topography. In the real world, these processes can have a substantial effect on the larger spatial scales, especially regarding the heat and momentum budgets. Pragmatic solutions are needed to close the existing gap. These are the so-called 'parameterizations', i.e., additional terms on the right-hand side of Eq. 29.1 and eventually additional prognostic equations.

Conceptual aspects of parameterizations are described in Arakawa (2004), and the most commonly used types are explored in Stensrud (2007). Parameterization aims to obtain a closed system for predicting the weather. The closure consists of quantitative statements about the statistical behavior of the 'unresolved state', in the sense that it feeds back to the 'resolved state' and that it is a function of the 'resolved state'. To some extent, parameterizations embrace physical reasoning, for example, the key role of buoyancy in convection parameterizations (Arakawa 2004). Nevertheless, there is ample scope for various theoretical concepts, which are exchanged and compared within the international meteorological community. It is acknowledged that parameterizations are indispensable in atmospheric models, but also a major source of forecast uncertainties. Especially questionable is the artificial separation of spatial scales, which is not necessarily accompanied by a clear separation of physical processes.

Synthesis

In the practical model development process, the dynamical core is set up first and tested in isolation by applying idealized tests, looking at stability and accuracy of the numerical solution (Sect. 29.3.1). Then the dynamical core is combined with already existing or nearly developed parameterization schemes, which have undergone credibility checks (Sect. 29.3.2). Having combined these model components and having specified an initial state[1] the full-fledged simulation is carried out. Its output is treated as a serious attempt at forecasting the weather as it occurs in reality. The combined model is tested on many different types of weather conditions, by performing numerous simulations and comparing their output to the weather observed in the real world (Sect. 29.3.3). Lessons learned from the validation of the combined version iteratively lead to a fine-tuning[2] of the parameterizations. Justification of the specifics in a parameterization setup remains very limited, but it can be shown whether, or to what extent, the combined and tuned simulation method is fit for weather forecasting.

As a last remark, we comment on the conceptual separation of the respective validation procedures. Obviously, different versions of validation procedures exist, and they pertain to isolated and combined model versions, respectively. However, their clear separation remains ambiguous, because the development and validation procedure has an iterative character in practice. As the separation is formally defined by the prognostic variables on the spatial grid, we can shed further light on the separation issue by describing the practical procedure of reducing the grid spacing. Finer grid spacing enables the dynamical core to explicitly account for processes on smaller scales than before, and the parameterization possibly needs to play a different role. In practice, the need for an adaptation depends on the atmospheric process. For example, if the grid spacing is reduced from 10 to 1 km, the simulation of deep convection needs considerable reformulation, whereas the interaction between solar radiation and cloud droplets is not revised. This is because deep convection takes place on scales of a few kilometers, whereas the origin of the radiation effect is on scales of micrometers. For the described grid refinement, the parameterization of radiation is 'grid invariant'. Under this condition, the iterative character of the development and validation procedure is much less pronounced, due to the weak interrelation between the parameterization design and the explicit simulation on the grid.

[1] Estimating the initial state is a complex and extensive task, called 'data assimilation'. Observations of the current state of the atmosphere enter a Bayesian estimation process, which uses a short-range forecast simulation as prior information. This estimation aims at the specific state of the atmosphere at present time in a manner that is physically consistent and spatially and temporally coherent (see Daley 1994 for more information).

[2] This tuning pertains to the simulation output as an entity. In a subsequent step outside the simulation (cf. Fig. 29.2), specific variables (e.g., temperature at 2 m height) are additionally corrected by statistical post-processing, which is applied separately to each variable.

29.2.2 Intended Use of the Simulation Output

As indicated in Sect. 29.1, atmospheric models are not only used in weather forecasting, but also in other fields such as air pollution scenarios or climate change scenarios. Even within the field of weather forecasting, the intended use is not unique. For example, it plays a role whether the forecast is intended to cover 1 day or 1 week and which kind of weather phenomena it is supposed to represent realistically. Given the intended use, developers of the model make informed decisions about various aspects of the model:

- set of prognostic variables and explicitly simulated processes (limited by the ability to describe the process, …)
- size of covered region, vertical extent of the simulation
- boundary values (the atmospheric state at boundaries of the covered region, sea surface temperatures, etc.)
- density of spatial grid, thickness of vertical layers
- accuracy of numerical solver
- representation of parameterized processes
- speed of computation
- time range of the forecast
- knowledge of an initial state, for all prognostic variables, on the selected grid and layers

Many of these aspects are interrelated and need to be balanced. The balance is also guided by validation, by expert knowledge of scientific principles and atmospheric predictability.

Regarding weather forecasting, applications span a wide range including public warnings (e.g., wind peaks, extreme temperatures, thunderstorms), services for the sector of transport (e.g., road ice, aviation-related forecasting), input for flood forecasts (e.g., severe rain). Some features (e.g., road ice, thunderstorms) are not explicitly issued by the raw simulation (cf. Fig. 29.2).

The computer simulation itself delivers the raw forecast data, which is available on the supercomputer. Then the raw forecast data is extracted, condensed, refined, tailored, and communicated. Statistical and empirical methods and partly also human experts (‘forecasters’) are involved. For example, official weather warnings in Germany are issued by meteorologists working on a 24/7 basis. Some of the products are directly distributed to various customers (‘product delivery’). Some are distributed internally, interpreted by meteorologists and weather advisors, and then communicated to the customers (‘service delivery’). A major intention of the raw simulation is to inform the subsequent process chain optimally.

In general, one would expect that in-house metrics of forecast quality should also reflect the quality criteria of customers. If not, this can be taken as an incentive to intensify the communication between providers and customers, for example, by expert and user feedback. In addition, it is worth mentioning that quality requirements of customers do not only pertain to the diagnosed differences between forecast and

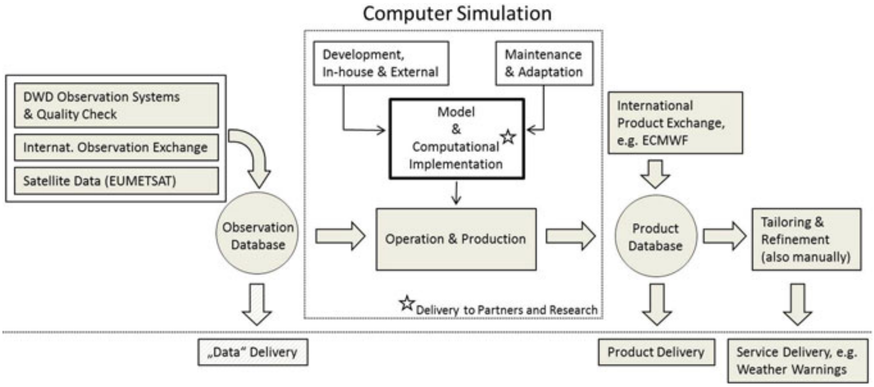

Computer Simulation
Development, In-house & External
Maintenance & Adaptation
DWD Observation Systems & Quality Check
Internat. Observation Exchange
Satellite Data (EUMETSAT)
Model & Computational Implementation ☆
International Product Exchange, e.g. ECMWF
Observation Database
Operation & Production
Product Database
Tailoring & Refinement (also manually)
☆ Delivery to Partners and Research
„Data" Delivery
Product Delivery
Service Delivery, e.g. Weather Warnings

Fig. 29.2 Weather Prediction at DWD, the national meteorological service in Germany. Adapted from the respective quality management handbook (Adrian 2016). The dotted box marks the computer simulation. The darker boxes belong to the time-critical 24/7 suite. The dotted horizontal line marks the deliveries which are also time-critical except the data delivery for long-term monitoring (such as climate). Another general delivery is the source code itself, as marked by the star within the bold box and the corresponding footnote ('delivery to partners and research')

observed weather events, but also to technical features such as timeliness, availability, readability, and to the selection and tailoring of weather information according to the intended use by the customer.

Looking at the intended use from an even broader perspective, weather forecasting ultimately intends to reach various stakeholders and to provide useful information in their decision-making context (Kox and Thieken 2017). When adopting this perspective, a quality assessment may look further down the 'value chain', i.e., simulation output, weather forecast, communication, perception, decision. This is not standard practice, but attempts in this direction have been made (Lazo et al. 2009). In recent years, it has been recognized that the assessment of the 'value chain' benefits from an interdisciplinary approach (e.g., meteorology, social sciences, psychology, economics). This area of research is beyond the scope of this chapter.

29.3 Validation Concepts

As mentioned in previous sections, we describe validation concepts targeted at the simulation output. Section 29.3.1 describes how the dynamical core is compared to idealized test cases, Sect. 29.3.2 very briefly outlines the validation of parameterizations, and Sect. 29.3.3 describes how the output of the full-fledged weather forecast simulation is compared to observed weather.

These different sections deal with very different complexity levels of the model. Accordingly, the various validation concepts are targeted at different expectations, so they highlight different aspects of 'quality' in the simulation output. In addition, various challenges come in different flavors, e.g., covering the formal scope of the model, generating reference data, defining quantitative measures.

In the following sections, the term 'validation' is not always applied. Section 29.3.1 applies the term 'verification' because the proper implementation of equations is tested for a number of cases with known solutions. Section 29.3.3 also applies the term 'verification', simply due to a tradition in the specific field of weather forecasting.

29.3.1 Idealized Tests for the Verification of the Dynamical Core

In the verification of the dynamical core, the proper design and implementation of the numerical solver are tested. In the context of atmospheric simulations, the probably most commonly used tools are idealized test cases. The key ingredients of this verification are *reference solutions* of the above-mentioned equation system (Euler equations or analogous equation systems, see below), which can be delivered for strongly simplified initial and boundary conditions.

A Closer Look at the Dynamical Core

It is commonly accepted that the Navier–Stokes equations for a rotating frame describe the flow of the earth atmosphere with very high accuracy at least as long as no diabatic heating Q (i.e., no phase changes and no radiative absorption or emission) takes place and the height range[3] is roughly below 100 km. As mentioned in the previous sections, phase changes of water, i.e., the formation of cloud droplets, raindrops, ice, and snow crystals and also their interaction with radiation need an extension of the compressible, non-hydrostatic Euler equations (29.1) in the form of several parameterizations. Since numerical simulations for weather and climate prediction have a resolution that is much larger than the Kolmogoroff scale length (about a few millimeters), one replaces the Navier–Stokes equations (that describe *molecular* diffusion) by the Euler equations plus an additional parameterization for turbulent diffusion.

So, besides the parameterizations, the numerical solver of the Euler equations (often called the 'dynamical core') is an important building block of an atmospheric simulation model. Moreover, due to the quite universal meaning of the Euler equations in almost all fluid dynamic applications (i.e., beyond atmospheric dynamics), it is reasonable to verify their proper implementation separated from the parameterizations.

In this context, it should be mentioned that not only in the past but also today simplified equations derived from the Euler equations are still used to formulate a dynamical core. Examples are the hydrostatic approximation (for large-scale flow),

[3] For height ranges above 100 km, the assumption of thermodynamic equilibrium begins to fail, and one needs another stage of gas description like the Boltzmann equation.

several degrees of the anelastic approximation (for smaller scale flow): the Boussinesq approximation, the 'standard' form (Ogura and Phillips 1962), the 'extended' form (Wilhelmson and Ogura 1972), and the pseudo-incompressible approximation (Durran 1998), or the 'unified anelastic and quasi-hydrostatic equations' (Arakawa and Konor 2009).

Reference Solutions
There are several possibilities to derive reference solutions:

- analytic solution under certain assumptions,
- numerical solution by a benchmark solver,
- manufactured solution.

While manufactured solutions are relatively widespread in the fluid dynamics community, this verification method is less often used by meteorological dynamical core developers. In the technique of manufactured solutions, one prescribes a 'solution', i.e., relatively simple but nontrivial fields in space and/or time, inserts them into the equation set and derives 'right-hand sides' of the equations just in a way that this solution exactly fulfills this extended equation system (see Chap. 12 by Roache in this volume). We see two reasons why the method is not used extensively by meteorological dynamical core developers. First, the additional implementation of the artificial 'right-hand sides' may introduce accidental coding errors which remain undetected. Second (and more severe), the artificially constructed solution may be far away from meteorological relevance. In atmospheric dynamics, nearly balanced states (i.e., hydrostatic, geostrophic) are highly relevant. Therefore, analytic solutions of the true Euler equations are often preferred, because they are closer to realistic meteorological flows.

More often used are reference solutions that are derived by a benchmark solver. Imagine that we use a 'highly enough resolved' numerical solution of an already existing 'well-known' and 'highly confidential' numerical solver. As indicated by the apostrophes, the drawback of the benchmark is a certain lack of transparency. However, the general advantage of such kind of reference solutions is their ability to consider relatively complex and strongly nonlinear flows. In any case, one must require that the reference solution has converged, implying that the equation system must contain additional diffusion terms which are also present in the numerical solver. Examples are the falling cold bubble test of Straka et al. (1993), the warm bubble test of Robert (1993), both leading to Kelvin–Helmholtz- or shear-instability phenomena, and for global models the Jablonowski and Williamson (2006) test case of a baroclinic instability. The practical availability of such reference solutions may consist in data files, e.g., containing a standard format like NetCDF or GRIB (the latter recommended by the World Meteorological Organization, WMO) that can be compared directly with the own solution via graphical plotting tools or by calculating error measures. However, in many cases, the developer subjectively compares the figures of a test case publication with his/her figures by 'eye norm'.

As an alternative, one can try to gain an analytic solution for a test case. The charm of this approach is the increased transparency compared to the benchmark approach since the derivation of this solution can be documented and therefore be reviewed by other scientists. From this viewpoint, the use of analytic solutions is probably the conceptually simplest verification approach. Additionally, the (relatively small) source code which calculates the analytic solution can easily be delivered.[4] However, nontrivial analytic solutions are in general hard to achieve, because the Euler equations are very complex. We discuss the following techniques to alleviate this problem: (1) subsets of the original equations, (2) approximated equation sets.

Considering subsets of the original equation set is relatively unproblematic when the solver contains a related 'switch' to mimic this simplification. Examples for those 'switches' are the following:

- the shallow atmosphere approximation, i.e., prefactors $1/r$ (r = radius of spherical coordinates) are replaced by $1/a$ (a = earth radius),
- no orography, i.e., all metric correction terms due to a curved terrain-following coordinate are switched off,
- 'flat earth', i.e., related spherical metric correction terms are switched off,
- no Coriolis term, i.e., the Earth rotation vector Ω or its projection $f = 2\Omega \sin\phi$ is just set to zero in the model.

Further subsets involve the consideration of 'advection' terms only, the consideration for sound expansion only, the assumption of stationarity, and tests of very small code parts. As a side remark, some of these groupings into subsets are connected to the use of symmetries for the solution of the equations (e.g., time invariance or rotational invariance).

Let us consider the subset of the (multidimensional) advection terms[5] for a scalar ϕ

$$\frac{\partial \phi}{\partial t} + \mathbf{v} \cdot \nabla \phi = 0. \tag{29.2}$$

in more detail. The motivation for considering this relatively small subset of the Euler equations is twofold. First, the advection process is very important for atmospheric flows and takes a significant part of computation time. Second, idealized tests with *exact analytic* solutions are well known. For example, for a given constant velocity field $\mathbf{v}(\mathbf{r}, t) = const$ any initial distribution $\phi(\mathbf{r}, t = 0) = \phi_0(\mathbf{r})$ fulfills the advection equation for any later time t by $\phi(\mathbf{r}, t) = \phi_0(\mathbf{r} - \mathbf{v}t)$. More complicated exact solutions can be constructed by superposition principles, symmetry arguments in higher dimensions, or by the method of characteristics for general first-order partial differential equations. One quickly recognizes that the choice of $\mathbf{v}$ relative to the coordinate axes of the model has an impact on how well the model simulation agrees

[4] In this context, we probably need to elucidate the term 'analytic' which may be beyond elementary functions like sines or logarithms. The term 'analytic' refers to a solution that is given as a series or in the form of integrals, and its calculation may require numerical methods, too. However, the calculation of integrals or series is, in general, a much more robust operation compared to the numerical solution of differential equations.

[5] In other fluid dynamic areas, these terms are called convective terms. Since the term 'convection' is otherwise used, 'advection' is preferred in the meteorological literature.

with this exact solution. Therefore, many model developers prefer the ‘solid-body rotation test’ with a prescribed velocity field $\mathbf{v}(\mathbf{r}) = \mathbf{k} \times \mathbf{r}$ around an axis $\mathbf{k}$. Again, the exact analytic solution of the advection problem is an appropriate rotation of the initial solution ϕ_0 around the same axis. Also, tests with an instationary velocity field have been proposed (LeVeque 1996), and a superposition of this field with a solid-body rotation on the sphere is used (Lauritzen and Thuburn 2012).

The subset of terms responsible for sound expansion (for the compressible Euler equations) is another option, because simple analytic solutions exist in an isothermal gas (so-called N-waves in 3D problems). The assumption of stationarity, i.e., neglecting all time derivatives may also be useful in finding solutions, for example, Staniforth and White (2007) present exact steady, axisymmetric, balanced zonal solutions for the nonlinear Euler equations.

The most simple case of subsets are single parts of the numerical solver like the divergence operator or the curl operator. Analytic solutions are easily available. For such small code parts the denomination ‘unit test’ may be used: the entry point of a test routine needs relatively little information (often only one field) and the result field is also easily available. Still, for the assessment of a successful passing of this unit test one has to allow tolerance ranges, a problem that is discussed below.

An analytic solution can sometimes be derived from an approximated equation set in the sense of equation sets mentioned above (often in combination with some of the ‘switchable’ simplifications listed above). Examples are the Boussinesq-approximated linear solution of wave expansion in a channel by Skamarock and Klemp (1992) or the also Boussinesq-approximated solution of nonlinear flow over a mountain by Long (1953). Nevertheless, it is still quite difficult to find analytic solutions for these equation sets. Moreover, a general drawback of such solutions lies in the fact that it can be hard to assess if a deviation toward the solution of the dynamical core under inspection is induced by an error or by the approximation; they are ‘un-controlled’ approximations.

A certain exception is the method of linearization around a basic state or perturbation theory in a broader sense. This approach allows to derive analytic solutions almost in a recipe like manner. Additionally, it has the advantage to be a *controlled* approximation. In other words, by reducing the perturbation parameter, higher order terms become increasingly unimportant, and the approximated solution converges to the true one. Examples are solutions of the wave expansion in channels of the *un-approximated*, flat Euler equations by (Baldauf and Brdar 2013) for regional models (see also Fig. 29.3) and of shallow atmosphere approximated equations (Baldauf et al. 2014) for global models.

Comparing the Reference Solution with the Solution of the Simulation

After having identified a reference solution, it must be compared to the solution of the simulation properly. This can be done via

- ‘eye norm’,
- convergence criteria,
- and additionally by checks of global conservation properties.

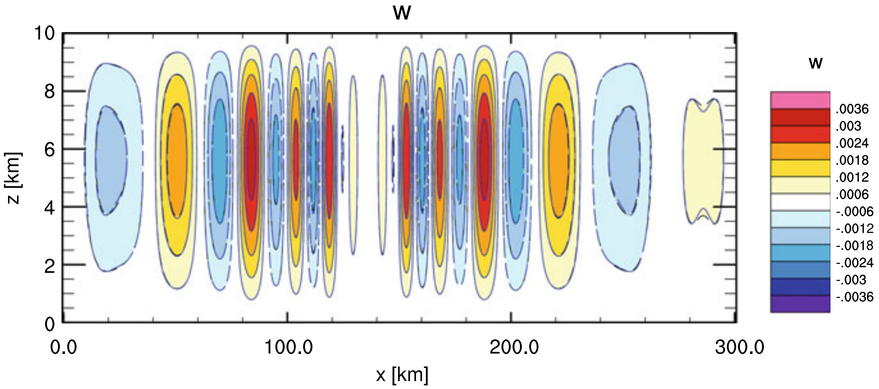


Fig. 29.3 Comparison of the vertical velocity w between a simulation with the regional model COSMO (colors and dashed lines) with a horizontal resolution of dx = 250 m and the analytic solution of Baldauf and Brdar (2013) (lines)

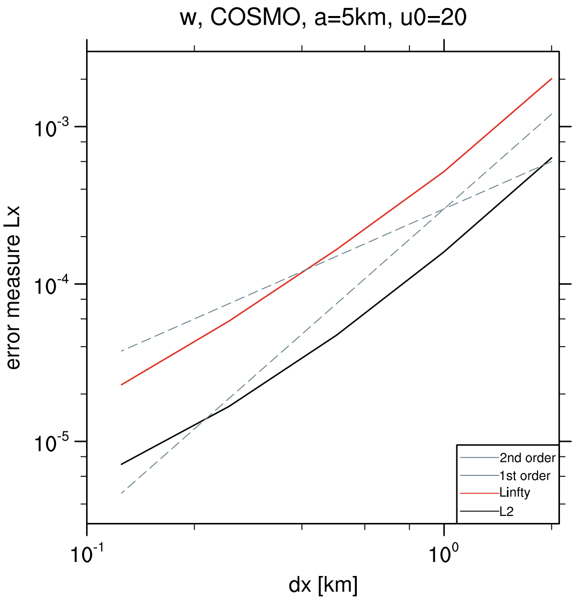
Comparison by 'eye norm' is based on 1D line plots, 2D field plots, or 2D cross-sections of 3D fields. Since the simulated solution will deviate in any case from the exact reference solution, some tolerance must be accepted (e.g., Fig. 29.3). In many cases, an experienced developer subjectively decides whether a particular test has been successfully passed.

Convergence criteria are much less subjective, but only applicable when the reference solution is exactly known (i.e., in the sense of the reduced equation subsets described above). In this case, the simulated solution must *converge* to the reference solution when the spatial and temporal grid mesh size is gradually reduced. Ideally, this convergence rate (it must be at least of first order for a consistent numerical scheme) is known from the construction of the numerical scheme itself and can be measured by numerical experiments in running the model with different spatial and temporal resolutions. In practice, error measures are determined (most often L_2 or L_∞-error) between the exact reference solution and the simulated solution for several temporal and spatial resolutions. The verification procedure checks whether the expected convergence occurs (e.g., Fig. 29.4).

Additionally, checks of general (most often global) properties are carried out. A simple example is again the subset of the advection terms. Advection with an arbitrary divergence-free velocity field has at least the following two properties: an initially constant field $\phi_0(\mathbf{r}) = const$ should remain constant, and for arbitrary initial fields the total mass should not change. Whereas any semi-Lagrangian solver easily fulfills the first property, it has problems with the second one. Inversely any finite-volume solver by construction fulfills the second property but has problems with the first one. Similar checks can be done for global integrals of other variables than mass, if they are globally conserved.

Many of the idealized tests mentioned before have time-dependent solutions and therefore depend on the initial state. Whereas the initial state in an analytical

Fig. 29.4 Same test case as in Fig. 29.3: error measures L_2 (black line) and L_∞ (red line) for the numerical solution against the analytic solution given in Baldauf and Brdar (2013). The two dashed lines denote first- and second-order convergence, respectively; one recognizes a convergence rate of roughly 1.5



reference solution is well defined, its numerical implementation is less well defined. That already a proper initialization can be a principal problem with idealized tests is illustrated by the following example. For the Euler equations, some test cases assume a hydrostatically balanced atmosphere which afterward is selectively disturbed, for example, by warm or cold bubbles. To establish a reasonable test setup, it is necessary to achieve the hydrostatic balance also numerically. In other words, the initial pressure and density fields must be close to the initial analytic state *without* inducing vertical accelerations just by the special numerical treatment of pressure gradient or buoyancy terms. For most of the dynamical core formulations using a terrain-following coordinate, such an exact balance cannot be fulfilled numerically in hilly terrain. Consequently, evaluation of the test results have to deal with the question of whether a difference between two numerical methods lies in the numerics or just in differences of the initial balancing procedures.

Recommended Collection of Tests

To sum up: The previous paragraphs have explained how the dynamical core may be verified by idealized test cases. Such a verification procedure is often complex and time-consuming. Consequently, in practice, a specific selection of tests is carried out for a new model development project. The selection should be generally accepted, sufficiently small, sufficiently simple to carry out and interpret, and sufficiently demanding for the dynamical core. Such a collection of tests has been compiled for global atmospheric simulations, see, for example, the Dynamical Core Model Intercomparison Project (DCMIP).[6] For regional atmospheric simulations,

[6] www.earthsystemcog.org/projects/dcmip.

this has also been achieved, for example, during the several 'SRNWP-workshops on nonhydrostatic modeling'.[7] Similar collections can be found in the literature (e.g., Giraldo and Restelli 2008).

29.3.2 Validation of Parameterizations

The basic concept of a particular parameterization scheme (e.g., deep convection, turbulence) usually results from physical reasoning, intensive observation campaigns, and special simulations which are usually higher resolved (e.g., Randall et al. 2003).

Regarding observation campaigns in the field, diagnostic studies are a common procedure, e.g., Yanai and Johnson (1993). The spatial density of the observation network corresponds to the 'large scale', i.e., the large-scale budgets are observed. These can also be diagnosed based on the prognostic equations, so the respective residuals are interpreted as the effect of the 'unresolved state'.

Furthermore, observations can be used to drive a so-called 'single-column' simulation. This method is based on the notion that parameterizations mainly describe processes acting along the vertical axis and only rarely describe interactions between neighboring grid points in the horizontal directions. In a single-column simulation, the full-fledged atmospheric simulation is reduced to a single grid point in horizontal space and still extends into the vertical. The horizontal dynamical flow is prescribed by observations. Then the result of this simulation is evaluated by additional observations, assessing the capability of the parameterizations to provide realistic forecasts.

Apart from the single-column simulation, another type of special simulation can be applied to developing the basic concept of a parameterization. These are atmospheric simulations with an extraordinarily fine grid spacing, the so-called 'cloud-system-resolving' simulations or 'large eddy' simulations. Their computational costs are prohibitively large for time-critical forecasting, but the simulations can produce highly resolved synthetic data sets. In contrast to 'coarser' simulations, their output contains individual cloud elements in convective systems and individual large eddies in turbulent flow. The forecast of these elements are not credible in a deterministic sense, but a statistical analysis becomes possible, especially because the individual elements cover a sufficiently wide range of time and space scales. The statistical analysis helps in understanding the underlying physical processes and in testing the parameterizations within the 'coarse-grid' simulations of time-critical weather forecasting (Randall et al. 2003). Obviously, the synthetic data sets are associated with some caveats. They do not entirely rely on first principles, but also contain parameterizations pertaining to their 'unresolved state'. Observations collected in special field programs are used to 'certify' that a high-resolution simulation is a reliable tool for the simulation of a particular regime, e.g., Heinze et al. (2017).

[7] http://www2.mmm.ucar.edu/projects/srnwp_tests/index.html.

29.3.3 Comparison to Observations

While the previous sections have looked at the model components in isolation, this section focuses on the simulation output of the comprehensive model including all ingredients (cf. Fig. 29.1). The output of this kind of simulation is a serious attempt at forecasting the weather as it occurs in reality.

Weather forecasting is in the lucky situation that the target of the forecast (e.g., the weather in 1–10 days) can be observed in due time so that success and failure of the forecast are known to a reasonable extent. Since several decades, very many national and international weather forecasting centers estimate the quality of their weather forecasts quantitatively and systematically. The results show that weather forecasts have improved substantially during the past 50 years (cf. Fig. 29.5 and also Bauer et al. 2015).

A systematic comparison between simulation output and 'observational data' is usually based on a number of weather forecasts, usually taken from a series of subsequent days (e.g., several weeks or months) with a weather forecast started at each of these days. In the terminology of the meteorological community, this retrospective comparison is referred to as 'verification'.

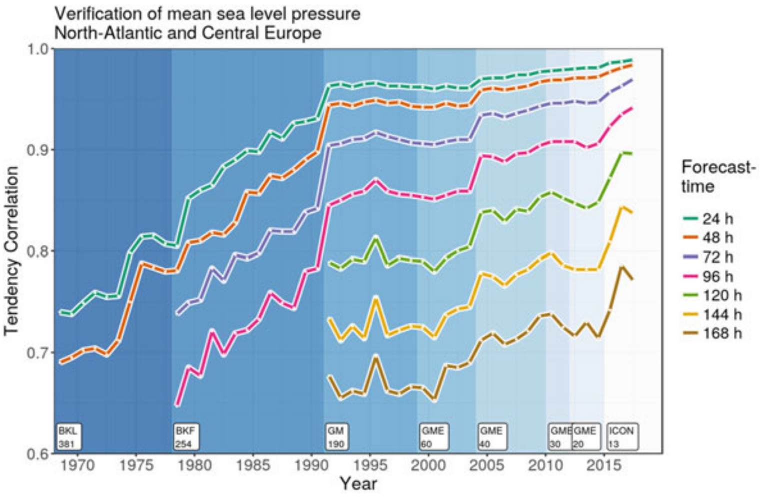


Fig. 29.5 Improvement of weather forecasts at DWD over the past 50 years. Meteorological variable: Mean sea level pressure. Measure for quantifying the difference between simulation output and the model analysis (i.e., the closest approach to observations): Tendency correlation coefficient, i.e., the correlation of anomalies. In this case, forecasts and analysis were corrected by the mean sea level pressure at forecast initialization. Global averages of the coefficient were obtained by arithmetically averaging correlations at each 1.5° × 1.5° area within the verification region. Yearly correlations were obtained by averaging daily correlations. Model which was used to produce the simulation output: the weather prediction model at DWD; type and version which was operational at that time. Since 1991, these models cover the globe. This verification study refers to a region confined to the North Atlantic and Mid Europe. Adapted from the annual report by DWD (DWD Jahresbericht 2016)

When used in an interdisciplinary context, the term ‘verification’ can be misleading. First, observations are certainly not considered as ‘perfect’ or ‘true’. Second, the comparison to observations is not considered as the overarching and sufficient quality criterion when setting up an atmospheric simulation (cf. Sects. 29.2.1 and 29.3.1). And last but not least, the meteorological notion of ‘verification’ is by no way related to procedures of ‘source code verification’. Therefore, some meteorologists prefer the term ‘evaluation’ or ‘validation’, but the tradition of using the term ‘verification’ is very established and still predominant.

A description of common practice in ‘verification’ can be found in Wilks (2011), Jolliffe and Stephenson (2011), or Casati et al. (2008). The design of a systematic comparison between simulation output and observations involves the following ingredients:

- identify available observations, including their quality control,
- pairwise matching of forecast value and observed value,
- identify a forecast aspect of interest given an application and/or the development process,
- select quantitative measures or graphics related to the relevant aspect,
- identify a benchmark forecast or a reference level of skill.

Observations are available due to a national and international observation network, including ground stations, radiosondes, ship and aircraft measurements, buoys, satellites, and radar. For example, Germany has an observing network consisting of 180 ground stations (plus 1800 stations working in an honorary capacity), 10 radiosonde stations, and 17 weather radars. In addition, measurements are taken by hundreds of ships and aircrafts, and the European weather satellite Meteosat. This is embedded in an international data exchange, so DWD also receives the data acquired worldwide.

On the one hand, the observation network is used for deriving an initial state for the time-critical prediction process, and, on the other hand, for the retrospective quality control of the forecast. In principle (and sometimes in practice), the initial state specification (i.e., the so-called ‘analysis’) could be also used in the retrospective quality assessment of forecasts, but this comes with severe caveats because the computer simulation itself already plays a key role within the specification of its initial state (Casati et al. 2008).

Quality control of observations consists in the elimination of very unlikely values, the correction of known errors (e.g., due to instrument limits or the observation site), and the estimation of uncertainty. The last point, the estimation of uncertainties, is still often neglected in validation. However, in recent years, forecast accuracy has become better and better, so the weather forecasting community is now starting to tackle this question as well (e.g., Bowler 2008).

The traditional way of comparing the simulation output with observations matches the time and location of an observation with the time and location of a forecast value available on the model grid. As a result of the matching procedure, a set of observation–forecast pairs becomes available. Various matching procedures exist, e.g., nearest point, bilinear interpolation. Characteristics of the simulation setup such as grid size and the grid-averaged orography are usually taken into account. Furthermore, it is beneficial if the observation is representative for an area as opposed to a very specific point (otherwise, the so-called representativeness errors arise). Concerning ground stations, this is partly attained by very thoughtful placement of observing sites. With decreasing grid size of the simulation, the representativeness error becomes less and less relevant. Recently, more indirect observations, mostly by remote sensing observations from radar (microwave reflectivity), lidar (backscatter signals) or satellites (e.g., photographs of the atmosphere in different electromagnetic wavelength windows) are compared with equivalent forecast products which are derived by so-called 'forward operators' within the simulation. These operators produce synthetic versions of radar, lidar and satellite fields which are compatible with the atmospheric state as seen by the simulation.

Theoretically, the comparison between observations and simulation output could look at a myriad of variables, locations, quality aspects. Unfortunately, and to a large extent, options are already reduced by the number and types of available observations. However, some degree of reduction is also desirable to highlight 'relevant' forecast aspects. The identification of 'relevant' aspects may be guided by forecast applications or by research questions posed by developers. The reduction of aspects (either desired for some reason or dictated by the observation network) may lead to a focus on specific variables (e.g., temperature, precipitation), vertical levels (e.g., at the surface, in 100 m height), specific regions (e.g., over complex orography, over land/sea), a specific range of values (e.g., close to freezing point, severe rainfall amounts). It may be of interest to aggregate the values (e.g., minimum during the day, the sum over a hydrological catchment area) or to look for specific patterns (e.g., a sudden change in time). And last but not least, it can be very helpful to specifically look at certain conditions (e.g., a specific season, short or long forecast lead times, convective situations).

Given the applications, forecast aspects are of interest which eventually affects a decision-making process in the subsequent process chain (e.g., the forecaster decides to issue a weather warning, leading to protective action by stakeholders). Because of the development process, it is desirable to find and eliminate the source of potential errors which propagate through the simulation and eventually affect an application. For example, errors at 5 km height are certainly of interest, because they may cause errors near the surface. However, due to the high complexity and nonlinearity, it can be very challenging to attribute the shortcomings of a forecast to a particular flaw in the model, e.g., to particular weakness in a parameterization scheme or a missing physical process. This is why experts sometimes start with a subjective assessment of a few selected cases to identify certain aspects which are then assessed more objectively.

For the 'objective' comparison between simulation output and observations, a number of quantitative measures are available, either in the form of graphics or 'scores' (see, for example, Jolliffe and Stephenson 2011 and Wilks 2011). It is common practice to apply many scores rather than just one or two. They can be categorized by looking at the formulation of the forecast first. The forecast can be formulated as (1) a continuous variable (e.g., the temperature value), (2) a binary event (e.g., temperature below freezing point: yes or no), (3) a probabilistic information (e.g., probability of the event ‘below freezing point’: 0–100%). More information on probabilistic forecasting is provided in Sect. 29.4.

Some measures of forecast quality can be found in the standard literature of statistics, especially those for continuous variables. A very obvious example is the sum of pairwise differences between the previously matched observations and forecasts (i.e., the forecast ‘bias’). Further obvious examples are the mean square error (see example in Fig. 29.6) and the Pearson correlation coefficient (see Chap. 18 by Saam in this volume). In addition, there is a multitude of measures which are confined to

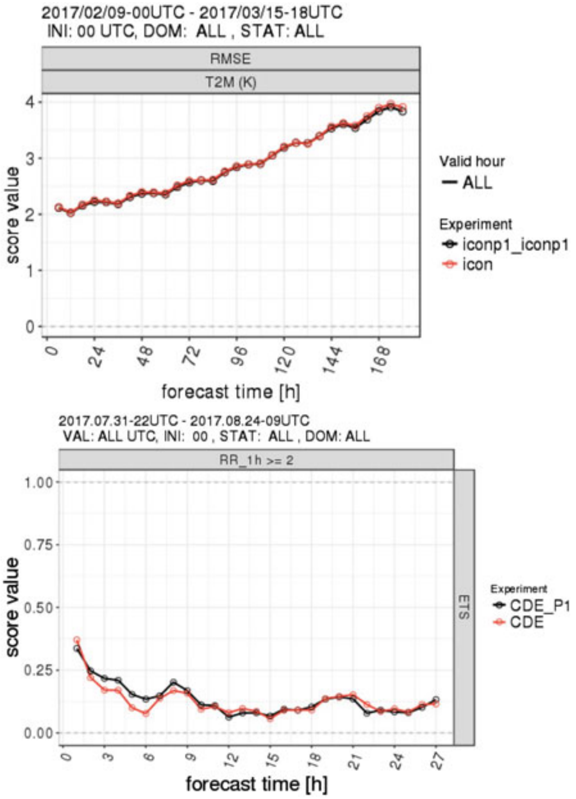


Fig. 29.6 Examples of verification results for a continuous variable (top) and a binary event (bottom). Top: The results refer to the forecast variable ‘temperature in 2 m height’. Forecast quality is estimated by the root mean square error, calculated at each observation site and averaged spatially. The match between forecast and observation includes horizontal bilinear interpolation of three model grid points and a correction due to their deviations in altitude. Verification results are shown as a function of forecast time for two versions of the forecast model ICON. Bottom: The results refer to the binary event ‘precipitation sum (accumulated over 1 h) greater than 2 mm: yes/no’. The measure of forecast quality is the ‘Equitable Threat Score’. Verification results are shown as a function of forecast time for two versions of the regional forecast model COSMO

Table 29.1 The contingency table summarizes the bivariate sample of observed/forecast events

	Observed YES	Observed NO
Forecast YES	Hits	False alarms
Forecast NO	Misses	Correct rejections

the meteorological community, especially the scores for binary events (e.g., Heidke Skill Score, Equitable Threat Score, see example in Fig. 29.6). Many of these measures summarize the bivariate sample of observed/forecast events as presented in the 'contingency table' (Table 29.1).

Some measures directly relate to estimated properties of the underlying joint, marginal, and/or conditional distribution (e.g., probability that the event is forecast under the condition that it is observed). Probabilistic forecasts are rated by yet another set of measures which typically look at 'reliability', 'discrimination' or 'resolution', 'sharpness'. Typical limitations of various scores are also well known (e.g., Gaussian assumption, spurious correlations due to trends or the diurnal cycle) and often the question is posed whether a particular score has desirable characteristics, for example being 'proper' which prevents hedging (Bröcker and Smith 2007).

Once the forecast has been compared to observations via some quantitative measure, the estimated quality of the forecast is very often related to the respective quality of a benchmark forecast. Typical choices of benchmark forecasts are the output of a slightly different simulation method, the observed weather of the previous day ('persistence forecast'), or a random draw from climatological conditions. This way, a reference level of 'forecast skill' is identified.

Taking another look at the matching procedure described above, there is a pitfall called the 'double penalty problem'. Imagine that simulation A can capture short wavelengths of a spatial field (e.g., a small-scale trough in the pressure field) and that these short wavelengths are also present in the observations, but with a slight phase shift compared to simulation A. The standard matching of forecast and observation pairs shows that simulation A predicts the low-pressure event where it was not observed, and it does not predict the event where it was observed. In a nutshell, the forecast of simulation A is wrong at all locations. Imagine that the benchmark simulation B, on the contrary, presents a spatially smoother forecast without any low-pressure events. Then the benchmark simulation B may be rated as the better forecast because it is correct at those locations where the low-pressure event was not observed. From a certain perspective, this result is counterintuitive, so it serves as the motivation for another set of measures, the so-called spatial methods. They can be classified into scale-separation, field-deformation, feature-based, and neighborhood methods (Gilleland et al. 2009; Ebert 2009).

Another critical point is the sampling issue, which exists in several flavors. One is spatial under-sampling associated with those observation types which cover only few locations compared to the vast number of grid points of the simulation (e.g., ground stations, radiosondes). Another one is the number and selection of forecast days that are contained in the verification period. When two simulation types are compared in terms of their measured forecast quality (e.g., Fig. 29.6), the statistical significance of the estimated differences can be of interest (e.g., attained by bootstrapping or significance tests). Reaching statistical significance may require a very thoughtful design of the simulation experiment, considering computational costs (i.e., number of days) and selection of days (e.g., representing a specific weather regime or season). Furthermore, if the improvements are targeted at rare weather events, the detection of forecast improvements faces additional challenges and pitfalls (e.g., Lerch et al. 2017). Statistical significance is harder to reach and, as another challenge, many traditional verification scores are not suitable which has led to the development of novel measures (e.g., Ferro and Stephenson 2011).

29.4 Uncertainty Estimation via Ensemble Forecasting

As with many other predictions, the prediction of weather is prone to errors and uncertainties. Especially in weather forecasting, the ‘perfect’ computer simulation is principally out of reach. This is not only a flaw of the computer simulation itself but also a characteristic of the real atmosphere which is a classic example of a so-called ‘chaotic system’ (cf. Bauer et al. 2015 and the original paper by Lorenz 1963). Tiny errors in the initial state grow rapidly during the time evolution of the atmosphere. The predictive power of the initial state vanishes when the forecast looks further and further into the future. Although its underlying equations are deterministic, the system appears ‘unpredictable’, in the sense that it is extremely sensitive to details in the initial state which cannot be specified with such tremendous accuracy.

In theory, the problem of simulating the uncertainty can be posed regarding equations describing the time evolution of the probability density function of the atmospheric state vector. Such equations are known: the Liouville equation for the growth of initial uncertainty, or a form of Fokker–Planck equation if uncertainties in the process formulation are also taken into account (Ehrendorfer 1997).

In practice, however, the solution of these equations is hampered by the large dimensionality of the atmospheric system. Instead, it is possible to attain a pragmatic ‘Monte Carlo sampling’ of the phase space of the future atmospheric state (Leutbecher and Palmer 2008). Initial condition uncertainty is addressed by producing an initial condition sample and by starting a deterministic simulation from each element of the sample. As a result, an ‘ensemble’ of forecasts is available. Each scenario, a so-called ‘member’ of the ensemble, is supposed to represent a reasonable possibility of the atmospheric evolution. Typically, such an ensemble comprises 10–50 members and the ensemble members move apart with increasing forecast time. They convey an idea of the day-specific predictability of the weather, and they provide the basis for probabilistic forecasts (e.g., 5% probability that temperature is below freezing point). The translation from ensemble members into a probability forecast may involve additional information sources (e.g., Bröcker and Smith 2008).

In addition, uncertainties in the representation of parameterized processes (cf. Sect. 29.2.1) are often addressed with explicit perturbations in the formulation of the parameterization. An ideal design would involve perturbations of all simulation ingredients which are subject to uncertainty as far as it is known. An all-embracing implementation of this vision is not common practice but remains a guiding idea.

Ensemble forecasts and the resulting probabilistic forecasts also undergo the standard quality assessment described in Sect. 29.3.3. As expected, even the simulation of forecast uncertainty is not perfect. For example, a forecast of '30% chance of rain tomorrow' can be compared with observations by collecting many of such '30% forecasts'. Then it can be quantified in how many of these cases the observation detected rain versus the number of cases with 'no rain'. Many further quality criteria can be checked, such as the ability to discriminate between observed forecast errors given a high or low forecast confidence. More information on the comparison of probabilistic forecasts to observations can be found in Jolliffe and Stephenson (2011) and Wilks (2011).

There certainly are limitations of the ensemble method. First, not every origin of forecast uncertainty is known and therefore the ensemble method cannot be designed to represent them properly. Second, the realistic simulation of error growth can be hampered by imperfections in the model itself. Third, the number of members in the ensemble is limited by the available computer power or, in other words, has to be traded against the intrinsic complexity of the deterministic simulation method. This can result in a dilemma, for example, regarding convective precipitation. On the one hand, the representation of convective processes benefits from a very dense spatial grid, because convection takes place on small scales. On the other hand, convection is very prone to chaotic behavior, so the ensemble approach is vital and it requires many members to capture also the very hazardous events associated with low probabilities. To some extent, limitations of the ensemble simulation output can be alleviated within the subsequent process chain (Fig. 29.2). Human experts or statistical post-processing schemes can partly account for those deficiencies which are known from experience or past data (e.g., Wilks and Hamill 2007). When assessing the quality of the ensemble approach, a pragmatic benchmark can be generated by collecting several deterministic weather forecasts (e.g., Hamill 2012). These are available when several weather agencies produce forecasts for the same time range and overlapping forecast regions.

Another active area of research is the practical use and relevance of uncertainty information, for example, in the formulation of weather warnings and in decision-making by stakeholders (e.g., see Kox et al. 2015 and references within). Respective validation concepts are beyond the scope of this chapter.

29.5 Discussion and Summary

Weather forecasting is in the fortunate situation that the output of weather simulations can be compared to observations in a retrospective manner. In the community of weather forecasting, this kind of quality assessment is traditionally called 'verification' which may lead to confusion in interdisciplinary dialogues. The comparison between simulation output and observations is recognized as a very fruitful method, but challenges remain and are subject of current debates:

- Due to the tremendous size of the simulation output (i.e., including all grid points, levels, and variables), the observation network is certainly not able to cover the entire simulation output. Therefore, quality can be assessed to some extent, but this may not match questions arising from the intended use or posed by developers.
- The match between observations and forecasts remains difficult, e.g., due to low representativeness and quality of observations, or slight shifts in location and altitude. The general means to tackle this problem is the use of forward operators.
- Furthermore, expert reasoning is required regarding the number and selection of days within a verification period. For example, the period can be too short (i.e., lack of statistical significance), or it can be too diverse (i.e., improvements for a specific weather situation are obscured).
- Further challenges arise when assessing the ability to forecast extreme and rare weather events.
- In addition, the tendency to use finer and finer spatial grids in weather simulations has introduced the 'double penalty' problem. For example, it is difficult to compare the quality of fine-grid simulations with the quality of coarse-grid simulations.
- The quality of a forecast can be rated according to a multitude of aspects, and each of them can be quantified by many different measures. This leads to a variety of results which require interpretation and priorities, especially when there is a practical need to identify the 'superior' model version.

In terms of uncertainty estimation, a method called 'ensemble forecasting' has made its way into operational centers during the past 25 years. Multiple simulations are started and result in an ensemble of possible outcomes which aim at representing the predictability of the day. The known sources of forecast uncertainty are addressed by sampling from a realistic range of options, which feed into the respective simulations. Due to the chaotic nature of the atmosphere, initial conditions are recognized as a major source of uncertainty, together with the imperfections of the model. The outcome of the ensemble forecast is translated into probabilistic forecasts and their quality can be assessed in a statistical manner. Ensemble forecasting is well established today, but challenges remain. Active debate and research arises from several questions:

- Sources of forecast uncertainty are not known to a full extent.
- A realistic simulation of error growth may be hampered by imperfections of the model, so the representation of forecast uncertainties is a challenge.
- As the atmospheric system is high dimensional and nonlinear, a considerable number of ensemble members is required. Due to limited computing resources, this requires a thoughtful compromise regarding other requirements such as the complexity of the model.

The comparison between forecasts and observations does not only inform users of the forecast but also the developers, who wish to translate the diagnosed shortcomings into improvements of the model. In this respect, the architecture of the atmospheric model plays a role, especially its partitioning into the ‘dynamical core’ and ‘parameterizations’. The proper implementation of the ‘dynamical core’ is usually tested via idealized test cases, for example, by deriving analytic solutions for equation subsets, or by linearizations around a basic state. This kind of verification is also well established, and some aspects are typically under discussion:

- In practice, the number of tests needs to be manageable. Expert reasoning is required to agree on a selection, typically balancing the following criteria: demanding for the dynamical core, meteorologically relevant, credible reference solution, sufficiently simple to interpret and carry out.
- The verification of a ‘dynamical core’ checks for several properties, so developers need to set priorities in their requirements.
- Directly comparing the quality of different ‘dynamical core’ formulations can be problematic, as some fulfill a specific check by definition.

After setting up the ‘dynamical core’, ‘parameterizations’ are added which represent the ‘unresolved state’ of the atmosphere. Their physical basis is much weaker and they are recognized as a relevant source of forecast uncertainty, but their role in the atmospheric model is indispensable. To a limited extent, evidence for a particular formulation is gathered by special observation campaigns and special simulation experiments. However, their ultimate eligibility and their fine-tuning only come into reach after putting the entire simulation together and testing its output against observations as described above. When shortcomings become visible, it can still be very difficult to translate them into improvements of the model, because the atmosphere is a high-dimensional and nonlinear system. Typically, the improvement and the assessment of the simulation method have an iterative character. Depending on the degree of ‘grid invariance’, the separation between the ‘resolved’ and the ‘unresolved state’ is recognized as an unsatisfying compromise and accompanied by a lively debate.

As a summary, validation of the simulation output and uncertainty estimation are well established in weather forecasting for many years. Some techniques and questions may be similar in other disciplines, especially when dealing with nonlinear partial differential equations, data in high-dimensional spaces, chaotic systems.

Acknowledgements We thank Felix Fundel and Ulrich Damrath (DWD) for delivering several verification plots. We also thank Nicole J. Saam and Claus Beisbart for inviting Susanne Theis to the Symposium ‘How to Build Trust in Computer Simulations’ (2015), funded by *VolkswagenStiftung*.

References

Adrian, G. (2016). *Qualitätsmangement-Handbuch für den Deutschen Wetterdienst (DWD)*. Deutscher Wetterdienst, Offenbach, Germany. Version 52.

Arakawa, A. (2004). The cumulus parameterization problem: Past, present, and future. *Journal of Climate*, *17*(13), 2493–2525.

Arakawa, A., & Konor, C. S. (2009). Unification of the anelastic and quasi-hydrostatic systems of equations. *Monthly Weather Review*, *137*(2), 710–726.

Baldauf, M., & Brdar, S. (2013). An analytic solution for linear gravity waves in a channel as a test for numerical models using the non-hydrostatic, compressible Euler equations. *Quarterly Journal of the Royal Meteorological Society*, *139*, 1977–1989.

Baldauf, M., Reinert, D., & Zängl, G. (2014). An analytical solution for linear gravity and sound waves on the sphere as a test for compressible, non-hydrostatic numerical models. *Quarterly Journal of the Royal Meteorological Society*, *140*, 1974–1985.

Bauer, P., Thorpe, I., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*, 47–55.

Bowler, N. E. (2008). Accounting for the effect of observation errors on verification of MOGREPS. *Meteorological Applications*, *15*(1), 199–205.

Bröcker, J., & Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, *22*(2), 382–388.

Bröcker, J., & Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus A: Dynamic Meteorology and Oceanography*, *60*(4).

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., et al. (2008). Forecast verification: Current status and future directions. *Meteorological Applications*, *15*(1), 3–18.

Daley, R. (1994). *Atmospheric data analysis*. Cambridge, UK: Cambridge University Press.

DWD (2016). *Deutscher Wetterdienst: Jahresbericht 2016*. Deutscher Wetterdienst, Offenbach, Germany. http://www.dwd.de.

Durran, D. R. (1998). *Numerical methods for wave equations in geophysical fluid dynamics*. New York: Springer.

Ebert, E. E. (2009). Neighborhood verification: A strategy for rewarding close forecasts. *Weather and Forecasting*, *24*(6), 1498–1510.

Ehrendorfer, M. (1997). Predicting the uncertainty of numerical weather forecasts: A review. *Meteorologische Zeitschrift*, *6*, 147183.

Ferro, C. A. T., & Stephenson, D. B. (2011). Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, *26*(5), 699–713.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., & Ebert, E. E. (2009). Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, *24*(5), 1416–1430.

Giraldo, F. X., & Restelli, M. (2008). A study of spectral element and discontinuous Galerkin methods for the Navier-Stokes equations in nonhydrostatic mesoscale atmospheric modeling: Equation sets and test cases. *Journal of Computational Physics*, *227*(8), 3849–3877.

Gramelsberger, G. (2010). Conceiving processes in atmospheric models–general equations, subscale parameterizations, and superparameterizations. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *41*(3), 233–241.

Hamill, T. M. (2012). Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Monthly Weather Review*, *140*(7), 2232–2252.

Heinze, R., Dipankar, A., Henken, C. C., Moseley, C., Sourdeval, O., Trömel, S., et al. (2017). Large-eddy simulations over Germany using ICON: A comprehensive evaluation. *Quarterly Journal of the Royal Meteorological Society*, *143*(702), 69–100.

Jablonowski, C., & Williamson, D. L. (2006). A baroclinic instability test case for atmospheric model dynamical cores. *Quarterly Journal of the Royal Meteorological Society*, *132*, 2943–2975.

Jolliffe, I. T., & Stephenson, D. B. (Eds.). (2011). *Forecast verification: A practitioner's guide in Atmospheric Science* (2nd ed.). Wiley.

Kox, T., & Thieken, A. H. (2017). To act or not to act? Factors influencing the general public's decision about whether to take protective action against severe weather. *Weather, Climate, and Society*, *9*(2), 299–315.

Kox, T., Gerhold, L., & Ulbrich, U. (2015). Perception and use of uncertainty in severe weather warnings by emergency services in Germany. *Atmospheric Research*, *158*, 292–301.

Lauritzen, P. H., & Thuburn, J. (2012). Evaluating advection/transport schemes using interrelated tracers, scatter plots and numerical mixing diagnostics. *Quarterly Journal of the Royal Meteorological Society*, *138*, 906–918.

Lazo, J. K., Morss, R. E., & Demuth, J. L. (2009). 300 billion served: Sources, perceptions, uses, and values of weather forecasts. *Bulletin of the American Meteorological Society*, *90*(6), 785–798.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., & Gneiting, T. (2017). Forecasters dilemma: Extreme events and forecast evaluation. *Statistical Science*, *32*(1), 106–127.

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*, 3515–3539.

LeVeque, R. J. (1996). High-resolution conservative algorithms for advection in incompressible flow. *SIAM Journal on Numerical Analysis*, *33*, 627–665.

Long, R. R. (1953). Some aspects of the flow of stratified fluids - Part 1. A theoretical investigation. *Tellus*, *5*(1), 42–58.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141.

Lynch, P. (2008). The ENIAC forecasts: A re-creation. *Bulletin of the American Meteorological Society*, *89*(1), 45–55.

Ogura, Y., & Phillips, N. A. (1962). Scale analysis of deep and shallow convection in the atmosphere. *Journal of the Atmospheric Sciences*, *19*, 173–179.

Parker, W. S. (2010). Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *41*(3), 263–272.

Randall, D., Krueger, S., Bretherton, C., Curry, J., Duynkerke, P., Moncrieff, M., et al. (2003). Confronting models with data: The GEWEX cloud systems study. *Bulletin of the American Meteorological Society*, *84*(4), 455–469.

Robert, A. (1993). Bubble convection experiments with a semi-implicit formulation of the Euler equations. *Journal of the Atmospheric Sciences*, *50*, 1865–1873.

Skamarock, W. C., & Klemp, J. B. (1992). The stability of time-split numerical methods for the hydrostatic and the nonhydrostatic elastic equations. *Monthly Weather Review*, *120*, 2109–2127.

Staniforth, A., & White, A. A. (2007). Some exact solutions of geophysical fluid-dynamics equations for testing models in spherical and plane geometry. *Quarterly Journal of the Royal Meteorological Society*, *133*, 1605–1614.

Stensrud, D. J. (2007). *Parameterization schemes: Keys to understanding numerical weather prediction models*. Cambridge University Press.

Straka, J. M., Wilhelmson, R. B., Wicker, L. J., Anderson, J. R., & Droegemeier, K. K. (1993). Numerical solutions of a non-linear density current: A benchmark solution and comparisons. *International Journal for Numerical Methods in Fluids*, *17*, 1–22.

Wilhelmson, R., & Ogura, Y. (1972). The pressure perturbation and the numerical modeling of a cloud. *Journal of the Atmospheric Sciences*, *29*, 1295–1307.

Wilks, D. (2011). *Statistical methods in the atmospheric sciences* (3rd ed., Vol. 100). Academic Press.

Wilks, D. S., & Hamill, T. M. (2007). Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, *135*(6), 2379–2390.

Yanai, M. and Johnson, R. H. (1993). Impacts of cumulus convection on thermodynamic fields. In *The representation of cumulus convection in numerical models*, number 46 in Meteor. Monogr. (pp. 39–62). American Meteorological Society.

Chapter 30
Validation of Climate Models: An Essential Practice

Richard B. Rood

Abstract This chapter describes a structure for climate model verification and validation. The construction of models from components and subcomponents is discussed, and the construction is related to verification and validation. In addition to quantitative measures of mean, bias, and variability, it is argued that physical consistency must be informed by correlative behavior that is related to underlying physical theory. The more qualitative attributes of validation are discussed. The consideration of these issues leads to the need for deliberative, expert evaluation as a part of the validation process. The narrative maintains a need for a written validation plan that describes the validation criteria and metrics and establishes the protocols for the essential deliberations. The validation plan, also, sets the foundations for independence, transparency, and objectivity. These values support both scientific methodology and integrity in the public forum.

Keywords Climate · Modeling · Verification · Validation · Science · Society · Quantitative · Qualitative · Community

30.1 Introduction

This chapter addresses the evaluation and validity of climate models. This subject has been addressed from the point of view of several disciplines: natural science, philosophy, computational science, software engineering, and law. The ultimate conclusion of this chapter is that an essential practice of climate model validation is needed to support the scientific, political, and societal uses of the scientific investigation of the Earth's climate.

The genesis of this chapter is the management, during the 1990s, of the Data Assimilation Office at the National Aeronautics and Space Administration's

R. B. Rood (✉)
Department of Climate and Space Sciences and Engineering,
University of Michigan, Ann Arbor, MI 48109, USA
e-mail: rbrood@umich.edu

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_30

(NASA's) Goddard Space Flight Center. The Data Assimilation Office[1] developed global weather and climate models that merged observations with model predictions. This process is called data assimilation.

Because the products of the Data Assimilation Office were to have routine applications in NASA's missions and scientific programs, it was required that they have a transparent and peer-reviewed validation process. The first version of the validation plan is described in the Data Assimilation Office's Algorithm Theoretical Basis Document (Data Assimilation Office 1996). This formalized validation process was institutional and beyond the testing and evaluation that occurred in the day-to-day activities of scientists and computational experts.

NASA has a strong culture of verification and validation for hardware, software, and observational data (for example, National Aeronautics and Space Administration (NASA) 2016). Extension of this culture to products and predictions from weather and climate models was, on the surface, self-evident. However, many scientists maintained that models could not be validated.

An influential paper by Oreskes et al. (1994) sets the formal argument that, in general, numerical models of geophysical phenomena cannot be validated. The argument is twofold. First is that "the climate" cannot be observed in its entirety. Second is that models are nonunique estimates of possible climate states. There are many threads to be followed in this argument, including that even if one were able to entirely observe "the climate" and the model happened to represent that instant, did the model do it for the right reasons? At the core of these arguments is that discrete numerical representations of the climate are always estimates with associated errors. As these models are constructed, they are designed to account for these errors; model performance is always a function of compensating errors.

The echoing of the statement that weather and climate models "cannot be validated" does not serve the discipline well. It belittles the consuming efforts of a large community of scientists and software engineers, who spend their time in many forms of testing and validation. Given the societal uses of weather and climate models, ranging from alerts of tornado risks days in advance to requiring changes in the world's energy systems to limit environmental warming, the notion that such models cannot be validated provides an unstable foundation for end users. It also contributes to a stable foundation of political argumentation that model-based predictions are too uncertain on which to base policy (Edwards 2010, Chaps. 15 and 16; Lemos and Rood 2010)

Focusing only on the roles of models and validation in the scientific method, the conclusion that models cannot be validated is at odds with scientific practice. Though people often view "science" as the domain of factual truth, the outcomes of scientific investigation are not "facts." Rather, the scientific method is the foundation for the exploration of natural phenomena with the outcomes being knowledge and a description of the uncertainties of that knowledge. The process of validation substantiates the uncertainty descriptions. Facts are, perhaps, knowledge with vanishingly small uncertainty, a rare outcome in the study of complex, natural systems. That models

[1] Now Global Modeling and Assimilation Office (https://gmao.gsfc.nasa.gov/).

cannot be validated is a conclusion that is meaningful in an abstract sense, perhaps, as an asymptotic approach to unknowable truth. However, such an unbounded interpretation of models stands at odds with verifiable evidence of the valid use of models and their ubiquitous and successful applications in society.

This chapter is organized as follows. The next section outlines some of the philosophical discussions of climate model validation and the development of community validation efforts by climate scientists. This is followed by the definition of terms that describe the use of climate models in the practice of scientific investigation. Then, there is a deconstruction of how weather and climate models are built, evaluated, and deployed. The definitions and the deconstruction are then synthesized to describe a general approach to the roles of testing, evaluation, verification, and validation in climate science. In the concluding discussion, the crucial role of validation in scientific organizations is described. The end conclusion is that validation is an essential practice of climate science, vital not only to the credibility and legitimacy of the scientific investigation but also to the applications of models in problems of decision-making in management and policy.

30.2 Climate Model Validation: Emergence of Definition and Community Practice

The Oreskes et al. (1994) paper serves as the starting point for a quick review of the verification and validation of models of natural systems. Other chapters in this volume provide more complete discussions of verification and validation, errors, uncertainties, calibration, and methodologies. The chapters on Weather Forecasting and Uncertainty Quantification Using Multiple Models are directly relevant to this chapter. Therefore, only an outline relevant to climate models is provided here.

Norton and Suppe (2001) discuss the credibility of climate models and point out that all of modern science relies on models. This is true, even, for what we define as observations. This is especially important for satellite observations, which are core to climate model evaluations. As will be discussed later, the reliance on models to determine the “observations” confounds the issues of independent sources of information for evaluation purposes.

Climate modeling is classified as computational science (Post and Votta 2005) and relies upon computational fluid dynamics. There is a rich literature on verification and validation in computational fluid dynamics, much of which is directly related to both weather and climate modeling.

Importantly, there have been efforts to standardize the language, with, broadly, verification focused on the correctness of the computational implementation, and validation focused on comparison of simulations with observations of the natural or experimental states. Oberkampf and Trucano (2002) provide an extensive review of verification and validation in computational fluid dynamics. In their review, they describe multilevel strategies and break down the construction, testing, and validation of complex codes. Some of the details of their approach will be used later. Roy and Oberkampf (2011), focusing on emerging techniques of uncertainty quantification, describe a structured approach to verification and validation. They demonstrate their methods comparing simulations to measurements from a wind tunnel.

Roache (1998, 2016) separates verification into two types. The first type is a verification that the computational code does what it is intended to do. The second type is a verification, focused on computational solutions, that describes the expected uncertainties in the calculation. Validation is then the comparison of the code with measures of reality, which can be measures of nature or measures of experiments.

The distinction that verification refers to computational attributes of simulation science and validation refers to comparisons of simulations to observations will be used here (see Chap. 4 by Murray-Smith in this volume). We also accept that climate models can be validated, and that the process of quantifying and describing the predictive skill of models is "model validation." (Dee 1995). Dee (1995) states, also, that constructive approaches to a model validation "process requires not a binary criterion of (true or false, valid or invalid) but rather a continuous one."

There are a number of unique practices of climate model evaluation that have emerged from the international modeling community. This is, in part, a response to the political and societal uses of climate models and their implications for foundational changes to global energy practices, built infrastructure, and economic robustness (see Saam's chapter on User's Judgements in this volume).

Notably, the climate science community has developed a culture of model intercomparison projects (MIPs). Gates (1992) describes the Atmospheric Model Intercomparison Project (AMIP). Characteristics of AMIP included simulation design, model specification, and the goal of all modeling groups performing the same suite of simulations. Also, important to the intercomparison is objective evaluation by independent experts, which is often achieved by spanning a community of experts. That is, diagnostics are prescribed that all modeling groups have to provide, and the ultimate analysis and synthesis include scrutiny by others than the model developers. The Coupled Model Intercomparison Project (CMIP)[2] was founded in 1995 and now focuses on the coupled atmospheric, ocean, land, ice, and biosphere models that are used for climate modeling. The CMIP experimental design changes from one community-wide experiment to the next. CMIP design and use are highly motivated by the needs for international assessments of climate change, such as those under the auspices of the Intergovernmental Panel on Climate Change.[3]

Sundberg (2011) investigates the culture of model intercomparison projects. A finding of Sundberg is that model intercomparison projects serve both social and scientific functions. The projects define credibility within a community by defining the type of experiments that the models are expected to be capable of and, ultimately, the standards of performance in those experiments. Climate model evaluation is distinguished by comparisons with past observations to establish the credibility of

[2] https://www.wcrp-climate.org/wgcm-cmip.

[3] http://www.ipcc.ch/.

future, unobserved, states. The observational-based analysis provides metrics, which emerges as defensible standards that have the endorsement of the dominant portion of the community. Sundberg (2011) contends that one purpose of intercomparison projects is to establish modeling as a pillar of scientific investigation on par with observational and theoretical (analytical) methods of investigation.

A number of researchers have studied the institutional practices of climate model evaluation. Guillemont (2010), interviewing at both European and United States institutions, concludes that there is "no systematic protocol for evaluating models." However, it is clear that the practice of climate model evaluation at all of the centers involves many of the same steps. These steps address issues of both software development and scientific development. They span the complexity of the system, the different scales that need to be represented, and the richness represented by the observations.

Complexity of both climate models and the Earth's climate is a recurring theme in the efforts to evaluate and establish the validity of climate models. Lenhard and Winsberg (2010) maintain that the complexity of climate models conflated with the history and practice of climate model development pose fundamental challenges to model validation. They conclude that "analytic understanding" of climate models in the sense of being able to link climate model successes or failures to specific shortcomings in the sub-models that represent specific physical processes is difficult, unlikely, and perhaps impossible. This leads to an evaluation strategy that looks, as a system, at the performance of climate models, realism as expressed by the observations, and consistency of the models and observations with the theory on which the models are built. Lenhard and Winsberg (2010) maintain that for the foreseeable future, climate model credentials will rely on expert interpretation of many simulations by many models, that is, the results of a plurality of models.

The emergence of community intercomparison projects promotes the development of shared standards of evaluation. The practice establishes the essential role of observations in the evaluation process. This contributes to the credibility of model simulations, by enabling a form of evaluation that is more rigorous than model-to-model comparisons, which occur in less data-rich disciplines.

A culture of verification and validation emerges from climate modeling community, which includes both observations and simulations. The models, originally designed as simplified representations of nature, become, themselves, complex systems whose behavior is difficult to describe. Evaluation, verification, and validation are, then, multilayered processes that cross disciplines and which use many sources of observations and many types of models. Verification and validation are vital aspects of the construction and applications of climate models, and these processes are so ingrained into the cultures of modeling centers, they are often not specifically recognized (Shackley 2001).

30.3 Definition of Terms

This section defines some key terms to formalize the structure of climate model validation.[4] Relevant material is found Chap. 2 by Beisbart and Saam and in Chap. 4 by Murray-Smith in this volume.

There is a need to define terms to provide a stable foundation for communications as well as to comply with the principles of scientific investigation and to support scientific organizations. The challenges of defining terms are made more difficult because there are needs to establish both the computational and natural science credibility of models. There are often ambiguities in language, because meaning is based on the background and goals of individuals and expertise groups.

Evaluation is a general term that includes both quantitative measures and qualitative analysis of a model's ability to address its design goals. Validation follows from the comparison of model simulations with observations of nature or experiments to establish the accuracy of the natural science of the model. Accuracy is informed by quantitative, often statistical, measurement of the suitability to address a specific application. Verification is associated with the computational integrity of the code and might include comparisons with analytic test problems as well as comparisons to high-fidelity computations. Testing is defined as part of verification and validation. That is, testing checks the performance, quality, reliability—generically, some attribute in a way that is narrowly defined compared to the model as a whole (Clune and Rood 2011).

"Systems" validation is defined as a comparison with an established baseline of simulations from an earlier release of the modeling system. For example, a comparison might be made with a portfolio of simulations of historical sets of observations. "Scientific" validation is a more open-ended process focused on the model's ability to address classes of physical processes or predictive problems for which it was designed.

The categories of system validation and scientific validation suggest another way to classify validation practice. Systems validation considers a candidate model; that is, a model under development intended to improve upon previously validated models. Comparison is made with observations as well as with the baseline version of the model.

Statistical methods are used to quantify spatial and temporal behavior, i.e., mean, bias, and variability. Statistics-based validation does not provide much information on the robustness of underlying physical, chemical, or biological processes. That is, the validation result does not say whether or not the model's answer is obtained for the right reasons; cause and effect is not evaluated. Process-based validation focuses on the representation of phenomena. Process-based validation often relies on the collection of extraordinary datasets from a quasi-isolated event that is characteristic

[4] Gettelman and Rood (2016) provides an introduction to climate science and climate modeling. Gettelman, A., and Rood, R. B. (2016), *Demystifying Climate Models: A Users Guide to Earth Systems Models,* Springer, Berlin, Heidelberg, pp. 274. The book is open source and available electronically at http://www.demystifyingclimate.org/, which also includes a list of errata.

of common types of events. An example might be to trace the evaporation of water from the Earth's surface to its return to the surface as precipitation in a thunderstorm. This process-based approach informs whether answers are obtained for the right reason.

Turning attention to the computational aspects of a model, verification can also be broken down into many steps and processes. Unit tests are fine-grained, low-level tests to assure that the programmer has, in fact, programmed instructions or algorithms correctly. Systems verification might include the ability to represent problems with known analytic solutions or to manipulate synthetic data with known properties. Another verification strategy is to compare a model simulation that has been developed as a benchmark through, perhaps, a calculation at an extraordinary resolution with a highly accurate numerical method that is too expensive to be run routinely (e.g., Jablonowski and Williamson 2006). In the verification process, tests also focus on bitwise reproducibility, checkpoint restarts, and parallel versus sequential computational fidelity. Clune and Rood (2011) describe verification practice more completely.

As described above, there are multiple steps of verification and validation that comprise the whole of the evaluation process. The steps of verification and validation span a range of complexity, which could be described as hierarchical. However, the steps are better viewed as interactive, part of the iterative, deliberative process, as opposed to a chain of hierarchical activities streaming up or down a decision tree (see also Chap. 4 by Murray-Smith in this volume).

The multilayered, iterative evaluation process uses different types of models. These model types and their use in practice are described more fully in Rood (2010). The primary and implicit focus, here, is the comprehensive, physical model. Such models use the first-principle laws of conservation to represent the climate. The conservation laws are drawn from classical physics and require that energy, momentum, and mass be conserved.

It is important to note that in weather and climate modeling, the term "physics" is often used to mean those processes that act on local spatial scales, as contrasted to fluid dynamical processes that occur on nonlocal spatial scales. The fluid dynamical processes and local-scale processes represent the conservation laws, and both are elements of the physical model—often called by climate modelers the "dynamics" and the "physics" (see also Chap. 29 by Theis and Baldauf in this volume).

The different types of physical models that find their use in evaluation are comprehensive, mechanistic, and heuristic. Comprehensive models seek to model all of the relevant interactions in a system. Mechanistic models prescribe some variables or boundary conditions, and the system evolves relative to the prescribed parameters. The first "climate" models were atmospheric models with the land, ocean, and ice at the surface specified as boundary conditions. As climate models have evolved, complexity has increased in incremental ways with coupling of atmospheric models with land, ocean, and ice models. Today, a climate model and the most advanced weather models are made of coupled component models.

Heuristic models follow, for example, from limits at large spatial- or time-averaged scales. They describe correlated behavior based on fundamental theoretical considerations. That a comprehensive model compares well with heuristic models at the comparable scales provides a measure of consistency, which is defined as an evaluation of whether the correlated behavior of variables is consistent with underlying first-principle considerations. Consistency is an important complement to measures of accuracy such as mean, bias, and variability.

There are also statistical models of the climate. Statistical models are extensively used to define the local-scale "physics" and their accumulated effects in physical models. They often rely on intensive observing campaigns that develop statistical relationships between observed variables of an evolving dynamical system. This leads to parameterizations, and the term local-scale parameterization will be used to describe the finest structure of model decomposition used here. Related to parameterization, the term algorithm will be used to represent numerical formulation of physical processes and functions that are directly derived from the underlying equation set (see Chap. 41 by Frisch and Chap. 29 by Theis and Baldauf in this volume).

Statistical models, more generally, predict future behavior based on past, observed behavior. Statistical models are used, for example, to predict sea surface temperatures in the Tropics from 1 year to the next (e.g., Johnson et al. 2000). Statistical models rely on having adequately observed behavioral relationships and for that behavior to remain the same (stationary) with time. That comprehensive models represent observed statistical behavior is a technique used in evaluation and validation.

Below is a list of selected terms:

- Physically based (physical) model: uses first-principle laws of conservation energy, momentum, and mass to represent and predict weather and climate.
- Component model: physically based model of atmosphere, ocean, land, ice, chemistry, biology, etc. A discipline-based model of a major subdiscipline of climate science.
- Coupled model: a model built from connected component models—that is, a climate model
- Application: the end use of a model, for which the model is designed.
- Evaluation: a general term to describe quantitative measures and qualitative analysis of a model's ability to address its application(s).
- Testing: checks the performance, quality, reliability—generically, in a way that is narrowly defined compared to the model as a whole.
- Verification: associated with the computational integrity of the code, and includes comparisons with analytic test problems, synthetic data, and high-fidelity computations.
- Benchmark: a routine test using synthetic, numerical, or observational data that establishes standards or performance—part of verification or systems validation.
- Validation: comparison of model simulations with observations of nature or experiments to establish the accuracy of the natural science of the model.
- Systems validation: a comparison with observations from an established baseline of simulations from an earlier release of the modeling system.

- Scientific validation: the process of assessing, by comparison with observations, a model's ability to address classes of geophysical problems (applications) for which it was designed.
- Statistics-based validation: determination of mean, bias, and variability of a candidate model relative to observations or previously validated model
- Process-based validation: investigation of model representation of quasi-isolated phenomena to analyze cause and effect.

30.4 Model Construction, Observations, Assimilation: Roles in Validation

In the ideal practice of science, observed phenomena are investigated with controlled experimentation. There is the notion that the experiment is confirmed or refuted by independent observational data. Such objective purity is rare; absolutism is not possible.

In weather and climate science, controlled experimentation of the natural system is not possible. In fact, observations are difficult to make; direct observations of "the climate" are rare. Temperature, the most familiar and iconic measurement of weather and climate, might come from thermometers, gases trapped in layers of ice, growth rings in trees, or radiation measured by space-based satellites. In all of these cases, a model of some type enters into assigning temperature to an observable.

The practice of computational science to investigate and predict the Earth's climate is placed in four elements: observations, infrastructure, models, and assimilation. These elements are related to each other; however, those relationships are not hierarchical, leading from one step to another. Rather they exist in an ecosystem, dependent upon the particular attributes of the application being addressed. Evaluation becomes an iterative, deliberative process, which requires diligence and peer-based scrutiny to assure the integrity of science-based investigation.

Of the four elements, observations are at the foundation. Scientific investigation relies on measured phenomena, observations. Models rely on observations. The observations of climate and climate change are many. The incomplete definition of climate as "average weather" suggests the importance of wind, temperature, and water. However, climate science and comprehensive models, ultimately, require measurements of many (>100) independent and derived observables to describe the air, ocean, ice, land, chemistry, and biology and their interactions. As we learn more about climate change and its impacts, we learn that new types of measurements are needed. Hence, observations of the "climate" do not sit as a distinct, complete, independent body of knowledge; models and their applications steer observational needs. Conversely, many of the observations require models or model components in their production.

The explicit mentioning of modeling infrastructure is warranted because of the complexity of climate models and the distribution of expertise across institutions.

Climate science evolves and emerges from many different fields of natural science—meteorology, oceanography, hydrology, glaciology, etc. (Edwards 2010, Chap. 7). As a result of the many disciplines involved in climate science, the many institutions, the independently developed computer codes, the inherent uncertainties, the societal consequences, and other sources of complexity, infrastructure becomes part of the scientific credibility and robustness of climate science. Infrastructure encompasses organizing structures and services, often focused on communication of information within computer codes, institutions, and people. Of specific interest is the software and hardware infrastructure required for computational science.

Two types of software infrastructure are introduced. The first is the infrastructure to support the coupling of the component models that make up climate models (Theurich et al. 2016). In this case, the infrastructure serves to bring order to model coupling, which has important consequences for the scientific method (see also Chap. 39 by Lenhard in this volume). First, there is the ability to do controlled simulation experiments, where component models can be changed one at a time to investigate, for example, the sensitivity of projections to the choice of ocean model. Second, there are questions of coupling methodology that need to be investigated and have scientific consequences. Therefore, infrastructure requires the verification of its computational integrity and enters into the portfolio of climate model components that require validation.

The second type of software infrastructure is that which facilitates model analysis and model intercomparison. Examples of this type of infrastructure include the Earth System Grid Federation[5] (Williams et al. 2016), which provides services for the Coupled Model Intercomparison Project. This infrastructure contributes to validation in several ways. In a formal sense, model simulations are made broadly accessible, hence, open to independent scrutiny. Models from several institutions are brought into a common methodology of evaluation and intercomparison, with the evaluations carried out by scientists who were not model developers. Finally, standard tools are built, collected, curated, and provided, which supports rigor and objectivity in the community. This infrastructure supports transparency, independence, and objectivity—all parts of model validation.

In the practice of climate science, models are used in two primary roles (Rood 2010). The first is diagnostic when the models are used to determine and to test the processes representing a set of observations. In this case, observations determine whether or not the processes are well known and adequately described; the model is validated with observations of the process. The second role is prognostic when the model is used to make a prediction.

A climate model can be viewed as a coupled composite of component models. Each of the component models can be viewed as the representation of a "process", in this case, the atmospheric processes, the oceanic processes, etc. Taken in isolation, the atmosphere model is made of many processes, for example, different types of clouds, the transfer of energy through the atmosphere by radiation, and turbulence. This reduction-based approach to model building is called process splitting and is a

[5] http://esgf.llnl.gov/.

standard way to build models (e.g., Strang 1968). The strength of the approach is that problems become tractable. The weakness of the approach is that the theories and algorithms that describe the processes are developed with some degree of isolation. They have to be connected, coupled, in the formation of the model as a whole. It is difficult to assure physical consistency.

This introduction of how models are built provides context for the relationship between models and data, and hence, validation. A diagnostic, process-focused examination of an isolated thunderstorm might rely on unique, high-quality observations. These observations might be used with a statistical model to define the parameterization that represents how heating at the Earth's surface, turbulence near the Earth's surface, leads to updrafts that cause clouds and thunderstorms. Hence, observations are used to guide the definition of model processes; they define local-scale parameterizations. Then, the model is used to predict future states, and different observations, likely with vastly different temporal and spatial attributes, are used to measure success and failure.

The use of observations to both construct and evaluate climate models hints at the intertwined relationships between observations and simulations that must be managed and disentangled in the validation process. The entanglement of models and observations becomes even greater when the fourth element of practice, assimilation, is considered.

Assimilation is the melding of model predictions with observations (Rood 2010). Originally used to provide the initial condition for weather forecasts, assimilation has become a core practice of weather and climate science. Many studies use assimilated data products as "observations." Weather forecasts are accurate enough in time ranges of hours to days that they are used to generate estimates of observations of sufficient accuracy to provide quality control of monitoring observing systems (e.g., Stajner et al. 2004). Such predictions also provide first guesses of observations to assist in, for example, retrieval of geophysical parameters from space-based observations of radiance.

The most powerful attribute of data assimilation in climate studies is to fill in the gaps. This gap-filling ranges from filling in the spatial and temporal gaps of observing systems, to estimating processes that are not observed. Of course, these data-influenced estimates of "observations" are reliant upon the model parameterizations, which were, originally, defined with the help of other observations. The broad use of assimilated datasets known as reanalyses in model validation raises philosophical and practical concerns that make it incumbent upon expert peer review to inform the legitimacy, credibility, and integrity (Cash et al. 2003) of the validation process.

30.5 Validation of Climate Models in Practice

The evaluation and validation of climate models is a core activity of the practice of climate change. Flato et al. (2013) describe the evaluation process and results used in the evaluation of the models used in the Intergovernmental Panel on Climate Change's Assessment Report 5.

The verification and validation of weather and climate models consider many criteria (see Chap. 24 by Liu and Yang in this volume). These include

- the correctness of a set of equations to represent phenomena;
- the accuracy of the representation of those equations with discrete mathematics suitable for digital computers;
- the correctness of the implementation on the computers;
- the construction, by coupling, of comprehensive models from component models in which functions and physical processes have been represented in a split or granular fashion;
- the ability of component and coupled models to represent observations with correct physical, chemical, and biological processes; and
- the ability of the coupled model to represent the conservation of energy, mass, and momentum,

The verification and validation processes are not purely quantitative as there are expert judgments and management of information that is a balance of positive and negative attributes. In climate model validation, it is also important to consider the attention that the validation process will receive in the public discourse about the societal uses of climate simulations.

This section provides a structure for the verification and validation process. It opens by establishing transparency, independence, and other values that are critical to the scientific method as well as public scrutiny. Then, the issues of identifying suitable observational validation data are discussed. A process anchored around a documented validation plan is introduced. First, the attributes of validation that require deliberation and expert analysis are introduced. Then, quantitative analysis is described as layers characterized by increasing geophysical complexity.

30.5.1 Independence, Transparency, and Objectivity: Basic Values of Verification and Validation

Independence, transparency, and objectivity are values of the scientific method and validation. For climate science, these values have broader importance. The results from the investigation of the Earth's climate motivate societal interventions that are disruptive. Therefore, observational data, simulation data, and how they are validated become societal assets. This opens them up to the scrutiny that is far broader than science-based validation. They become part of political arguments, which often focus on aligning scientific uncertainty with political goals (Lemos and Rood 2010).

Model developers and model scientists, individually, are responsible for performing and documenting test procedures and results. However, when many people and institutions are providing model components, algorithms, and local-scale parameterizations, their individual efforts at verification and validation do not assure that the collective whole is validated. Therefore, in modeling centers, it is essential to develop organization-wide and model-system-wide testing, verification, and validation procedures. The validation process and evaluation criteria need to be documented and results must be available for scrutiny by those not directly involved in, for example, building the model. This suggests two principles of validation, independence, and transparency.

The validation process needs to be designed and agreed upon at the beginning of a model development cycle or a simulation experiment. Metrics need to be determined, as well as standards for comparison.

Testing and validation plans that can be executed and evaluated by experts, who are not directly involved in building and deploying a model, are necessary. Such independence serves to evaluate the robustness of logic and correctness of the implementation. Independent review is well suited to reveal confirmation bias, where a developer or scientist might have limited their evaluation once a result agreed with their expectations. Independent review brings different perspectives and different expertise bases to an evaluation; it addresses issues of conflicts of interest.

Transparency and the documentation of models and their validation support another attribute of scientific investigation, reproducibility. Reproducibility and peer review by the scientific community are part of the practice of the scientific method and are part of the scientific validation.

All of these principles of validation aim at objectivity and the development of trust that conclusions are based on evidence of quantitative measures. The end result, in this case, is a determination that a model is suitable for its application. There is a description of what has been concluded and descriptions of uncertainties, perhaps, including a description of unsuitable applications of the model.

30.5.2 Identification of Independent Observational Data

As described in the previous section, observations and simulations are intertwined. Therefore, a priori expectations that observational data and simulation data are independent of each other must be evaluated as part of the validation process. This subsection considers the issues regarding the independence of simulation and observational data. The controversy concerning the relationship between satellite temperature measures and simulation is used as an example (Mears and Wentz 2017; Santer et al. 2017). Lloyd (2012) provides a philosopher's perspective of the controversy.

Detectors on satellites measure electromagnetic radiation at specific frequencies. To measure temperature from space relies on understanding the absorption and emission of radiative energy in the Earth's atmosphere. In order to relate the space-based measured radiances to temperature, a radiative-transfer model is required. The equations of the radiative-transfer model are the same for the observational application and the climate model. The details of the radiative-transfer model implementation for temperature determination and the one used in a weather or climate model will be different.

Validation approaches for satellite observations have a fundamental difference from climate model validation. The validation of observations is a problem of reduction, which ultimately might be the comparison of a set of independent measurements at a single point in space and time. Instrument validation looks toward less complexity. A climate model, however, looks toward more complexity. As component models are combined into climate models, more complexity is included. It is a problem of expansion.

In the case of validating satellite temperature observations, the narrowing view supports deductive conclusions about the quality of the satellite observations. Successful validation of the satellite temperature provides quantitative information about radiative-transfer models, and hence, information relevant to the correctness of analogous equations and their computational implementation in climate models.

The controversy over discrepancies between observed satellite temperature trends and climate model trends (Lloyd 2012) is framed by their being multiple algorithms for calculating satellite-based temperatures. The discrepancies between models and different calculations of observational information help to define research directions for both observational and simulation scientists. If there is convergence of the models and observations, then confidence in conclusions increases. If there is divergence, then errors are exposed, which can be corrected. In either event, the intertwined roles of models and observations challenge both the researcher and the communication of the research for societal use.

In cases when there are not truly independent observations, there are strategies that withhold some observations to assure their independence. There are other strategies that isolate models based on their use in data analysis and assimilation. That is, a model used to calculate merged model–observation assimilation products is not, then, used for climate predictions. This is in contrast to weather forecasting, where the assimilation is used to provide the best possible initial state of the weather prediction. The use of assimilation products in climate model validation, always, carries philosophical concerns. Some observational datasets have too big a role in model development to allow them to be used in validation; they are only measures that the model was implemented correctly. Such observations have transitioned from validation to verification. For the purpose of this chapter, it is assumed that due diligence has been exercised to assure simulation–observation independence.

30.5.3 Deliberative Validation and Expert Judgment

The goals of testing, verification, and validation are to assess correctness at all stages of development and implementation. The validation process communicates the trustworthiness of models to their users. Therefore, the validation process must address the principles of independence, transparency, and objectivity.

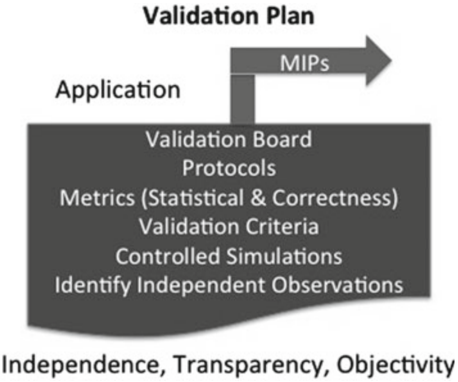

Validation Plan
MIPs
Application
Validation Board
Protocols
Metrics (Statistical & Correctness)
Validation Criteria
Controlled Simulations
Identify Independent Observations
Independence, Transparency, Objectivity

Fig. 30.1 Structure of a validation plan. The selection of the application of the model defines the details of the validation plan. The plan aspires to assure the values of Independence, Transparency, and Objectivity. The plan assumes that the modeling organization will participate in the community-based model intercomparison projects (MIPs)

The first step in validation is the development of a validation plan. The elements of a validation plan are highlighted in Fig. 30.1 and described below. If the validation criteria are known at the beginning, then the definition is added to the validation process. Transparency is provided to both developers and end users. During development, it is always the case that scientists and developers can identify further improvements. The plan, therefore, defines an endpoint, based on how well the model performs at a particular snapshot and assures that the model addresses particular user needs.

The determination of evaluation criteria, metrics, and standards of comparison requires careful consideration of the purpose, the application, of the model. Example applications might be weather forecasting, seasonal forecasting, decadal climate projections, and multi-century climate projections.

The presence of an application gives the model purpose, an anchor in reality; it relieves the model from the impossibility of representing an unknowable truth. Increasingly, organizations seek to use unified modeling systems for a range of applications. This has both scientific and management motivations. A validation plan that spans the suite of applications advances unified modeling systems; however, it causes tensions when model development improves one application at the expense of another. In this case, deliberations that consider management and organizational priorities are required; protocols for managing these deliberations are part of the validation plan. This is but one place where expert judgment contributes to climate model validation (see Saam's chapter on User's Judgment in this volume).

Adherence to a validation plan improves the ability of an organization to allocate human and computational resources. A well-documented testing and verification procedure eases coordination within an organization and collaborations with external organizations. An organization is better able to meet goals within budget and on schedule.

Within the plan, independence benefits from the definition of a validation team. The validation team should be largely independent of model developers. Model developers, as well as end users, are an essential part of writing the validation plan. Their presence helps to assure the relevance of the validation criteria and metrics. Model developers are also essential in the analysis to understand cause and effect. However, the exercise of scoring and ranking model performance should fall to an independent group. Such independence contributes to objectivity and is consistent with the scientific method.

Statistical evaluation gives quantitative measures of accuracy. However, there are nuanced scientific considerations that need to be considered in the validation plan. A new local-scale physics parameterization, fluid dynamics scheme, treatment of topography, etc. can represent a significant improvement in the correctness of the equation set or their numerical representation, i.e., a science-based improvement. Such an algorithm might improve the realism of features such as fronts, that is, simulated frontal passages "look like" nature's frontal passages (see Chap. 16 by Meyer in this volume). It is possible, perhaps even likely, that the first implementation of the scientific improvement will lead to decreased performance in some metrics. Indeed, in the hands of an expert calibrator, less scientifically correct schemes can be modified to meet specified statistical measures. However, in the end, the correctness of the equations and their representation as numerical algorithms improve the basic construction of the model.

The seeming paradox of a "more correct" model leading to less accurate statistical scores occurs because the balance among algorithmic and parametric approximation errors is changed (see Chap. 5 by Roy in this volume). The validation team is, therefore, sometimes faced with a judgment call of accepting a lower scoring model with an improved scientific basis. Such a decision has long-term consequences. The validation plan, therefore, needs to consider the balance between potential quantitative degradation versus more robust future development.

The design of the validation plan should assure that the model is susceptible to quantitative validation. In an ideal world, models submitted for validation would evaluate a small number, perhaps single, changes. However, this is not practical. Scientific development moves forward in the component models as well as the technical development of the model infrastructure. Significant validation occurs with component models, leaving the challenge of multiple changes being tested in a coupled environment. Analysis of the expected outcomes of the individual changes needs to be posited and included in the evaluation criteria. Again, protocols to manage the reality of multiple changes in multiple components and what can and cannot be tolerated in validation are required.

Final validation requires that the coupled model be validated. This is expensive and validation strategies continue to evolve. A manageable subset of coupled model test cases needs to be defined based on application priorities. Adjudication of conflicting information will be required.

The validation plan needs to anticipate the role of a model in community efforts in validation; that is, model intercomparisons such as the Coupled Model

Intercomparison Project.[6] Model intercomparisons have a strong influence on model validation and greatly enhance the confidence in climate model results. The intercomparisons are an important foundation for describing the uncertainty (see Chap. 34 by Knutti et al. in this volume). Therefore, experimental design and validation criteria for simulations extend outside of institutions to allow community-based experiment and validation protocols.

30.5.4 Quantitative Evaluation

With the foundation and scope of a validation plan set, protocols for testing, verification, and validation need to be defined. The validation plan described in Data Assimilation Office (1996) will be used in concert with the verification and validation structures described in Oberkampf and Trucano (2002).

Model categories are organized in relation to geophysical complexity. With regard to validation, the algorithms and parameterizations are least complex and can be evaluated and validated with data from isolated process experiments and, in some cases, with analytic solutions. The process models, which represent a quasi-isolated, highly observed geophysical feature, are composites of algorithms and parameterizations. An example of a process study is the growth and decay of a type of Arctic cloud (e.g., Roesler et al. 2017).

The component models require observations that span their domain of purpose, atmosphere, ocean, etc. The coupled models span multiple domains, with the most comprehensive model requiring observations representative of the entire system. The need to manage this complexity through the design of controlled simulation experiments and the design of validation exercises that have realizable metrics is self-evident. The requirement for an application or suite of applications to limit the complexity is, likewise, self-evident.

Figure 30.2, pictorially, describes quantitative evaluation in layers of increasing complexity. The left panel shows a notional four-layered structure presenting the construction of a model. At the bottom layer are local-physics parameterizations and algorithms. The next layers represent composites of these parameterizations into process models and component models. The top layer is a fully coupled climate model. These layers of models need to undergo both verification and validation.

30.5.4.1 Verification

Verification has two major goals. The first is to assure the algorithms are correctly implemented and doing what they are intended to do. The second is through comparison with analytic and well-described benchmark cases to characterize uncertainty

[6] https://www.wcrp-climate.org/wgcm-cmip.

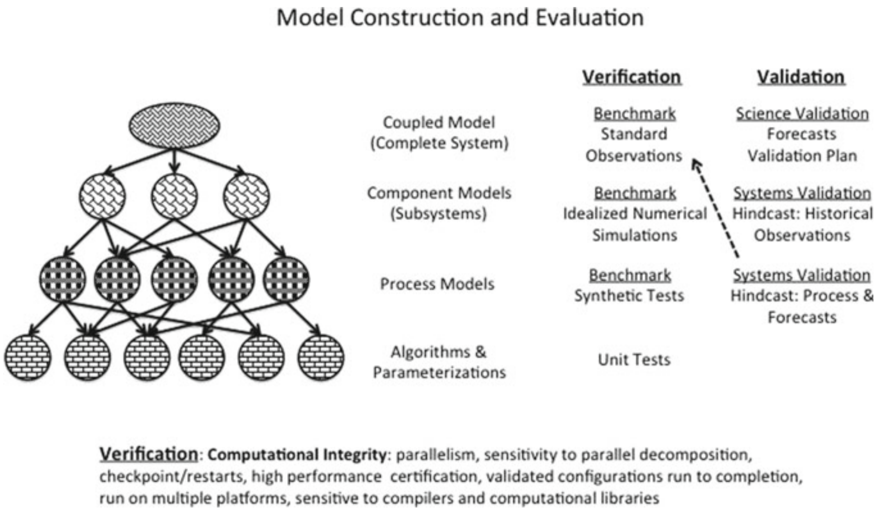

Model Construction and Evaluation
Verification
Validation
Coupled Model (Complete System)
Benchmark Standard Observations
Science Validation Forecasts Validation Plan
Component Models (Subsystems)
Benchmark Idealized Numerical Simulations
Systems Validation Hindcast: Historical Observations
Process Models
Benchmark Synthetic Tests
Systems Validation Hindcast: Process & Forecasts
Algorithms & Parameterizations
Unit Tests
Verification: Computational Integrity: parallelism, sensitivity to parallel decomposition, checkpoint/restarts, high performance certification, validated configurations run to completion, run on multiple platforms, sensitive to compilers and computational libraries

Fig. 30.2 Linking model structure to verification and validation. Following Oberkampf and Trucano (2002). The dashed arrow from Validation to Verification suggests that as Systems Validation with certain datasets evolves to a level of maturity that they no longer represent unique model quality; those tests move to verification. That is, they become benchmark standards that all models are expected to achieve

associated with the numerical representation of the science-based equations set (see Chap. 11 by Rider in this volume).

The verification process assures the computational integrity of the implementation. The targets of such testing include parallelism, checkpoint/restarts, and performance. In addition, it is important to check that model configurations run to completion, run on multiple platforms, are sensitive to parallel decomposition, and are sensitive to compilers or computational libraries (Clune and Rood 2011). These tests are not just of computational consequence as some applications rely on simulations with slightly altered initial conditions. It is important to know whether or not results differ due to computational differences or science-inspired differences.

With regard to benchmarks and test cases, at the local-physics parameterization and algorithm level, there are synthetic tests, the possibility of analytic tests, and well-defined numerical tests (see Saam's chapter on Benchmarks in this volume). For example, does a remapping scheme and its inverse remapping return the original field—is mass conserved? There is also the potential to check algorithms with narrowly defined observation-based tests, whose solutions are established benchmarks. These tests can be defined as unit tests in that they are fine-grained—at the building block level. Unit tests assure the quality of the building blocks. Errors revealed at the unit test level support efficient model development.

At the next level of complexity, when fine-grained parameterizations and algorithms are integrated together into subsystems and systems, verification tests become more challenging. There are few analytic tests at this level of integration. There is the potential to develop rigorous tests using synthetic data, which might verify successful implementation of computational code and perhaps simple (for example, linear) scientific measures. At this stage, an infrastructure that supports the ability to configure models of different complexities, e.g., process-based models or mechanistic models, is important as some fields have intensive-observation-campaign problems and datasets whose solutions are well characterized. The ability to perform these tests provides insight into both computational and scientific qualities. Such tests might be viewed as minimal standards or benchmarks as all models are expected to do the benchmarks well. Proceeding to the highest levels of complexity, component models and coupled models, there is a need for benchmark calculations relative to previously characterized simulations; however, standards are likely to be institutional rather than community-wide. This level of systems testing, which often involves observational data sets, will be deferred to systems validation. There is still research and experience needed to develop routine testing strategies and test problems for coupled systems.

30.5.4.2 Validation

Validation is establishing the suitability of a model for an application by comparisons of simulations to observations. At the lowest levels of complexity, there are often comparisons with observations specifically collected to define and test parameterizations and processes. These comparisons with observations have, effectively, moved across the transition from validation to verification. If a state-of-the-art representation is not achieved in these tests, then the parameterizations are not accepted as credible. The rest of the discussion will focus on systems validation and scientific validation.

30.5.4.3 Systems Validation

Systems validation is appropriate at the component model and coupled model levels. From the perspective of the coupled model, the validation of component models can be described as subsystem validation.

Using the atmospheric model as an example, systems validation is made up of a series of baseline simulations designed to investigate performance on a class of problems that represent its applications. Such simulations might be a set of 10-day weather forecasts from standard specified initial conditions that include all seasons (see Chap. 29 by Theis and Baldauf in this volume).

Longer simulations of the atmospheric model with specified sea surface temperatures allow the investigation of the onset of model bias and the ability to simulate several modes of climate variability, such as the El Niño–La Niña cycle.[7] Such

[7] See Chap. 9, Gettelman, A., and Rood, R. B. (2016), *Demystifying Climate Models: A Users Guide to Earth Systems Models,* Springer, Berlin, Heidelberg, 274 pp. http://www.demystifyingclimate.org/.

simulations generally rely on observations collected after 1979, when the satellite observing system became global and persistent. Simulations are compared to observations as well as large archives of simulations that have established model credibility measures. Analysis tools such as the Taylor diagram (Taylor 2001) provide statistical measures against a range of geophysical quantities that have been determined to provide a foundational measure of the climate. These tools also document changes from one generation of models to the next.

Extending the atmospheric models to include atmospheric chemistry introduces another set of baseline simulations. The field of ozone science has been out front developing integrated standard measures, which are designed to represent the combined effects of transport and chemistry. The diagnostics of Douglass et al. (1999) rely on strong theoretical constraints, that is, heuristic models. Such diagnostics can be automated from the standardized output and provide quick and profound measures of model performance.

Each component modeling discipline and some coupled models (e.g., chemistry-transport) will have a set of standard simulations that can be performed and analyzed in a reasonable amount of time (weeks to months). This will establish the credibility of the components and justify implementing, testing, verifying, and validating the performance in coupled systems. At some level, the component-level system-level evaluation can be automated. An excellent example of publically available automated validation information for the Community Earth System Model can be found online.[8]

At the coupled model level, a similar approach is used. In models designed for seasonal or the El Niño–La Niña forecasting, the ability to forecast historical archives of the El Niño–La Niña events is a natural focus. The El Niño–La Niña problem is one where statistical models also play an important role, as coupled physical models do not definitively establish the state of the art.[9]

Hindcasting, also known as backcasting, is a primary method of model evaluation and validation. It is critical to choose a historical time period when it can be established that there are adequate, independent observations to support validation.

With regard to climate models designed for century-scale applications, much attention is paid to the simulation of twentieth century, or more generally the post-industrial to the current time. Concurrent with the commerce of the industrial revolution, weather observations spread across the globe—the observational record greatly improved. The focus on the twentieth century allows examination of important modes of air–land–sea interactions, response to volcanoes, and some aspects of solar variability. Longer timescale variability associated with oceans and ice are not fully represented in the twentieth-century record.

A possible disadvantage of the twentieth-century record from the point of view of the validation scientist is that there are many human-caused alterations to the environment that influence global signals. Aside from greenhouse gas emissions, there are land-use changes, emissions of particulate pollution, policies to control particulate pollution, and composition changes that led to extreme events such as

[8] http://www.cgd.ucar.edu/amp/amwg/diagnostics/plotType.html.

[9] http://iri.columbia.edu/our-expertise/climate/forecasts/enso/2017-June-quick-look/.

the ozone hole. These changes mean that we do not have a highly instrumented, "natural," historical period to serve as a control. On the other hand, modeling the transient behavior associated with all of these environmental alterations provide valuable model tests.

Simulations of the last thousand years, which capture the onset of large carbon dioxide release and other influences of a growing human population, are also routine parts of validation. For these longer simulations, there is a greater reliance on proxy measures of climate, for example, tree rings, and lake sediments.

Hindcasts focused on isolated events allow full-system, process-based investigation. The archetypical example is a well-measured volcanic eruption (e.g., Robock 1983). Another example is an El Niño–La Niña event. Though still occurring within the global environment, these events are relatively short-lived (<5 years) and involve heating and cooling, water vapor responses (i.e.,*feedbacks*), and atmosphere–land–ocean–biological responses. Satellite observations provide global measurements of key variables. Hence, these events emerge as quasi-controlled test cases, which influence many key climate variables and exercise model processes and their interactions.

With this level of verification and systems validation, it is justified for an organization to release a modeling system for broader use. However, further scientific validation better substantiates credibility.

30.5.4.4 Scientific Validation

Scientific validation is the process of assessing by comparison with observations a model's ability to address classes of geophysical problems (applications) for which it was designed.

If the application of the model includes forecasts in a routine or operational mode, then forecast or prediction experiments are used as validation. Prediction-based validation is common in weather forecasting (see Chap. 29 by Theis and Baldauf in this volume). The basic idea is that a candidate model is scored against an existing model on how well they verify with future forecasts. Compared with hindcasts, these forecast cases have not been part of the validation data and, hence, represent states that are new to the model. If, in a statistically significant number of cases, the candidate model performs better than the previous version of the model cases, then the candidate model is validated for its forecast application.

Weather forecasting is in some ways unique because the short timescale of the needed forecasts allows the validation process to be concluded in weeks to months. For longer timescales and climate projections, it is not possible to wait for future states to be realized. Therefore, other methods of scientific validation are invoked.

For a coupled model intended for a portfolio of climate applications, the validation plan should identify a small number of metrics (<10) that the scientific improvements in the candidate model are expected to address. The priority metrics are largely based on improvements of documented deficiencies in previous versions of the model. These deficiencies are not simply those revealed by statistical measures, but, more importantly, those revealed by scientific investigation of the previous version of the model. These scientific investigations occur as communities of users exercise the model over, on the order, of 18–24 months. This is timescale appropriate for deliberative research and peer review as well as development and validation of a model. The development and validation process, presently, support a 3–5-year span in releases of modeling systems.

For scientific validation, the validation plan needs to identify classes of problems that are priority foci, for example, climate variability, hydrometeorology, and stratospheric ozone. These are each complex interrelated simulation problem. Physical consistency, as informed by correlative behavior, takes on a high value in this evaluation, that is, processes related to cause and effect. Improvement in the representation of processes stands along with statistical measures of mean, bias, and variability.

In the validation exercise, it is certain that some metrics will improve and some will degrade. At this point, it is when a pre-negotiated validation plan, reliance on the application priorities, and independence of a validation board stand to bring closure to a validation exercise. Validation becomes a deliberative process, balancing strengths and weaknesses, relative to objective measures of skill and expert judgment of the robustness of process representation. The validation results become the foundation of the uncertainty description as well as part of the next development and validation phase.

30.6 Discussion

This chapter has deconstructed and described an organized approach to the practice of climate model verification and validation. On one hand, the interactions between observations, simulation, computational approximations, and scientific correctness substantiate the arguments of Oreskes et al. (1994) that, in an absolute sense, climate models can never be proven to have gotten the right answer for the right reason. On the other hand, the comprehensive testing and evaluation of weather and climate models provide a high degree of confidence that weather and climate models provide useful information for planning and practice. Climate scientists and software engineers have developed a culture of verification and validation that establish a model's credibility and legitimacy.

Guillemont (2010) noted, across modeling institutions, both the similarity of validation practice and the lack of a formalized approach. This chapter provides definitions in an effort to describe, formally, climate model verification, and validation. The construction of models from components and subcomponents is discussed, and the construction is related to verification and validation. In addition to quantitative measures of mean, bias, and variability, it is argued that a measure of physical consistency is required. Physical consistency is evaluated as correlative behavior that is related to underlying physical theory.

The more qualitative attributes of validation are discussed. Specifically, the challenge of improved "science" leading to degraded quantitative skill is discussed. The role of realism with model weather "looking like" observed weather is introduced.

There are tensions when a model is required to address a portfolio of applications, and there is inconsistent improvement across the portfolio. The consideration of these issues leads to the need for deliberative, expert evaluation as a part of the validation process.

The chapter maintains a need for a written validation plan that describes the validation criteria and metrics and establishes the protocols for the essential deliberations. The validation plan, also, sets the foundations for independence, transparency, and objectivity. These values support both scientific methodology and integrity in the public forum.

The roles of community-informed validation and software infrastructure are also discussed. Shared software infrastructure helps to manage the complexity of multiple disciplines and multiple institutions. Likewise, shared analysis software and protocols contribute to the development of standardized methods and scores. Community-based intercomparisons contribute strongly to uncertainty descriptions.

Perhaps more so than might be expected, the chapter highlights the roles that representation of weather and weather forecasting plays in climate model development. This contributes to the discussion of the need to assure that observational data and simulation data are independent. This independence is challenged in many datasets, especially those which rely on data assimilation.

30.7 Conclusion

Daniel Farber, a law Professor at the University of California Berkeley, analyzed whether or not climate models were characterized well enough to justify societal responses to mitigate climate change and use models in adaptation planning. Farber concludes that with the model intercomparisons and the national and international assessments (Farber 2007):

> Climate scientists have created a unique institutional system for assessing and improving models, going well beyond the usual system of peer review. Consequently, their conclusions should be entitled to considerable credence by courts and agencies.

Predictions and projections will always be uncertain, which is a fact of scientific investigation (Lemos and Rood 2010). Given that climate science is embracing more complexity with each generation of models and observations, it is unlikely that uncertainty will be reduced in an absolute sense. Uncertainty reduction is not required to use climate predictions and projections in planning and practice. Uncertainty is always present in decision-making. Verification and validation frame the uncertainty description for the application.

The basic results of climate science that the Earth will accumulate heat relative to pre-industrial times, that the air and ocean will warm, that ice will melt, that sea level will rise, and that the weather will change are known with virtual certainty. The foundation of that conclusion does not lie on the increasingly complex climate models described here. The foundation relies on the basic principles of conservation of energy. Increasing carbon dioxide and other alterations to the Earth by humans cause solar energy to be held near the Earth's surface. That energy heats the Earth's surface, and there must be consequences of that heating.

The consequences of that heating are complex. Climate models are the best tool to inform us about those consequences, their interactions, and their impacts. Climate models allow us to anticipate and to plan. Climate models allow us to explore policy options. Indeed, climate models provide perhaps the most knowable aspects of what the next century will be like.

Acknowledgements I thank the editors, Claus Beisbart and Nicole J. Saam, for the opportunity to contribute this chapter and for their efforts in putting together this volume. I thank Cecelia Deluca for reading an early version of the manuscript and many discussions on modeling infrastructure, verification and validation, and insights into modeling culture.

References

Cash, D. W., Clark, W. C., Alcock, F., Dickson, N. M., Eckley, N., Guston, D. H., et al. (2003). Knowledge systems for sustainable development. *Proceeding of the National Academy of Sciences*, *100,* 8086–8091.

Clune, T. L., & Rood, R. B. (2011). Software testing and verification in climate model development. *IEEE Software, 28,* 49–55. https://doi.org/10.1109/MS.2011.117.

Data Assimilation Office (DAO). (1996). *Algorithm Theoretical Basis Document Version 1.01*, Data Assimilation Office, Goddard Space Flight Center. Retrieved from https://eospso.gsfc.nasa.gov/sites/default/files/atbd/atbd-dao.pdf.

Dee, D. P. (1995). A pragmatic approach to model validation. In *Quantitative skill assessment for coastal ocean models*. American Geophysical Union (pp. 1–14).

Douglass, A. R., Prather, M. J., Hall, T. M., Strahan, S. E., Pasch, P. J., Sparling, L. C., et al. (1999). Choosing meteorological input for the global modeling initiative assessment of high-speed aircraft. *Journal Geophysical Research, 104,* 27545–27564.

Edwards, P. N. (2010). *A vast machine*. Cambridge, MA, USA: The MIT Press.

Farber, D. A. (2007). *Climate models: A user's guide*. Berkeley, CA, USA, UC Berkeley Public Law Research Paper No. 1030607.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2013) Evaluation of climate models. In T. F. Stocker, D. Qin, G. -K. Plattner, M. Tignor, S. K. Allen, J. Boschung, et al. (Eds.) *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change.* Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Gates, W. L. (1992). AMIP: The atmospheric model intercomparison project. *Bulletin of the American Meteorological Society, 73,* 1962–1970.

Gettelman, A., & Rood, R. B. (2016). *Demystifying climate models: A users guide to earth systems models*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-48959-8.

Guillemot, H. (2010). Connections between simulations and observation in climate computer modeling. Scientist's practices and "bottom-up epistemology" lessons. *Studies in History and Philosophy of Modern Physics, 41,* 242–252.

Jablonowski, C., & Williamson, D. L. (2006). A baroclinic wave test case for dynamical cores of General Circulation Models: Model intercomparisons. *NCAR Technical Note NCAR/TN-4691STR*, National Center for Atmospheric Research, Boulder, CO (89 pp).

Johnson, S. D., Battisti, D. S., & Sarachik, E. S. (2000). Empirically derived Markov models and prediction of tropical Pacific sea surface temperature anomalies. *Journal of Climate, 13,* 3–17.

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Modern Physics, 41,* 253–262.

Lemos, M. C., & Rood, R. B. (2010). Climate projections and their impact on policy and practice. *Wiley Interdisciplinary Reviews: Climate Change*, *1*, 670–682. https://doi.org/10.1002/wcc.71.

Lloyd, E. A. (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.

Mears C. A., & Wentz F. J. (2017). A satellite-derived lower tropospheric atmospheric temperature dataset using an optimized adjustment for diurnal effects. *Journal of Climate.* Early online release https://doi.org/10.1175/JCLI-D-16-0768.1.

National Aeronautics and Space Administration (NASA). (2016). Independent Verification and Validation Framework. IVV 09-1, Version: P. Retrieved from https://www.nasa.gov/sites/default/files/atoms/files/ivv09-1-verp.doc.

Norton, S. D., & Suppe, F. (2001). Why atmospheric modeling is good science. In C. A. Miller & P. N. Edwards (Eds.), *Changing the atmosphere: Expert knowledge and environmental governance* (pp. 67–105). Cambridge, MA, USA: The MIT Press.

Oberkampf, W. L., & Trucano, T. G. (2002). *Verification and validation in computational fluid dynamics*, *SAND2002 – 0529*. Albuquerque, NM, USA: Sandia National Laboratories.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science, 263,* 641–646.

Post, D. E., & Votta, L. G. (2005). Computational science demands a new paradigm. *Physics Today, 58,* 35–41.

Roache, P. J. (1998). Verification of codes and calculations. *AIAA Journal, 36,* 696–702.

Roache, P. J. (2016). Verification and validation in fluids engineering: Some current issues. *ASME Journal of Fluids Engineering, 138*, 11.

Robock, A. (1983). El Chichón provides test of volcanoes' influence on climate. *National Weather Digest, 8,* 40–45.

Roesler, E. L., Posselt, D. J., & Rood, R. B. (2017). Using large eddy simulations to reveal the size, strength, and phase of updraft and downdraft cores of an Arctic mixed-phase stratocumulus cloud. *Journal Geophysical Research, 122,* 4378–4400.

Rood, R. B. (2010). The role of the model in the data assimilation system. In W. Lahoz, B. Khattatov, & R. Menard (Eds.), *Data assimilation: Making sense of observations* (pp. 351–379). Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-540-74703-1_14.

Roy, C. J., & Oberkampf, W. L. (2011). A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering, 200,* 2131–2144.

Santer, B. D., Solomon, S., Pallotta, G., Mears, C., Po-Chedley, S., Fu, Q., et al. (2017). Comparing tropospheric warming in climate models and satellite data. *Journal of Climate, 30,* 373–392.

Shackley, S. (2001). Epistemic lifestyles in climate change modeling. In C. A. Miller & P. N. Edwards (Eds.), *Changing the atmosphere: Expert knowledge and environmental governance* (pp. 107–133). Cambridge, MA, USA: The MIT Press.

Strang, G. (1968). On the construction and comparison of difference schemes. *SIAM Journal on Numerical Analysis, 5,* 506–517.

Stajner, I., Winslow, N., Rood, R. B., & Pawson, S. (2004). Monitoring of observation errors in the assimilation of satellite ozone data. *Journal Geophysical Research, 109,* D06309. https://doi.org/10.1029/2003JD004118.

Sundberg, M. (2011). The dynamics of coordinated comparisons: How simulationists in astrophysics, oceanography and meteorology create standards for results. *Social Studies of Science, 41,* 107–125.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal Geophysical Research, 106,* 7183–7192.

Theurich, G., DeLuca, C., Campbell, T., Liu, F., Saint, K., Vertenstein, M., et al. (2016). The earth system prediction suite: Toward a coordinated US modeling capability. *Bulletin of the American Meteorological Society*, *98*, 1229–1247.

Williams, D. N., Balaji, V., Cinquini, L., Denvil, S., Duffy, D., Evans, B., et al. (2016). A global repository for planet-sized experiments and observations. *Bulletin of the American Meteorological Society, 98,* 803–816. https://doi.org/10.1175/BAMS-D-15-00132.1.

Chapter 31
Validation of Agent-Based Models in Economics and Finance

Giorgio Fagiolo, Mattia Guerini, Francesco Lamperti, Alessio Moneta and Andrea Roventini

Abstract Since the survey by Windrum et al. (Journal of Artificial Societies and Social Simulation 10:8, 2007), research on empirical validation of agent-based models in economics has made substantial advances, thanks to a constant flow of high-quality contributions. This Chapter attempts to take stock of such recent literature to offer an updated critical review of the existing validation techniques. We sketch a simple theoretical framework that conceptualizes existing validation approaches, which we examine along three different dimensions: (i) comparison between artificial and real-world data; (ii) calibration and estimation of model parameters; and (iii) parameter space exploration. Finally, we discuss open issues in the field of ABM validation and estimation. In particular, we argue that more research efforts should be devoted toward advancing hypothesis testing in ABM, with specific emphasis on model stationarity and ergodicity.

Keywords Agent-based models · Validation · Calibration · Sensitivity analysis · Parameter space exploration

JEL codes C15 · C52 · C63

G. Fagiolo (✉) · M. Guerini · F. Lamperti · A. Moneta · A. Roventini
Istituto di Economia, Scuola Superiore Sant'Anna, Pisa, Italy
e-mail: giorgio.fagiolo@santannapisa.it

M. Guerini · A. Roventini
OFCE - Sciences Po, Paris, France
e-mail: mattia.guerini@sciencespo.fr

F. Lamperti
FEEM, Milano, Italy
e-mail: francesco.lamperti@santannapisa.it

A. Moneta
e-mail: alessio.moneta@santannapisa.it

A. Roventini
e-mail: andrea.roventini@santannapisa.it

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_31

31.1 Introduction

Modeling economies as complex systems using agent-based models (ABMs) is a relatively recent approach in economics (LeBaron and Tesfatsion 2008; Farmer and Foley 2009; Battiston et al. 2016; Turrell 2016). Nevertheless, since the 80s of the past century, it has increasingly been attracting many scholars belonging to several subfields, becoming both a complement and a substitute for more traditional economic-modeling methodologies (Schelling 1969, 1971; Epstein and Axtell 1996; Axelrod 1997). For example, ABMs are nowadays considered as a valid and effective competitor of standard Dynamic Stochastic General Equilibrium (DSGE) models in macroeconomics (see e.g., Dosi et al. 2010; Assenza et al. 2015; Hommes 2013, and the survey Fagiolo and Roventini 2017). Likewise, ABMs of financial markets are often considered better than traditional models, which are based on the efficient-market hypothesis in explaining the statistical properties of buy-and-sell high-frequency dynamics (cf. e.g., Franke and Westerhoff 2012; Leal et al. 2016).

Existing literature agrees that ABMs in economics provide two main added values, as compared to their orthodox counterparts (Dawid and Delli Gatti 2018; Dieci and He 2018). First, ABMs allow for more descriptive richness, as they describe ecologies of agents, locally interacting through nonobvious network structures, learning to use incomplete information, and competing within imperfect markets. Second, the modeler developing an ABM has typically more flexibility in both input and output validation of its model.

This second feature of ABMs has always attracted a lot of attention and has generated, especially in the past years, a booming number of contributions. Back in 2007, the influential article by Windrum et al. (2007) attempted to survey ABMs validation methods, concluding that a lot of work would have been needed in order to fully develop a satisfactory set of techniques that consistently take ABMs to the data. In fact, many developments have occurred in the past 10 years, which this Chapter tries to review. We go along such developments distinguishing between three different dimensions: (i) calibration and estimation of model parameters; (ii) comparison between artificial and real-world data.

The chapter is organized as follows. First, we offer an introduction to the most-diffused practices in building and running agent-based models in economics (Sect. 31.2). In Sect. 31.3, we also sketch a simple theoretical framework that conceptualizes existing validation approaches. Sect. 31.4 provides a critical review of the literature, whereas in Sect. 31.5 we describe the most recent trends as to validation techniques in ABMs. Finally, Sect. 31.6 concludes with some critical considerations on future work.

31.2 Agent-Based Computational Economics: Common Practices

Notwithstanding the existence of different types of agent-based models, which have been developed by various subfields within economics, such as macroeconomics, industrial dynamics, finance, asset pricing, etc., one can identify some general patterns and common practices in the building process, under a common umbrella that we refer to as Agent-Based Computational Economics (ACE).

31.2.1 The Development of a Typical Agent-Based Model

Researchers typically do not know the *true* data generating process of phenomena under study, which we refer to as the real-world DGP ($rwDGP$). This can be seen as a very complicated, multiparameter, stochastic process that governs the generation of a unique realization of some time series and stylized fact that we can empirically observe and estimate. The goal of the modeler is, therefore, to provide a sufficiently good approximation of the $rwDGP$ by using an ABM. Naturally, the model releases a simplified DGP, which we refer to as the model-DGP ($mDGP$) and which should provide a meaningful explanation of the causal mechanisms generating the set of observed stylized facts, and, more generally, a good representation of the data (Heine et al. 2005). The empirical validation of an ABM is then the process by which one evaluates the extent to which the $mDGP$ is a good representation of the $rwDGP$.[1]

The most adopted procedure for the development of an ABM is the *indirect calibration approach* (see Windrum et al. 2007).[2] This procedure is composed of four separate steps. The first consists in the identification of some real-world stylized facts of interest that the modeler wants to explain. In the second, one specifies the model, the time line of the events, the microlevel dynamic equations which embody the individual agents' behavior, the set of parameters, and the set of random disturbances. Validation and the hypothesis testing are performed in the third step in order to compare model's output with the observations obtained from real- world datasets. Finally, there could be a fourth step, where the ABM is employed for policy analysis exercises, implemented by changing some of the behavioral equations (e.g., capital requirements for macroprudential policy, as in Popoyan et al. 2017) or some of the parameters (e.g., tax rates for fiscal policies, as in Dosi et al. 2010). In what follows, we will explore these four steps in more detail.

Stylized facts identification. The starting point of most ABMs is the identification of a set of micro and macro stylized facts and empirical regularities (e.g., static or dynamic correlations, empirically observed distributions, etc.); see also Chap. 15 by Meyer in this volume. For the sake of generality, let us define as a stylized fact any possible type of measurable unconditional object that can be investigated by means of some econometric exercises or more generally by statistical techniques.

[1] The validation process might also take different perspectives. In particular, as reported by Burton and Obel (1995), the model's assumptions and abstractions have to be judged accordingly with the model's purpose. In this paper, we mostly focus on validation of policy-oriented, descriptive agent-based economic and financial models.

[2] However, also other viable strategies are available: see, for example, the calibration approach proposed by Werker and Brenner (2004); Brenner and Werker (2007) and the history friendly models developed by Malerba et al. (1999).

In such unconditional objects (see Brock 1999), the causal generating mechanism, or data generating process (DGP), is unclear or too complex to be explained by a simple, low-dimensional system of dynamic equations. Examples of micro and macro stylized facts that have been empirically identified and replicated by means of ABMs in different fields encompass fat-tailed distributions of returns, endogenous emergence of flash crashes, long-run coexistence of heterogeneous investing rules in finance; fat-tailed distributions of firm growth rates, Zipf distribution of firm size, negative correlations between prices and market shares in monopolistic competitive markets in industrial dynamics; investment lumpiness, Okun and Beveridge curve, cyclical co-movements of variables in macroeconomics.

Model specification. After having singled out a set of possibly interlinked stylized facts, one can try to find an explanation of the underlying causal forces, i.e., learning and describing the exact form of the real-world DGP, or at least a sufficiently accurate approximation of it. This is the ultimate objective of any ABM. The great advantage of ABMs vis-á-vis traditional ones commonly employed in economics and finance derives from its generative bottom-up approach genuinely rooted in evolutionary, complex-system theories (more on that in Lane 1993; Tesfatsion 2006; Fagiolo and Roventini 2012, 2017). This indeed allows the researcher to take into account the complex dynamics of a system that is populated by heterogeneous and boundedly rational agents possessing a partial and possibly biased amount of information about the global system in which they live. However, agents are adaptive and learn in order to survive and prosper in such an uncertain framework following some forms of "Simonesque" (see Simon 1991) satisfying principle.[3] Obviously, also when ABMs are developed to approximate the $rwDGP$, the number of degrees of freedom is high and different researchers can follow alternative routes according to their different expertise, backgrounds, and theoretical hypothesis about the underlying generating process.[4]

Output validation. After the modeler has specified the behavioral equations of the actors populating the system, the ABM takes the form of a high- dimensional, discrete-time stochastic process. Indeed, a part of the ACE community (especially in financial and asset pricing models) has strongly relied on Markov processes theory and on statistical physics tools in order to reduce the dimensionality of the model and eventually—under specific circumstances—to analytically solve the simplest model. But in general, as their complexity is high, ABMs are usually simulated by means of extensive Monte Carlo (MC) exercises in which the random seed is

[3] In that there is a major departure with respect to neoclassical models, where the (representative) agent has axiomatic preferences and maximizes some smooth objective function with an easily computable bliss point.

[4] This is also one of the critiques that is usually addressed to ACE. Since ABMs do not stick to some generally accepted axiomatic rule of behavior, they introduce discretionary choices that the modeler shall take. We will see how practitioners have coped with this issue in Sect. 31.4.2.1. A possible solution to discipline the construction phase of an ABMs has been put forward by Grimm et al. (2006) and is called the ODD protocol (from "Overview, Design concepts, and Details").

modified along the MC dimension.[5] Once such MC exercises are performed and the synthetic data collected, the researcher can verify whether the model is able to generate unconditional objects which are not statistically significantly different from the ones previously observed in real-world datasets.[6] Naturally, all these unconditional objects can be related to micro and macro variables.

Policy analysis. Once the model has been validated and proved to be able to account for the micro and macro empirical regularities under study, it can then be employed as a policy laboratory. Indeed, the impact of different economic policies in alternative scenarios can be studied by (i) varying some parameters, in particular, those related to policy-maker interventions or to some broad institutional setting (e.g., tax rates); (ii) modifying initial conditions related to agents' state variables (e.g., income distribution, firms' technology); (iii) changing some agents' behavioral rules and interaction patterns (accounting e.g., for different market setups); (iv) introducing macro and/or micro heterogenous shocks (e.g., innovation or climate-damages shocks). These can be interpreted as *exogenous* policy changes, which allow a researcher to evaluate their effects in a fully controllable environment, where treatment effects can be easily identified, and endogeneity issues are almost absent.

In what follows, we will focus mostly on validation, discussing more in depth what is the relationship between an ABM, its inputs and outputs. In particular, the interpretation of the ABM as a process that transforms a set of inputs into outputs, poses two relevant questions: (i) how a *ceteribus paribus* variation of one input affects the output (a detailed discussion will be presented in Sect. 31.4.2.1), and (ii) to which extent the generated output is a good approximation of the real-world phenomenon that the modeler aims to explain (discussion in Sect. 31.4.2).

31.2.2 Inputs of Agent-Based Models

In ABMs, we can characterize two broad categories of inputs: initial conditions and parameters.

Initial conditions. They determine the values of macro and agents' state variables at the beginning of the simulation. In small- scale ABMs, which are typically characterized by a deterministic skeleton and may possess at least one computable deterministic fixed point, initial conditions can be set at the equilibrium or in some contour of it (see Brock and Hommes 1997; Westerhoff and Dieci 2006; Guerini 2013; Guerini et al. 2017) and then the ABM can be used to locally study the dynamics of the system.[7] However, in complex stochastic models, characterized by high

[5] As stated in Turrell (2016), the first agent-based model was developed in the 30s by the physicist Enrico Fermi in order to study the transport of neutrons through matter. Fermi's agent-based technique was later called Monte Carlo method (Metropolis and Ulam 1949).

[6] In Sect. 31.4.2, we will discuss the tools available for the verification and validation of ABMs.

[7] One can also study the basins of attraction of the dynamical system to study the robustness with respect to initial conditions.

levels of dimensionality, fixed points or statistical equilibria may not exist or not being known to the modeler. In such a framework, the selection of initial conditions can become a nontrivial issue, affecting the ergodicity and dynamics of the system, its output and more generally the very validity of the model. Different solutions are proposed in the ACE literature. The first one initializes the model in a homogeneity situation, where all the state variables of the agents are set equal to some economically meaningful values (see Dosi et al. 2010, 2013, 2015). The second solution instead draws initial conditions of a category of agents from a specific distribution, possibly grounded on some empirical regularity (e.g., fat-tailed firm size distribution as in, e.g., Bianchi et al. 2007, 2008a). Finally, if rich enough datasets are available, one can employ them to directly impute initial conditions values (Hassan et al. 2008).

Parameters. They can fix some macro conditions, determine the size of agents' reactions to events, or they characterize the distributions from which stochastic decisions are taken by agents or shocks are drawn. In many economic and finance ABMs, parameters are of particular interest because they might drive the dynamic of the system to different statistical equilibria, they characterize some specific policy relations or some particular institutional arrangement that the modeler wants to investigate. Parameters are usually calibrated, or they can be estimated if appropriate data is available (see Sect. 31.4.1). Moreover, several methods allow to perform *sensitivity analysis* exercises in order to map the model responses to parameter variations (see Lee et al. 2015; Dosi et al. 2017c). These techniques will be discussed in more detail in Sect. 31.5.

31.2.3 Outputs of Agent-Based Models

ABMs can generate both microlevel and aggregate outputs.

Microlevel output. The output of an ABM is composed of MC (the number of Monte Carlo simulations) panel datasets containing different micro variables for a set I of agents over a specified time window T. Therefore, the data can be collected in the following form:

$$Z_{m,k} \in \mathbb{R}^{K \times MC}, \quad Z_{m,k} = \left\{z_{m,k,i,t};\ i = 1, \ldots, I;\ t = t_0, \ldots, T\right\}, \quad \forall k \in K, \tag{31.1}$$

where m denotes a specific Monte Carlo run, k indicates a micro-variable, i represents the agent cross-section dimension, and t captures the time dimension. As an example, in a macroeconomic ABMs these variables can represent household income or consumption levels, firm prices, capital, profits, etc.

Aggregate output. The output of each Monte Carlo simulation m contains also some time-series variables, which emerge from the aggregation along the agent cross-section dimension. These aggregate series (denoted by an upper bar) take the following form:

$$\bar{Z}_{m,h} \in \mathbb{R}^{H\times MC}, \quad \bar{Z}_{m,h} = \left\{\bar{z}_{m,h,t};\ t = t_0, \ldots, T\right\}, \quad \forall h \in H, \tag{31.2}$$

where h denotes the macro variable observed at different time steps t. For example, in a macroeconomic ABM, one can aggregate the micro variables concerning households, firms to compute GDP, price indexes, or the unemployment level.

31.2.4 Relation Between Input and Output

For micro and for aggregate variables, the simulated synthetic datasets can generate a number of stylized facts or statistical properties that the modeler can compare with those obtained from the empirical analysis of the corresponding real-world dataset. This is the core of the indirect calibration approach presented above and the first validation test that an ABM must satisfy. The similarity between model-generated and real-world data constitute the essence of the validation problem for ABMs, and it will be extensively discussed in Sects. 31.4 and 31.5. For the moment, let us only anticipate that in the past decade, different strategies have emerged tackling a set of related, but slightly different issues.

For any validation method, one should consider that in ABMs, the set of generated micro and macro variables $\left\{Z_{m,k}, \bar{Z}_{m,h}\right\}$ are not intrinsic features of the model itself, but are emerging properties coming from the complex interaction between model institutional arrangements and model inputs. Therefore, the statistical properties of the output might exist only conditionally on the selected initial conditions, parametrization, the chosen random seed, and the selected institutional arrangement. This means that a stylized fact that has been obtained under a specific set of inputs, might not necessarily hold true under different arrangements, and robustness analysis must be performed before using ABMs for policy analysis exercises.

31.3 Agent-Based Model Validation: Theoretical Framework

Validation of computer simulation models encompasses a variety of interrelated issues and concepts. Manson (2002) distinguishes between *output validation* and *structural validation*. The latter asks how well the simulation model represents the (prior) conceptual model of the real-world system, while the former assesses how successfully the simulations' output exhibits the historical behaviors of the real-world target system. Further, output validation can be directly related to what Leombruni et al. (2006) define as *empirical validity* of a model, i.e., validity of the empirically occurring true value relative to its indicator. Following Rosen (1985), let us consider two parallel unfoldings: the evolution of the system (an economy, a market, etc.) and the evolution of the model of the system. If the model is correct, properly calibrated and initial conditions have been fixed according to the initial status of the real system, the simulation should mirror the historical evolution of the real-world system with respect to the variables of interest. This is exactly the assessment of the relationship between simulated and empirical data that constitutes the focus of this chapter. However, there are many other validity issues that we do not explicitly address. For example, Leombruni et al. (2006) discuss *theoretical validity* (the validity of the theory relative to the simulation), *model validity* (the validity of the model relative to the theory), *program validity* (the validity of the simulating program relative to the model), and *operational validity* (the validity of the theoretical concept to its indicator or measurement).[8]

Using Fig. 31.1, let us present our view and definition of model validation (cfr. Windrum et al. 2007). Assume that the modeler knows (from a preliminary simulation study, or from some ex ante knowledge about the model under scrutiny) that the real-world system is ergodic, and that the $rwDGP$ displays a sufficiently stationary behavior for a time period after T^*. Further, let us assume that for a particular set of initial conditions, micro and macro parameters, the $mDGP$ runs until it reaches some form of stable behavior, which can be further summarized by a set of statistics $S = \{s_1, s_2, \ldots\}$. Thus, each realization of the model will produce different values of the summary statistics s_j. Then, one must perform a sufficiently high number of independent Monte Carlo runs to estimate the distributions of those statistics, from which moments can be computed.[9] Such moments will depend on the initial choices that were made in terms of parameters and initial conditions. However, by exploring a sufficiently large number of points in the space of initial conditions and parameter values, and by computing, at least, the first two moments ($E(s_j)$ and $V(s_j)$) at each point, one could gain a deep understanding of the behavior of the model, test the robustness of the results, and identify the set(s) of parameters providing the most relevant dynamics. Modelers and practitioners can make use of the uniquely observable real-world micro and macro time series and, under the assumptions outlined above, compute their longitudinal moments. Hence, the statistical properties of the artificial data and of the real-world can be compared and this constitutes what we call *empirical validation*. In that, a relevant issue concerns the availability of suitable real-world data for validation; in general, the economic and financial literatures tend to use "macro-level" data (e.g., time series of GDP or stock prices). In the future, we believe that the increased data availability and computing power will push toward the systematic inclusion of more informative microlevel data, i.e., data at the level of agents.

Summing up, validating a simulation model amounts to assessing the extent by which the model structure represents the data generating process that underlies the societal reality but that cannot be directly observed. Validation involves use of both

[8] In agent- based modeling, some of the standard validity aspects that are relevant in many fields of numerical simulations are not an issue; for example, systems are always represented in discrete time and, hence, discretization errors are not possible. Further, low emphasis is usually posed on code verification.

[9] See also Secchi and Seri (2017) on the issue of selecting the number of times a computational model should be run.

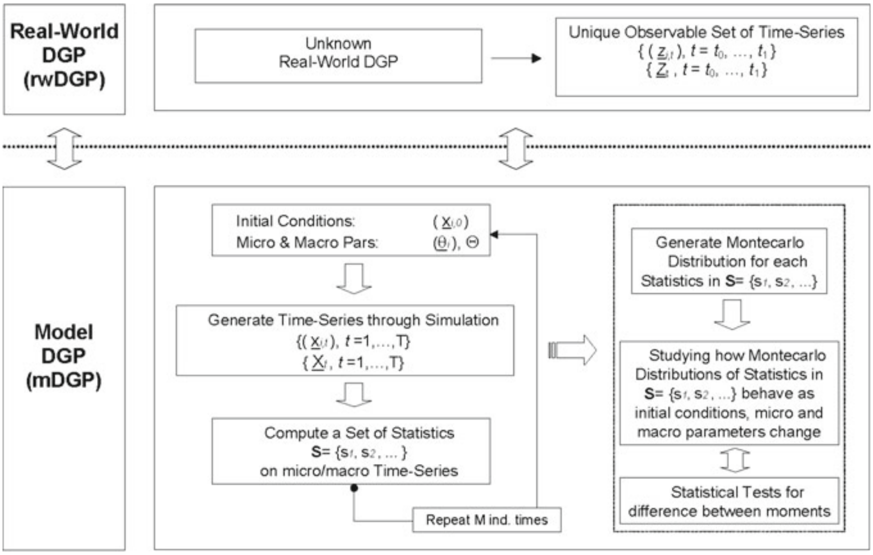

Real-World DGP (rwDGP)
Unknown Real-World DGP
Unique Observable Set of Time-Series
$\{(\underline{z}_{i,t}), t = t_0, \dots, t_1\}$
$\{\underline{Z}_t, t = t_0, \dots, t_1\}$
Model DGP (mDGP)
Initial Conditions: $(\underline{x}_{i,0})$
Micro & Macro Pars: $(\underline{\theta}_i), \Theta$
Generate Time-Series through Simulation
$\{(\underline{x}_{i,t}), t = 1,\dots,T\}$
$\{\underline{X}_t, t = 1,\dots,T\}$
Compute a Set of Statistics
$\mathbf{S} = \{s_1, s_2, \dots\}$
on micro/macro Time-Series
Repeat M ind. times
Generate Montecarlo Distribution for each Statistics in $\mathbf{S} = \{s_1, s_2, \dots\}$
Studying how Montecarlo Distributions of Statistics in $\mathbf{S} = \{s_1, s_2, \dots\}$ behave as initial conditions, micro and macro parameters change
Statistical Tests for difference between moments

Fig. 31.1 A procedure to study the output of ABMs

artificial data derived from the model and real observational data. Validation, in this general sense, differs from a strict notion of both demonstration and falsification, because our approach is not framed in a binary framework (i.e., to reject vs. not to reject the model), but comes in degrees, which also allows us to judge how a model performs empirically relatively better than another (for a discussion of falsification and validation see also Chap. 5 by Beven and Lane and Chap. 26 by Roache in this volume). In many circumstances, indeed, what is needed by the practitioner is, instead of a response on whether to accept or reject a model, a measure of how good the model is, which allows the researcher to choose among alternative model specifications.

31.4 Agent-Based Model Validation: Literature Review

The most general classification scheme for agent-based models (ABM) according to their level of empirical validity has been proposed by Axtell and Epstein (1994) and Barde and van der Hoog (2017) and consists of four levels:

- **Level 0**: the model is a caricature of reality, as established through the use of simple graphical devices (e.g., allowing visualization of agent motions).
- **Level 1**: the model is in qualitative agreement with empirical macrostructures, as established by plotting, e.g., the distributional properties of agent population. This is the easiest way to matching stylized facts.

- **Level 2**: the model produces quantitative agreement with empirical macrostructures, as established through on-board statistical estimation routines.
- **Level 3**: the model exhibits quantitative agreement with empirical microstructures, as determined from cross-sectional and longitudinal analysis of the agent population.

Publishing standards for economic and finance ABMs require at least that satisfaction of Level 1.[10] Under the Level 1 approach, an agent-based model gets validated through a statistical comparison of unconditional objects: stylized fact observed in real-world data and emergent properties derived from the simulation environment. This amounts, therefore, at replicating a large number of possible micro and macro stylized facts characterizing the phenomena of interest.[11]

Current developments in empirical validation of ABMs shows a progression from Level-1 to Level-2 models, as the mere replication of empirical regularities and other unconditional objects are increasingly replaced or supplemented by quantitative estimation. Such a fresh stream of research requires the models to generate series that exhibit the same dynamics (Marks 2007; Lamperti 2018b), the same conditional probabilistic structure (Barde 2016b), or the same causal structures (Guerini and Moneta 2017) as those observed in the real-world data. Furthermore, new methods to estimate and calibrate the parameters of ABMs have been developed with the aim of minimizing the distance between some distributional properties of the real simulation outcomes.

We claim that such new contributions will bring agent-based models on the same ground of advancement of the DSGE literature. Indeed, the emerging literature on validation and estimation of ABMs represents the ACE counterpart of the progresses occurred in the estimation of DSGE models and well represented by the works of Del Negro and Schorfheide (2006); Canova and Sala (2009); Paccagnini (2010); Fernández-Villaverde et al. (2016).

Notwithstanding the possible overlaps between calibration, estimation, and validation strategies,[12] in what follows we propose a classification based on the central aim of each procedure. We, therefore, present in Sect. 31.4.1 calibration and estimation procedures, which are essentially exercises for tuning model parameters or understanding the likelihood that a parameter is responsible for simulation results. We then discuss in Sect. 31.4.2 the validation procedures, which evaluate how the inputs or outputs of simulated models resemble some well-defined real-world statistical properties.

[10] Level 0 models can be somehow accepted if their aim is merely exploratory rather than descriptive.

[11] See, for example, Dosi et al. (2010, 2013, 2015, 2016a) for replication of business cycle and growth stylized facts; Dosi et al. (2017a) for accounting of labor-market micro and macro regularities; Popoyan et al. (2017) for the reproduction of many credit and interbank market properties; Lamperti et al. (2018a, b) for capturing coevolution of economic fundamentals with energy and emission quantities; Pellizzari and Dal Forno (2007); Leal et al. (2016) for simulating financial market booms and busts.

[12] For a discussion of calibration and testability, see Chap. 40 by Frisch in this volume.

31.4.1 Calibration and Estimation

Notwithstanding the fuzzy difference between calibration and estimation, in what follow we discuss the two approaches as both aim at solving the same class of problems (in line with Hansen and Heckman 1996). Calibration and estimation exercises have peculiar difficulties in Agent-Based modeling: the complex microeconomic interactions and the presence of ubiquitous nonlinearities (even in the simplest models) do not allow one to obtain a closed-form solution of the likelihood function and of the moments conditions. Therefore, one must resort to indirect inference or other simulation methods.

31.4.1.1 Indirect Inference

Indirect inference (Gourieroux et al. 1993) is the standard approach that has been developed for the estimation of small-scale agent-based models, characterized by relatively few parameters and short computational time.[13] Indirect inference allows one to estimate or to make inferences about the parameters of a model by means of simulation methods. It has been considered the preferred estimation choice since the very first ABM estimation attempts (see e.g., Winker and Gilli 2001, 2004).[14] Also in the ABM framework, one could try to employ the Generalized Method of Moments approach (GMM), as in the very stylized models by Alfarano et al. (2005, 2006). However, in most of financial and economic Agent-Based Models, the moment function is completely unknown and one has to approximate it via Monte Carlo simulation exercises. In such a framework, the consistency and efficiency of the parameters estimates strongly depend on the level of approximation of the moment generation function.

Following Chen et al. (2012), the procedure of the Method of Simulated Moments (MSM) can be summarized as follows:

> We first choose a vector of parameter values to generate the simulated time series by running the agent-based model with this chosen set of parameter values. We then compare some statistics (moments) of this simulated time series, the simulated moments, with those using real data, the sample moments. The difference between the two is used to form a distance function (the objective function). The MSM is purported to minimize the distance by searching over the entire parameter space.

Formally, one must estimate the vector of parameters θ^* that solves the following minimization problem:

$$\text{argmin}\ \mathcal{L}(X^{RW}, X^{AB}; \theta) \tag{31.3}$$

[13] Benchmark models are, for example, the Brock and Hommes (1998) asset pricing model and the Kirman (1991) speculative bubbles model.

[14] See also Boswijk et al. (2007); Bianchi et al. (2008b); Goldbaum and Mizrach (2008); Franke (2009); de Jong et al. (2010); Franke and Westerhoff (2012); Chiarella et al. (2014); Platt and Gebbie (2016).

where X^{RW} and X^{AB} represent, respectively, the set of chosen moments observed in the real-world data and their counterpart derived from the ABM simulation.

This procedure is sufficiently general and in principle it is applicable to any type of ABM, but three drawbacks make it unfeasible in practice when the model complexity increases, and the simulation time becomes a relevant constraint. First, an analytical solution for the problem of minimization of the approximated distance function is rarely available, forcing one to rely on numerical approximations. Second, moment selection is arbitrary and different choices may lead to differing estimation results. Third, the procedure is computationally intensive as one needs to run a sufficient number of Monte Carlo simulations of the model for each instance of the parameter space, and then evaluate the distance between the generated moments and those observed in sampled real-world data.

Very close alternatives to the MSM for estimating an agent-based model is the Simulated Minimum Distance (SMD) approach, which has been adopted by Fabretti (2013) and by Grazzini and Richiardi (2015) and the Simulated Maximum Likelihood (SML) by Kukacka and Barunik (2017).

31.4.1.2 Bayesian Approaches

As documented in the previous section, most of estimation and calibration works have been following a frequentist approach. However, after the popularization of Bayesian methods for the estimation of DSGE models (see Fernández-Villaverde and Rubio-Ramírez 2007; Fernández-Villaverde et al. 2016), Bayesian inference techniques for estimating ABMs have been introduced in Grazzini et al. (2017). In general, the adoption of Bayesian strategies should reduce the discretionary choices involved in the somehow ad hoc selection of the moments to be taken into consideration, the auxiliary model, or in any other metric that allows to evaluate the distance between the real and the simulated time series. Moreover, Bayesian approach could be more asymptotically efficient as it exploits the information provided by the whole distribution of the data and not only those of some specific moments.

However, Bayesian methods are not exempted from these issues. First, as documented by Canova and Sala (2009) and Fagiolo and Roventini (2012, 2017), the selection of the prior distribution can possibly generate an artificial curvature to the posterior distribution, when the likelihood tends to be flat, thus ending up in an interval calibration exercise. Second, the computational cost of Bayesian techniques is especially high when they are applied to ABMs for estimating the likelihood function. Such computational costs can be reduced by adopting efficient sampling schemes or likelihood function approximations, whose appropriateness should be evaluated on a case-by-case basis. However, as ABMs do not typically have closed-form solutions, a large number of Monte Carlo instances still need to be simulated (see Lamperti et al. 2018c).

31.4.2 Validation

31.4.2.1 Input Validation

The main focus of input validation has been (i) testing some of the behavioral assumptions typically included in Agent-Based models; (ii) selecting the initial conditions of the model under investigation; and (iii) exploring the parameter space. Let us now consider each of them.

Selection of behavioral rules. Well in line with behavioral economics (see e.g., the seminal contribution of Kahneman and Tversky 1979), the very first input validation exercises of ABMs has resorted to laboratory experiments, which allow the researcher to directly verify how an individual behaves in a controlled environment. Typically, these experiments have been used to test specific assumptions on agents' behavior embedded in small-scale ABMs (see Hommes 2011, 2013; Anufriev et al. 2016). Later, controlled laboratory experiments have been employed to estimate heuristic switching models (as in Anufriev and Hommes 2012; Assenza et al. 2013).

Instead, in more complex ABMs, specific behavioral assumptions cannot be directly tested, and other approaches have been adopted. We present here three of them that have allowed researchers to reduce the problem known in the literature as the "wilderness of bounded rationality." In the *adjustment heuristics* approach (Gaffeo et al. 2008; Assenza et al. 2015; Guerini et al. 2017), economic agents follow very basic economic principles in order to set some of their state variables. For instance, in these models, prices are fixed by the principle of excess demand. In the *management science* approach (see Dawid et al. 2016), the decisions of agents are modeled starting from the researches carried out in the management literature. More specifically, consumers and firm behaviors are modeled following, respectively, the indications provided by the marketing and firm strategy literature. Finally, the *empirical microeconomics* approach attempts to model the behavior of agents relying on microeconomic empirical evidence. This is the case, for example, in the "Schumpeter meeting Keynes" stream of models (Dosi et al. 2010, 2013, 2015).

Selection of initial conditions. Input validation can concern the selection of initial conditions of the model. Even simple and deterministic ABMs can display chaotic dynamics, wherein small deviations between two configurations may generate extremely different time series (see Brock and Hommes 1997, 1998; Hommes 2013). However, if the model is ergodic, it explores the whole state space and reaches a stationary distribution. The problem of sensitive dependence on initial conditions can be tackled in small- scale models, which are typically analytically solvable and where boundaries conditions and basins of attraction can be easily studied. On the contrary, it is still an open issue in more complex models, where the large support from which initial-condition values can be drawn implies huge computational costs.

Exploration of the parameter space. Apart from parameter estimation and calibration, which have been thoroughly discussed in the previous section, in Agent-Based models one may need to explore the parameter space in order to assess the impact of different parameters on the dynamics of the model and to perform policy analysis exercises. An increasing number of works have started to investigate the robustness[15] of a model by running Monte Carlo simulations under different parameter settings (Ciarli 2012; Salle and Yıldızoğlu 2014; Bargigli et al. 2016; Dosi et al. 2016b, 2017b, c). More on that in Sect. 31.5.3.

31.4.2.2 Output Validation

As introduced in Sect. 31.3, output validation is the process of evaluating the extent to which the outcome of a simulated model is a good representation of real-world observations. The baseline evaluation process focussing on the replication of stylized facts has been naturally embedded in most of Agent-Based models, which are often designed to account for phenomena unexplained by analytically tractable models.

Recently, more sophisticated statistical techniques have been developed to satisfy more stringent output validation requirements. In particular, they try to account for the "unconditional object" critique in Brock (1999) and to better discriminate among different ABMs reproducing the same set of stylized facts.

For instance, Marks (2013) employs three similarity measures—the Kullback–Leibler, the State Similarity Measure, and the Generalized Hartley Metric—to analyze and validate an ABM of brand rivalry in the general validation framework developed in Marks (2007)[16]. Barde (2016a, b) and Lamperti (2018a, b) develop two new similarity measures based on information theoretic criteria. Guerini and Moneta (2017) instead measure similarity by comparing the causal relations entailed in a Structural Vector Autoregression model estimated on both real and simulated data (cfr. Sects. 31.5.1 and 31.5.2). Following Grimm et al. (2005), all these compare model and data according to the patterns conceiving relevant information on the system under scrutiny.

Note that all these recently developed validation techniques focus only on aggregate time series, while most of ABMs have been been able to replicate *both* micro and macro stylized facts. We believe that the next challenge is to further extend the new approaches to validate ABMs also in terms of microeconomic behaviors.

31.5 A New Wave of Validation Approaches

The debate on ABM validation is still an open one and a novel wave of approaches has recently blossomed, offering to modelers and policymakers additional tools for the analysis of their models. This section outlines and discusses some of these contributions in relation to existing gaps in the literature.

[15] For robustness of the model, we here mean the stability of the results to small variations of the parameters. See also Lorscheid et al. (2012) and Thiele et al. (2014).

[16] See also Chap. 12 by Marks in this volume.

31.5.1 Validation As Replication of Time Series Dynamics

Output validation concerns the assessment of how successfully simulations from a model mirror the historical behavior of the real-world target system (cf. Sect. 31.3). In practice, this amounts at evaluating the degree of similarity between two or more time series. In most applications, the method of simulated moments and simulated minimum distance are used, but as we argued before, these might have some shortcomings when applied to ABMs. In order to overcome these shortcomings, Lamperti (2018b) has recently proposed a novel information theoretic criterion, called Generalized Subtracted L-divergence (*GSL-div*), that measures the degree of similarity between the dynamics observed in real data and those produced by the numerical simulation of a model. Contrary to simple summary statistics, the *GSL-div* has been constructed to compare time series on the basis of their patterns. Validation is achieved capturing the ability of a given model to reproduce the distributions of time changes (that is, changes in the process' values from one point in time to another) observed in the real-world series, without the need to resort to any likelihood function or to impose requirements of stationarity. The *GSL-div* provides a precise quantification of the distance between the model and data with respect to their dynamics in the time domain.[17]

The *GSL-div* can be estimated numerically following a simple, four-step procedure.

1. Time series (both real and simulated) are symbolized.
2. Patterns of symbols are observed through rolling windows of different length $l = 1, .., L$.
3. Distributions of patterns, f_l, are estimated for each windows' length.
4. The distance between distributions from real and simulated data are evaluated through an information theoretic criterion and, finally, aggregated.

The *GSL-div* has been tested to discriminate among different classes of stochastic processes, going from simple Autoregressive–Moving-Average (ARMA) models to random walks with drifts and structural breaks. Systematic comparisons with alternative measures of fit commonly used for calibrating ABMs in economics and finance (e.g., mean squared error (MSE), distance between moments, etc.; for an overview of these measures see Chap. 17 by Saam in this volume) has revealed that the *GSL-div* provides much more satisfactory performances. Such results point the adequacy of the approach to quantify the degree a simulation model mirrors real-world data.

Lamperti (2018a) applies the described approach to the analysis of a widely used financial market model with heterogeneous traders. He finds that the GSL-div can further improve the validation of the model with respect to criterion grounded on the minimization of the mean squared error as in Recchioni et al. (2015).

[17] For other interesting approaches on pattern-based validation see Barde (2016b) and Marks (2018).

31.5.2 Validation as Matching of Causation

Discovering causal structures is a relevant task for at least two interconnected reasons: (i) it allows understanding and explaining the origin and propagation of phenomena that are observed at some point in time; (ii) it provides information on available policy channels to be used for impacting the system. Given such premises, Guerini and Moneta (2017) claim that models employed to provide policy prescriptions should match the causal relationships observed in the real systems they represent. They propose a procedure to validate a simulation model by estimating and comparing the causal structures incorporated in the model with those obtained from real-world data.

The causation matching approach proposed by Guerini and Moneta (2017) follows a sequential procedure that can be divided into five steps.

1. Data harmonization and preparation.
2. Analysis of ABM properties.
3. Estimation of Vector Autoregressive (VAR) models.
4. Identification of the Structural Vector Autoregressive (SVAR) models.
5. Validation assessment.

In the first step, some simple transformations are performed to allow the comparison of empirical and artificial data (e.g., cutting simulated series to make them equally long as their real-world counterpart, removing trend, etc.). In the second step, two emergent properties of the series produced by the simulated model are analyzed (e.g., stationarity and ergodicity tests). In the third step, the reduced-form VAR model is estimated via ordinary least squares or accounting for co-integrated variables via the Johansen and Juselius (1990) procedure. In the fourth step, the structural form of the model is identified by means of the so-called PC (in case of Gaussian residuals, Spirtes et al. 2000) or VAR-LiNGAM (if residuals are non-Gaussian, Shimizu et al. 2006 and Hyvarinen et al. 2010) causal search algorithms.[18] Finally, in the last step, the two estimated causal structures are compared according to simple distance measures.

Guerini and Moneta (2017) also apply their approach to the well-known K+S macroeconomic agent-based model developed in Dosi et al. (2015). Causal structures from model simulations are compared to those obtained from U.S. data for the period 1959–2014. Results show that the model is able to capture between 65% and 80% of the causal relations entailed by a SVAR estimated on real-world data. Such a positive finding could be then compared to the results obtained when the procedure is also applied to different agent-based and DSGE models. In that, the causality-matching validation test is highly complementary to GSL-div employed to assess the replication of time series dynamics (cf. Sect. 31.5.1).

[18] VAR-LiNGAM stands for Vector Autoregressive Linear Non-Gaussian Acyclic Model.

31.5.3 Global Sensitivity Analysis via Kriging Meta-Modeling

The understanding of model's response to (possibly joint) changes in some parameter values or initial conditions is pivotal to assess the robustness of models' output as well as to draw robust implications from policy exercises. It also allows understanding whether variations in some parameters drive the model away from the empirical reality. However, sensitivity analysis in ABM can often involve high computational costs stemming from simulating the model for many vectors of parameters, initial conditions, and seeds of the pseudorandom number generating process. Salle and Yıldızoğlu (2014) have been the first to propose the combination of design of experiments (DoE) and kriging meta-modeling (Krige 1951; Van Beers and Kleijnen 2004) to address the issue within the economics literature. The strategy they propose is straightforward: DoE allows to minimize the sample size of parameter configurations under the constraint on their representativeness. The original model is, therefore, approximated with a meta-model, which is then employed to connect the parameters to the variables of interest at virtually zero computational costs. The meta-model is indeed a simplified version of the original model that can be more parsimoniously run to evaluate the effect of inputs (parameters) on model's output.

Building on such an approach, Dosi et al. (2017b, c) provide a global sensitivity analysis for a relatively simple model of industry dynamics and for a more complicated macroeconomic model. Their procedure runs as follow:

1. employ nearly orthogonal latin hypercubes (NOLH) to sample the parameter space;
2. develop a kriging meta-model (KMM) to approximate the original ABM[19];
3. perform Sobol variance decomposition to analyze the meta-model sensitivity to parameters;
4. draw three-dimensional surfaces to represent the response of the variable of interest in the meta-model to changes in parameters.

For example, Dosi et al. (2017c) study a model of industrial dynamics investigating how the distribution of firms' growth rates changes in response to different input variations ranging from the relevance of learning mechanisms to the strength of the selection process among competing firms. Kriging and Sobol decompositions have also been successfully employed to the more complex K+S model agent-based model to study the impact of structural reforms in the labor market (Dosi et al. 2016b) and to detect the emergence of hysteresis (Dosi et al. 2017b).

[19] Coupling NOLH with kriging meta- modeling has been frequently used to approximate the output of computer simulation models (see, for example, McKay et al. 1979; Salle and Yıldızoğlu 2014; Bargigli et al. 2016).

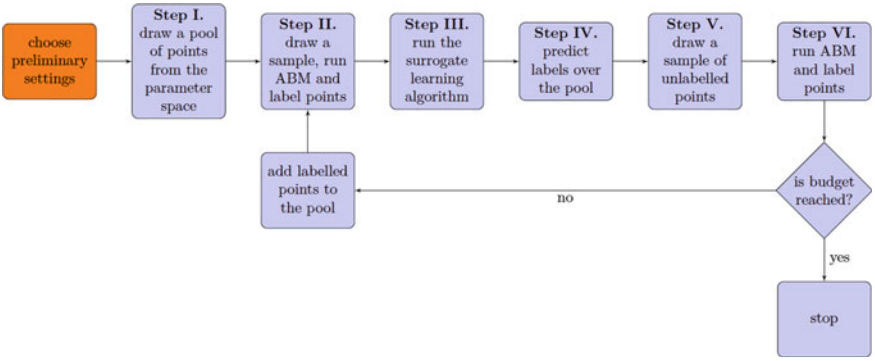

choose preliminary settings
Step I. draw a pool of points from the parameter space
Step II. draw a sample, run ABM and label points
Step III. run the surrogate learning algorithm
Step IV. predict labels over the pool
Step V. draw a sample of unlabelled points
Step VI. run ABM and label points
is budget reached?
no
add labelled points to the pool
yes
stop

Fig. 31.2 Schematic representation of the proposed procedure to learn a surrogate for an ABM. *Source* Lamperti et al. (2018c)

31.5.4 Parameter Space Exploration and Calibration via Machine-Learning Surrogates

Kriging constitutes a valuable meta-modeling technique to approximate the behavior of an ABM in a given region of the parameter space. However, one may need to extensively explore the parameter space to detect possible abrupt changes in the aggregate properties of the model or simply to have a general and precise overview of its behavior. Such broad explorations are usually either infeasible in terms of computational costs or tend to boil down to rough approximations obtained with small samples of learning points. This issue is addressed in Lamperti et al. (2018c), who explicitly tackle parameter space exploration and calibration of ABMs combining supervised machine-learning and intelligent sampling to build a surrogate meta-model, which is then used to classify parameter vectors according to the behavior they produce. The machine-learning surrogate dramatically reduces the computation time needed to perform large-scale explorations of the parameter space, while providing a powerful filter to gain insights into the complex functioning of agent-based models.

The original would be practically infeasible in terms of computational costs.

The learning process of a surrogate occurs over multiple rounds, as summarized in Fig. 31.2). The crucial part of the job is finding a precise approximation of the original model, which has to be learnt over samples of points selected to minimize the computational effort of the overall procedure. In particular, the surrogate training procedure involves three decisions:

1. choose a machine-learning algorithm to act as a surrogate for the original ABM;
2. select a sampling procedure to draw samples from the parameters space in order to train the surrogate;
3. select a score or criterion to evaluate the performance of the surrogate.

Extreme gradient boosted trees (XGBoost, see Chen and Guestrin 2016) are used as the predefined surrogate learning algorithm employed to form a random ensemble of classification and regression trees (CART, cf. Breiman et al. 1984). This choice allows the surrogate to learn nonlinear "knife-edge" properties, which typically characterize ABM parameter spaces. The sampling procedure builds a set of parameter vectors on which the agent-based model is actually evaluated in order to provide labeled data points for the training of the surrogate. Sets of parameter combinations are successively drawn according to a quasi-random Sobol sampling over the parameter space (Morokoff and Caflisch 1994). Finally, the quality of the surrogate approximation is measured through the true positive ratio (TPR), a standard classification accuracy indicator computed as the number of parameter vectors correctly predicted (by the surrogate) to satisfy the user-specified conditions over the total number of parameters in the "pool" truly satisfying them.

Lamperti et al. (2018c) provide two applications of the machine-learning surrogate approach employing a financial ABM and a model of endogenous growth. Results are encouraging. In the growth model (Fagiolo and Dosi 2003), parameter vectors delivering fat-tailed output growth-rate distributions are selected. The authors find that even for limited budget, the surrogate correctly classifies more than 80% parameter combinations, and computational costs are extraordinarily lower than those required by the original ABM.

31.6 Conclusions

Ten years after the influential article by Windrum et al. (2007), the issue of empirical validation of agent-based models (ABM) in economics and finance is still among the top items in the to-do list of researchers. This despite the fact that many advances have occurred, especially in the three key areas of: (i) calibration and estimation of model parameters; (ii) comparison of real-world and artificial data; and (iii) parameter space exploration.

This Chapter has attempted to critically survey such a recent literature focusing on developments in the above three areas.

Notwithstanding the huge effort made in advancing the frontier in ABM validation techniques, the process of developing a complete and coherent validation toolbox is still ongoing and some important issues are still to be better understood.[20]

First, the pros and cons of each different validation methodology are still not completely laid out in the literature. This is a pity, as a sort of if-then map would be extremely useful for practitioners aiming at picking the right tool in each specific situation. Projects developing such a map would be very welcome in the community, although it is of course clear that each tool is aimed at a specific task, and no validation technique is more general than the others. Relatedly, validation software packages should be developed to ease the adoption of the different existing techniques.

[20] The interested reader might want to look at Thiele et al. (2014) for a cookbook guiding model exploration and sensitivity and Grimm et al. (2005) for a pattern-oriented approach at model building and evaluation.

Second, and most importantly, more research efforts should be devoted toward advancing hypothesis testing in ABM. In particular, more robust statistical tests should be developed in order to better characterizing model stationarity and ergodicity, and to better understand how the failure of these properties might affect the problems of estimation, calibration, validation, and exploration.

To conclude, we believe that the development of better empirical-validation techniques is a never-ending process, which must naturally coevolve together with the developments of new models, new statistical techniques and with the increase in computational power (see also Chap. 18 by Robinson in this volume). In that, recent developments in machine-learning and the increase availability of big data could entail the next leap forward: machine learning offers indeed more flexible methodologies that allow one to minimize the number of assumptions when running an econometric model; big data, instead, allow one to perform more thorough comparisons of the model with the real-world situations, by extending validation also to the microlevel. All in all, these extensions would allow ACE models to progress from Level 2 to Level 3 in the Axtell and Epstein (1994) classification (see Sect. 31.4).

Furthermore, validation of ABMs will never tell whether a model is a correct description of the complex, unknown and non-understandable real-world data generating process. However, in a Popperian fashion, ABM validation techniques should eventually allow researchers to understand whether a model is a bad description of it.

Acknowledgements We gratefully acknowledge the support by the European Union Horizon 2020 research and innovation programme under grant agreement No. 649186 - ISIGrowth. Further, we express our gratitude to Francesca Chiaromonte, Giovanni Dosi, Mauro Napoletano, Marcelo Pereira, Amir Sani, and Maria Enrica Virgillito for helpful comments and discussions on the issues surveyed in this chapter.

References

Alfarano, S., Lux, T., & Wagner, F. (2005). Estimation of agent-based models: The case of an asymmetric herding model. *Computational Economics*, *26*(1), 19–49.

Alfarano, S., Lux, T., & Wagner, F. (2006). Estimation of a simple agent-based model of financial markets: An application to Australian stock and foreign exchange data. *Physica A: Statistical Mechanics and its Applications*, *370*(1), 38–42.

Anufriev, M., Bao, T., & Tuinstra, J. (2016). Microfoundations for switching behavior in heterogeneous agent models: An experiment. *Journal of Economic Behavior & Organization, 129*(C):74–99.

Anufriev, M., & Hommes, C. (2012). Evolutionary selection of individual expectations and aggregate outcomes in asset pricing experiments. *American Economic Journal: Microeconomics*, *4*(4), 35–64.

Assenza, T., Delli Gatti, D., & Grazzini, J. (2015). Emergent dynamics of a macroeconomic agent based model with capital and credit. *Journal of Economic Dynamics and Control, 50*(C):5–28.

Assenza, T., Heemeijer, P., Hommes, C., & Massaro, D. (2013). *Individual expectations and aggregate macro behavior*. Tinbergen Institute Discussion Papers 13-016/II, Tinbergen Institute.

Axelrod, R. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton University Press.

Axtell, R. L., & Epstein, J. M. (1994). Agent-based modeling: Understanding our creations. *The Bulletin of the Santa Fe Institute*, *9*(2), 28–32.

Barde, S. (2016a). Direct comparison of agent-based models of herding in financial markets. *Journal of Economic Dynamics and Control, 73*(C):329–353.

Barde, S. (2016b). A practical, accurate, information criterion for nth order markov processes. *Computational Economics*, 1–44.

Barde, S., & van der Hoog, S. (2017). *An empirical validation protocol for large-scale agent-based models*. Studies in Economics 1712, School of Economics, University of Kent.

Bargigli, L., Riccetti, L., Russo, A., & Gallegati, M. (2016). *Network calibration and metamodeling of a financial accelerator agent based model*. Technical report, Università Politecnica delle Marche.

Battiston, S., Farmer, J. D., Flache, A., Garlaschelli, D., Haldane, A. G., Heesterbeek, H., et al. (2016). Complexity theory and financial regulation. *Science*, *351*(6275), 818–819.

Bianchi, C., Cirillo, P., Gallegati, M., & Vagliasindi, P. (2007). Validating and calibrating agent-based models: A case study. *Computational Economics*, *30*, 245–264.

Bianchi, C., Cirillo, P., Gallegati, M., & Vagliasindi, P. (2008a). Validation in agent-based models: An investigation on the CATS model. *Journal of Economic Behavior & Organization*, *67*, 947–964.

Bianchi, C., Cirillo, P., Gallegati, M., & Vagliasindi, P. A. (2008b). Validation in agent-based models: An investigation on the CATS model. *Journal of Economic Behavior & Organization*, *67*(3–4), 947–964.

Boswijk, H. P., Hommes, C. H., & Manzan, S. (2007). Behavioral heterogeneity in stock prices. *Journal of Economic Dynamics and Control*, *31*(6), 1938–1970.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Brenner, T., & Werker, C. (2007). A taxonomy of inference in simulation models. *Computational Economics*, *30*(3), 227–244.

Brock, W. A. (1999). Scaling in economics: A reader's guide. *Industrial and Corporate Change*, *8*(3), 409–446.

Brock, W. A., & Hommes, C. H. (1997). A rational route to randomness. *Econometrica*, *65*(5), 1059–1095.

Brock, W. A., & Hommes, C. H. (1998). Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics and Control*, *22*(8–9), 1235–1274.

Burton, R. M., & Obel, B. (1995). The validity of computational models in organization science: From model realism to purpose of the model. *Computational & Mathematical Organization Theory*, *1*(1), 57–71.

Canova, F., & Sala, L. (2009). Back to square one: Identification issues in DSGE models. *Journal of Monetary Economics*, *56*(4), 431–449.

Chen, S.-H., Chang, C.-L., & Du, Y.-R. (2012). Agent-based economic models and econometrics. *The Knowledge Engineering Review*, *27*(2), 187–219.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

Chiarella, C., He, X.-Z., & Zwinkels, R. C. (2014). Heterogeneous expectations in asset pricing: Empirical evidence from the S&P500. *Journal of Economic Behavior & Organization, 105*(C):1–16.

Ciarli, T. (2012). Structural interactions and long run growth: An application of experimental design to agent-based models. *Revue de l'OFCE*, *124*, 295–345.

Dawid, H. & Delli Gatti, H. (2018). Chapter 2 - agent-based macroeconomics. In C. Hommes & B. LeBaron (Eds.), *Handbook of computational economics* (Vol. 4, pp. 63–156). Elsevier.

Dawid, H., Harting, P., van der Hoog, S., & Neugart, M. (2016). A heterogeneous agent macroeconomic model for policy evaluation: Improving transparency and reproducibility.

de Jong, E., Verschoor, W. F., & Zwinkels, R. C. (2010). Heterogeneity of agents and exchange rate dynamics: Evidence from the EMS. *Journal of International Money and Finance*, *29*(8), 1652–1669.

Del Negro, M., & Schorfheide, F. (2006). How good is what you've got? DSGE-VAR as a toolkit for evaluating DSGE models. *Economic Review*, (Q 2):21–37.

Dieci, R., & He, X.-Z. (2018). Chapter 5 - heterogeneous agent models in finance. In C. Hommes & B. LeBaron (Eds.), *Handbook of computational economics* (Vol. 4, pp. 257–328). Elsevier.

Dosi, G., Fagiolo, G., Napoletano, M., & Roventini, A. (2013). Income distribution, credit and fiscal policies in an agent-based keynesian model. *Journal of Economic Dynamics and Control*, *37*(8), 1598–1625.

Dosi, G., Fagiolo, G., Napoletano, M., Roventini, A., & Treibich, T. (2015). Fiscal and monetary policies in complex evolving economies. *Journal of Economic Dynamics and Control*, *52*, 166–189.

Dosi, G., Fagiolo, G., & Roventini, A. (2010). Schumpeter meeting keynes: A policy-friendly model of endogenous growth and business cycles. *Journal of Economic Dynamics and Control*, *34*(9), 1748–1767.

Dosi, G., Napoletano, M., Roventini, A., & Treibich, T. (2016a). Micro and macro policies in the Keynes+Schumpeter evolutionary models. *Journal of Evolutionary Economics*, forthcoming, 1–28.

Dosi, G., Pereira, M., Roventini, A., & Virgilito, M. E. (2017a). When more flexibility yields more fragility: The microfoundations of keynesian aggregate unemployment. *Journal of Economic Dynamics & Control*, *81*, 162–186.

Dosi, G., Pereira, M. C., Roventini, A., & Virgillito, M. E. (2016b). *The effects of labour market reforms upon unemployment and income inequalities: An agent based model* (LEM Working Papers 2016/27). Scuola Superiore Sant'Anna.

Dosi, G., Pereira, M. C., Roventini, A., & Virgillito, M. E. (2017b). *Causes and consequences of hysteresis: Aggregate demand, productivity and employment* (LEM Working Papers 2017/07). Scuola Superiore Sant'Anna.

Dosi, G., Pereira, M. C., & Virgillito, M. E. (2017c). On the robustness of the fat-tailed distribution of firm growth rates: A global sensitivity analysis. *Journal of Economic Interaction and Coordination*, 1–21.

Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up*. Brookings Institution Press.

Fabretti, A. (2013). On the problem of calibrating an agent based model for financial markets. *Journal of Economic Interaction and Coordination*, *8*(2), 277–293.

Fagiolo, G., & Dosi, G. (2003). Exploitation, exploration and innovation in a model of endogenous growth with locally interacting agents. *Structural Change and Economic Dynamics*, *14*, 237–273.

Fagiolo, G., & Roventini, A. (2012). Macroeconomic policy in DSGE and agent-based models. *Revue de l'OFCE, 0*(5), 67–116.

Fagiolo, G., & Roventini, A. (2017). Macroeconomic policy in DSGE and agent-based models redux: New developments and challenges ahead. *Journal of Artificial Societies and Social Simulation, 20*(1).

Farmer, D. J., & Foley, D. (2009). The economy needs agent-based modelling. *Nature*, *460*, 685–686.

Fernández-Villaverde, J., Ramírez, J. F. R., & Schorfheide, F. (2016). *Solution and Estimation Methods for DSGE Models* (NBER Working Papers 21862). National Bureau of Economic Research, Inc.

Fernández-Villaverde, J., & Rubio-Ramírez, J. F. (2007). Estimating macroeconomic models: A likelihood approach. *Review of Economic Studies*, *74*(4), 1059–1087.

Franke, R. (2009). Applying the method of simulated moments to estimate a small agent-based asset pricing model. *Journal of Empirical Finance*, *16*(5), 804–815.

Franke, R., & Westerhoff, F. (2012). Structural stochastic volatility in asset pricing dynamics: Estimation and model contest. *Journal of Economic Dynamics and Control*, *36*(8), 1193–1211.

Gaffeo, E., Delli Gatti, D., Desiderio, S., & Gallegati, M. (2008). Adaptive microfoundations for emergent macroeconomics. *Eastern Economic Journal*, *34*(4), 441–463.

Goldbaum, D., & Mizrach, B. (2008). Estimating the intensity of choice in a dynamic mutual fund allocation decision. *Journal of Economic Dynamics and Control*, *32*(12), 3866–3876.

Gourieroux, C., Monfort, A., & Renault, E. (1993). Indirect Inference. *Journal of Applied Econometrics, 8*(S):85–118.

Grazzini, J., & Richiardi, M. (2015). Estimation of ergodic agent-based models by simulated minimum distance. *Journal of Economic Dynamics and Control, 51*(C):148–165.

Grazzini, J., Richiardi, M. G., & Tsionas, M. (2017). Bayesian estimation of agent-based models. *Journal of Economic Dynamics and Control, 77*(C), 26–47.

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., et al. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological modelling*, *198*(1–2), 115–126.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., et al. (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, *310*(5750), 987–991.

Guerini, M. (2013). *Is the friedman rule stabilizing? Some unpleasant results in a heterogeneous expectations framework*. Technical report, Department of Economics and Finance Working Papers, Unicatt, Milan.

Guerini, M., & Moneta, A. (2017). A method for agent-based models validation. *Journal of Economic Dynamics and Control*.

Guerini, M., Napoletano, M., & Roventini, A. (2017). No man is an island: The impact of heterogeneity and local interactions on macroeconomic dynamics. *Economic Modelling*.

Hansen, L. P., & Heckman, J. J. (1996). The empirical foundations of calibration. *The Journal of Economic Perspectives*, *10*(1), 87–104.

Hassan, S., Pavon, J., & Gilbert, N. (2008). Injecting data into simulation: Can agent-based modelling learn from microsimulation. In *World Congress of Social Simulation*.

Heine, B.-O., Meyer, M., & Strangfeld, O. (2005). Stylised facts and the contribution of simulation to the economic analysis of budgeting. *Journal of Artificial Societies and Social Simulation, 8*(4).

Hommes, C. (2011). The heterogeneous expectations hypothesis: Some evidence from the lab. *Journal of Economic Dynamics and Control*, *35*(1), 1–24.

Hommes, C. (2013). *Behavioral rationality and heterogeneous expectations in complex economic systems*. Number 9781107564978 in Cambridge Books. Cambridge University Press.

Hyvarinen, A., Zhang, K., Shimizu, S., & Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, *11*, 1709–1731.

Johansen, S., & Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration. With application to the demand for money. *Oxford Bullettin of Economics and Statistics*, *52*, 169–210.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kirman, A. (1991). Epidemics of opinion and speculative bubbles in financial markets. In M. Taylor (Ed.), *Money and financial markets* (pp. 354–368). Blackwell.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, *52*(6), 119–139.

Kukacka, J., & Barunik, J. (2017). Estimation of financial agent-based models with simulated maximum likelihood. *Journal of Economic Dynamics and Control, 85*(C):21–45.

Lamperti, F. (2018a). Empirical validation of simulated models through the GSL-div: An illustrative application. *Journal of Economic Interaction and Coordination*, *13*(1), 143–171.

Lamperti, F. (2018b). An information theoretic criterion for empirical validation of simulation models. *Econometrics and Statistics*, *5*, 83–106.

Lamperti, F., Dosi, G., Napoletano, M., Roventini, A., & Sapio, A. (2018a). Faraway, so close: Coupled climate and economic dynamics in an agent-based integrated assessment model. *Ecological Economics*, *150*, 315–339.

Lamperti, F., Dosi, G., Napoletano, M., Roventini, A., Sapio, A., et al. (2018b). *And then he wasn't a she: Climate change and green transitions in an agent-based integrated assessment model*. Technical report, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy.

Lamperti, F., Roventini, A., & Sani, A. (2018c). Agent-based model calibration using machine learning surrogates. *Journal of Economic Dynamics and Control*, *90*, 366–389.

Lane, D. A. (1993). Artificial worlds and economics, part II. *Journal of Evolutionary Economics*, *3*(3), 177–197.

Leal, S. J., Napoletano, M., Roventini, A., & Fagiolo, G. (2016). Rock around the clock: An agent-based model of low- and high-frequency trading. *Journal of Evolutionary Economics*, *26*(1), 49–76.

LeBaron, B., & Tesfatsion, L. (2008). Modeling macroeconomies as open-ended dynamic systems of interacting agents. *American Economic Review*, *98*(2), 246–250.

Lee, J.-S., Filatova, T., Ligmann-Zielinska, A., Hassani-Mahmooei, B., Stonedahl, F., Lorscheid, I., et al. (2015). The complexities of agent-based modeling output analysis. *Journal of Artificial Societies and Social Simulation*, *18*(4), 4.

Leombruni, R., Richiardi, M., Saam, N. J., & Sonnessa, M. (2006). A common protocol for agent-based social simulation. *Journal of Artificial Societies and Social Simulation*, *9*(1), 15.

Lorscheid, I., Heine, B.-O., & Meyer, M. (2012). Opening the fiblack boxfiof simulations: Increased transparency and effective communication through the systematic design of experiments. *Computational and Mathematical Organization Theory*, *18*(1), 22–62.

Malerba, F., Nelson, R., Orsenigo, L., & Winter, S. (1999). 'History-friendly' models of industry evolution: The computer industry. *Industrial and Corporate Change*, *8*(1), 3.

Manson, S. (Ed.). (2002). *Validation and verification of multi-agent systems, in complexity and ecosystem management*. Cheltenham: Edward Elgar.

Marks, R. (2007). Validating simulation models: A general framework and four applied examples. *Computational Economics*, *30*(3), 265–290.

Marks, R. E. (2013). Validation and model selection: Three similarity measures compared. *Complexity Economics*, *2*(1), 41–61.

Marks, R. E. (2018). Pattern-based metrics for validating simulation model output. In C. Beisbart & N. J. Saam (Eds.), *Computer simulation validation. Fundamental concepts, methodological frameworks, philosophical perspectives*. Springer.

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239–245.

Metropolis, N., & Ulam, S. (1949). The monte carlo method. *Journal of American Statistical Association*, *44*, 335–341.

Morokoff, W. J., & Caflisch, R. E. (1994). Quasi-random sequences and their discrepancies. *SIAM Journal on Scientific Computing*, *15*(6), 1251–1279.

Paccagnini, A. (2010). *DSGE model validation in a bayesian framework: An assessment*. MPRA Paper 24509, University Library of Munich, Germany.

Pellizzari, P., & Dal Forno, A. (2007). A comparison of different trading protocols in an agent-based market. *Journal of Economic Interaction and Coordination*, *2*(1), 27–43.

Platt, D., & Gebbie, T. (2016). *Can agent-based models probe market microstructure*? Papers 1611.08510, arXiv.org.

Popoyan, L., Napoletano, M., & Roventini, A. (2017). Taming macroeconomic instability: Monetary and macro-prudential policy interactions in an agent-based model. *Journal of Economic Behavior & Organization, 134*(C):117–140.

Recchioni, M. C., Tedeschi, G., & Gallegati, M. (2015). A calibration procedure for analyzing stock price dynamics in an agent-based framework. *Journal of Economic Dynamics and Control*, *60*, 1–25.

Rosen, R. (1985). *Anticipatory systems: Philosophical, mathematical, and methodological foundations*. Oxford: Pergamon.

Salle, I., & Yıldızoğlu, M. (2014). Efficient sampling and meta-modeling for computational economic models. *Computational Economics*, *44*(4), 507–536.

Schelling, T. C. (1969). Models of segregation. *The American Economic Review*, *59*(2), 488–493.

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, *1*(2), 143–186.

Secchi, D., & Seri, R. (2017). Controlling for false negatives in agent-based models: A review of power analysis in organizational research. *Computational and Mathematical Organization Theory*, *23*(1), 94–121.

Shimizu, S., Hoyer, P. O., Hyvarinen, A., & Kerminen, A. J. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*, 2003–2030.

Simon, H. A. (1991). Bounded rationality and organizational learning. *Organization Science*, *2*(1), 125–134.

Spirtes, P., Glymur, C., & Scheines, R. (2000). *Causation, prediction, and search*. MIT Press.

Tesfatsion, L. (2006). Chapter 16 agent-based computational economics: A constructive approach to economic theory. In *Handbook of computational economics*, 2 (pp. 831–880).

Thiele, J. C., Kurth, W., & Grimm, V. (2014). Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using NetLogo and R. *Journal of Artificial Societies and Social Simulation*, *17*(3), 11.

Turrell, A. (2016). *Agent-based models: Understanding the economy from the bottom up*. Quarterly bulletin Q4, Bank of England.

Van Beers, W. C. & Kleijnen, J. P. (2004). Kriging interpolation in simulation: A survey. In *Simulation Conference, 2004. Proceedings of the 2004 Winter* (vol. 1). IEEE.

Werker, C., & Brenner, T. (2004). *Empirical calibration of simulation models* 0410. Papers on economics and evolution, Max-Planck-Institut für Ökonomik.

Westerhoff, F. H., & Dieci, R. (2006). The effectiveness of keynes-tobin transaction taxes when heterogeneous agents can trade in different markets: A behavioral finance approach. *Journal of Economic Dynamics and Control*, *30*(2), 293–322.

Windrum, P., Fagiolo, G., & Moneta, A. (2007). Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, *10*(2), 8.

Winker, P., & Gilli, M. (2001). Validation of agent-based models of financial markets. *IFAC Proceedings Volumes*, *34*(20), 401–406.

Winker, P., & Gilli, M. (2004). Applications of optimization heuristics to estimation and modelling problems. *Computational Statistics & Data Analysis*, *47*(2), 211–223.

Part VIII
Challenges in Simulation Model Validation

Chapter 32
Validation and Equifinality

Keith Beven

Abstract In this chapter, the concept of equifinality of model representations is discussed, from a background of model applications in the environmental sciences. Equifinality in this context is used to indicate that there may be many different model structures, parameter sets and auxiliary conditions that might appear to give equivalent output predictions or acceptable fits to any observation data available for use in model calibration. This does not imply that the resulting ensemble of models will give similar predictions when used to predict the future under some changed conditions. As new information becomes available to allow model validation, this can be used to constrain the ensemble of models within a Bayesian updating framework, although epistemic sources of uncertainty can make it difficult to define appropriate likelihood measures. It seems likely that the equifinality concept will persist into the future in the form of ensembles of (stochastic) model runs being used to estimate prediction uncertainties. However, more research is needed into the limitations of model structures, information content of data sets subject to epistemic uncertainties and means of evaluating and validating models in the inexact sciences.

Keywords Inexact sciences · Model ensembles · Epistemic uncertainties · Environmental models · Equifinality thesis · Audit trail

32.1 Introduction

There are a number of stages in the modelling process. Beven (2012a) distinguishes a perceptual model stage that may be purely qualitative in nature; a conceptual model stage, in which our perceptions about a system are approximated by equations; and a procedural model stage in which those equations are implemented as code, often with a further level of approximation. In the environmental sciences, which can be considered as representative of the inexact sciences, nearly all predictive models of

K. Beven (✉)
Lancaster Environment Centre, Lancaster University, Lancaster, UK
e-mail: k.beven@lancaster.ac.uk

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_32

this type involve parameter values that have to be estimated in some way before a model can be run for an application to a particular system. This will be the case even for models derived by deductive reasoning from process theories. This estimation may be purely on the basis of past experience and observations, though it is often difficult to relate parameter values determined by measurement to the effective values required to make a model successfully simulate the system response. Thus, it is a common practice to calibrate or tune the effective parameter values against some historical data set, where such data are available. This chapter discusses the possibility that there may be many different model representations that might be considered to give equivalent output predictions or, in particular in the context of model calibration, simulations that are considered to be acceptable in some sense, given the epistemic uncertainties involved in the modelling process. This is the equifinality thesis (Beven 2006).

Following model calibration, it is also generally considered as good practice to carry out an additional stage of model validation. This can involve the reconsideration of the theoretical basis of a model based on results from the calibration stage, but more generally involves a further evaluation against some "independent" data set for the system being simulated. The minimal evaluation, in this respect, is for another period of observations from the same site: the split-record validation test (Klemes 1986; Refsgaard and Knudsen 1996). A stronger and more independent test would be to evaluate model outputs against observed variables that have not been used in calibrating the model. It will be shown that in the context of the equifinality thesis, this can be a way of refining the ensemble of models considered as acceptable, as a form of Bayesian updating. It is suggested that this provides a way of testing models as hypotheses in the inexact sciences, while making allowance for the sources of uncertainty in the modelling process.

In the following, Sect. 32.2 introduces the equifinality concept. Empirical results from many computer experiments with a variety of hydrological models showed that equifinality is generic to this type of modelling exercise. It is the parameter *set* that fully reflects the interactions between parameter values in producing a sequence of model predictions that will be considered as acceptable or not (Sect. 32.3). Using environmental models as an example, Sect. 32.4 explains the particularities of the inexact sciences with respect to errors and uncertainties in the input and observational data, and the consequent limitations in using of traditional statistical hypotheses testing. Section 32.5 introduces the Generalised Likelihood Uncertainty Estimation (GLUE) method which preserves the effects of parameter interactions within the model structure. It allows for equifinality in both model structures and parameter sets. Section 32.6 argues that in GLUE, there is a framework that allows the choice of any sensible likelihood measure, including the recognition of periods of disinformation in the data used for model calibration v and evaluation. Section 32.7 turns to simulation validation in the inexact sciences. Here, the modeller is often caught between the desire to use as wide a range of conditions as possible in model calibration, and the need to check model predictions on some validation data set. The suggested ensemble framework for model evaluation provides a methodology for model validation (and invalidation), despite, the difficulties of properly evaluating the information content of data subject to epistemic uncertainties, and consequently defining an appropriate measure of acceptability in deciding whether a model is really fit for purpose.

32.2 The Origins of Equifinality Concepts

The word equifinality was originally used in the context of the General Systems Theory of Ludwig von Bertalanffy (1951, 1968) to describe the concept that open systems might reach a similar steady state starting from different initial conditions. The examples he used were mostly concerned with biological systems, but the concepts of General System Theory were introduced into meteorology by Thompson (1961) and geomorphology by Culling (1957) and Chorley (1962). In geomorphology, it was used to convey the idea that similar landforms might derive from different initial states and histories. Culling (1987) later modified his ideas to a more general description of stability of form derived from the theory of nonlinear dynamic systems.

Beven (1975, 1993, 2006, 2009) adopted equifinality for use with environmental models, where empirical results based on Monte Carlo simulation show that different model structures and different model parameter sets might lead to rather similar model outputs, and in particular, rather similar performance when compared with any available observational data. For such models, it is generally expected that there will be a gradation of acceptability from the "best" models that can be found, to those that are clearly not acceptable as simulators of the system of interest. In this context, the equifinality concept is intrinsically linked to model calibration and validation as parts of the modelling process, including the consideration of uncertainty in the available data. The equifinality thesis suggests that there will be no single model representation of an environmental system, but rather an evolving ensemble of models that are considered acceptable in the sense of being useful in prediction as new information becomes available over time (Beven 2006, 2012a, b).

This should not be a surprise given the epistemic uncertainties associated with model representations in the inexact sciences, but it still allows for a form of scientific methodology to be adopted by treating the ensemble of plausible models as multiple working hypotheses about how the system of interest is functioning. The interesting question then is how to test those hypotheses in different ways. Whether any of the models as hypotheses can be considered as fit for purpose is discussed earlier by Beven and Lane (Chap. 6 in this volume).

32.3 Equifinality as an Empirical Result

I first started running Monte Carlo experiments with hydrological models in about 1980 at the University of Virginia. At that time, I had access to a CDC6600 mainframe computer, which allowed many more runs to be made than previous computers I had been able to use. Modelling strategies in the hydrological community at that time were focussed on trying to find the optimum model of a catchment area or water resource system, including research on finding better methods of optimisation (e.g. Sorooshian and Gupta, 1995). My previous experience, however, suggested that any optimum model would be poorly defined, conditional on the period of data used in model calibration, the errors in that data, and the performance measure(s) or objective function(s) chosen by the modeller. Sometimes, the choice of measure seemed rather arbitrary, such as the use of an index based on the sum of squared model residuals (such as the Nash and Sutcliffe (1970) efficiency measure, a form of coefficient of determination based on the sum of squared residuals), even though the time series of residuals do not often seem random or independent in nature. In one study, we showed that an optimisation algorithm resulted in the model being used in a way that conflicted with the way it was supposed to represent the surface and subsurface runoff processes (Beven and Kirkby 1979). Discussion of appropriate measures to use for evaluating hydrological models continues to the present day (see for example, Schaefli and Gupta 2007; Smith et al. 2008; Gupta et al. 2009; Reusser et al. 2009; Gupta and Kling 2011; Mizukami et al. 2018).

So a brute force Monte Carlo method for model calibration, in which many different model parameter sets were chosen randomly from prior ranges or distributions, provided an interesting alternative strategy. Once the runs were made, there was the possibility of evaluating the outputs from those runs in a variety of different ways (although at that time storing and retrieving the results was expensive and slow, and generally involved writing to magnetic tapes). The impact of the initial results was dramatic, showing that for the types of performance measures being used at that time, there were very many model parameter sets, spread through the model parameter space that gave rather similar results when evaluated against past observations, indicating a form of equifinality of model representations. This was also the case for model output variables; many model parameter sets would give similar outputs.

An example of such results for a hydrological model is given in Fig. 32.1, which shows the results of many Monte Carlo simulations for a single model structure with different randomly chosen parameter sets. These scatter plots or dotty plots represent projections onto single parameter axes of point samples from a multidimensional surface defined by a single performance measure in the model space. Each point represents the results from a single model run. As point projections, they do not properly illustrate the nature of that surface which might be rather complex, involving multiple peaks and troughs or ridges and valleys that will reflect the complex interactions between parameters in the model outputs. In addition, similar values of the performance measure might be the result of the model simulating the functioning of the system in rather different ways (for example, with different dominant processes, or simply different residual compromises in achieving a similar value of performance).

However, three important points may be drawn from such plots. The first is that there often seems to be some upper limit of performance. Figure 32.1 shows this for a single model structure, but this can apply also to multiple model structures with more or less processes represented. It is known from fitting functions to data using statistical regression, that the more degrees of freedom, in general, the better the fit that will be obtained, but with the danger that the data might be over-fitted leading to problems in prediction and validation. It seems that it is not always the case that more complex model structures will give better fits to the evaluation data, even if they have more parameters that can be varied. This appears to be related to the limitations of the data sets being used for driving and evaluating the model, and the limitations of the model structure in representing the system of interest. At that time, little consideration was given to uncertainties and errors in the input and observational data, but allowing for such potential uncertainties will only increase the potential for equifinality.

Second, many parameters show that the highest values of the performance measure are spread across the range or distribution from which the values are sampled. This has been seen in such experiments with a wide range of environmental models for which Monte Carlo experiments are feasible. Indeed, it is often the case that the best

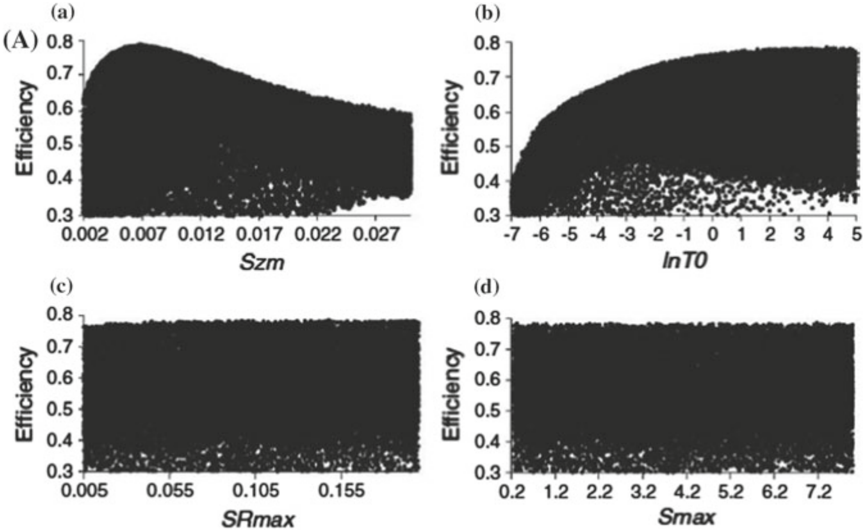


Fig. 32.1 Examples of equifinality in dotty plots for different types of model applications. Y-axes are proportional to a likelihood measure; X-axes are parameter values across the range of values sampled. Each dot represents one run of the model with randomly chosen parameter values. In all of these cases, the samples were taken from uniform distributions across the indicated ranges, and parameters were sampled independently. The plots, therefore, represent projections of points on the likelihood surface onto each parameter axis, **A** Results from a simple four-parameter rainfall-runoff model, evaluated against observed river discharges using the sum of squares based efficiency measure, for the Hafren, catchment, mid-Wales (after Page et al. 2007), **B** Results from a flood inundation model, evaluated using a fuzzy performance measure against water levels after a major flood in the Alzette River, Luxembourg (from Pappenberger et al. 2007), **C** Results from an advection–dispersion solute transport model evaluated against drainage concentrations for a conservative tracer and a non-conservative pesticide in the outflow from a large undisturbed soil column. v_e is the mean pore water velocity, D is the dispersion coefficient, R is the retardation coefficient and μ^E is the degradation coefficient for the pesticide. The error bars for R and μ^E indicate the results from a nonlinear least squares optimization (from Zhang et al. 2006)

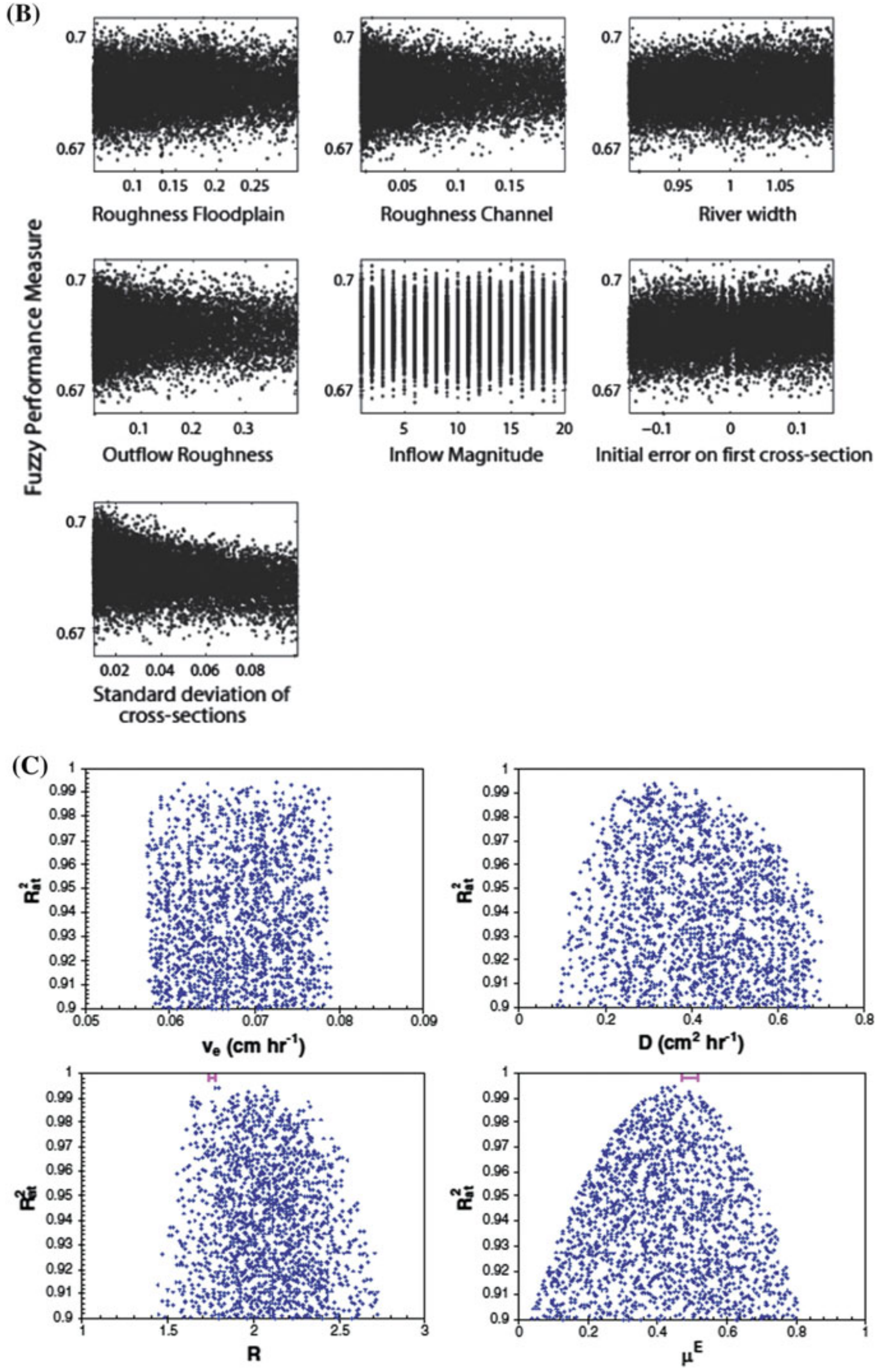


Fig. 32.1 (continued)

fits extend right to the edges of the specified range (as shown in Fig. 32.1), even if that range represents the modeller's choice of physically plausible values. This might also be related to the limitations of the model structure and data available.

The third point is that for any particular parameter value that shows good fits to the evaluation data, there will also generally be a full range of poorer fits conditional on the values of all the other parameters. This means that it is not (usually) individual parameter values that determine a good fit or bad fit, it is the parameter *set* that is important. It is the parameter *set* that fully reflects the interactions between parameter values in producing a sequence of model predictions. These interactions will reflect the nonlinear dynamics of the model representation and will not necessarily map to simple covariance structures, since the nature of the interaction might vary in different parts of the model space. Because of these complex interactions, there have been cases where the parameter set made up of the modal values of each posterior marginal parameter distribution over the ensemble of acceptable models has not itself produced acceptable simulations.

Note that, these are empirical results from many computer experiments with a variety of models. They show that equifinality is generic to this type of modelling exercise and focus attention on parameter *sets* in model performance. Changing the performance measure, or the period of calibration data, can modify the shape of the surface and result in different rankings of the parameter sets. There may also be Pareto trade-offs between multiple performance measures for a given parameter set (e.g. Yapo et al. 1998; Madsen 2003; Vrugt et al. 2003; Pokhrel et al. 2012). Note also that these issues apply to both deterministic and stochastic models.

32.4 Equifinality in Model Calibration in the Inexact Sciences

The empirical results from environmental models in the last section are examples of trying to apply models in what can be called the inexact sciences (e.g. Helmer and Rescher, 1959). The inexact sciences include environmental sciences that is subject to limited knowledge of their boundary conditions and processes. In the inexact sciences, we do not expect to achieve very good fits to the observations, since there will be errors in model structures, uncertainty in input and boundary condition data, and uncertainty in the observations used in model evaluation (Beven 2002, 2009). In addition, we do not expect these errors and uncertainties to have simple statistical structures. They will not be aleatory, but predominantly epistemic in nature. Model residuals will also be affected by the way in which input errors and uncertainties are processed through the nonlinear model dynamics. Thus, even if input uncertainties could be assumed to have a simple statistical structure, then the dependent output uncertainties from the model, which will affect the model residuals when compared with observational data, should be expected to have a nonstationary bias, variance and correlation structure. If the input uncertainties themselves are epistemic and complex, then this problem will be greater (Beven 2012b, 2016).

These issues make it difficult to assess the real information content of the data, particularly of data involving time series or spatial patterns rather than distributions of values alone. In some more extreme cases of epistemic uncertainty, the available input and evaluation data may not be physically consistent and so may not be informative in determining whether a model gives an acceptable simulation or not in calibration (e.g. Beven and Westerberg 2011; Beven et al. 2011a; Beven and Smith 2015).

In the inexact sciences, therefore, while we might want to treat models as potential working hypotheses to describe system functioning, testing those hypotheses in the face of epistemic uncertainties is proving to be challenging. Statistical methods of hypothesis testing will not generally be useful. They are not designed to deal with equifinality and complex interactions between parameters and nonstationarity in residual characteristics (unless they can be represented in simple functional forms, e.g. Renard et al. 2010; Schoups and Vrugt 2010). We might, however, be able to borrow some concepts from statistical hypothesis testing. In particular, we would wish to avoid making false positive and false negative errors (see also Chap. 6 by Beven and Lane in this volume). We might also wish to treat hypothesis testing and model calibration as a conditioning problem, starting with a prior distribution of plausible parameter sets and updating that ensemble as more observational data become available. A number of methods are available for doing so (e.g. Beven 2009).

32.5 Equifinality as Behavioural Model Ensembles

One of the earliest applications of Monte Carlo simulation to an environmental model was that of George Hornberger, Bob Spear and Peter Young, modelling algal blooms in Perth Inlet in Western Australia (e.g. Hornberger and Spear 1981; Spear and Hornberger 1980). Their model was relatively simple, linking algal productivity to nutrient status in the Inlet. One of the features of their analysis was the division of their ensemble of model runs into what they called “behavioural” and “non-behavioural” parameter sets. A model run was considered behavioural if it predicted an algal bloom in the Inlet; it was non-behavioural if it did not. The focus was on the parameter sets, rather than the individual parameters, since the effects of parameter interactions could be seen in whether there was an algal bloom or not.

This division into the two ensembles also provided a method for global sensitivity analysis. By looking at the marginal distributions for individual parameters in the behavioural and non-behavioural sets, then those parameters that showed a strong separation of the cumulative distributions could be inferred as the most sensitive, while those that showed little or no separation could be considered relative insensitive. The method has been widely applied, in part because it makes no strong assumptions about the nature of model residuals or parameter interactions, including to cases where there is no clear behavioural or non-behavioural indicator (like an algal bloom) but only relative measures of performance in fitting observations.

Clearly, this method incorporates the equifinality concept in a rather natural way (see van Straten and Keesman 1991; and Rose et al. 1991, for other early environmental modelling examples of this type of approach). If there are many different models that give more or less equally good results then they can be included in the behavioural ensemble. The behavioural set can then be used to make predictions about the system. If there are models that do not give acceptable results then they can be included in the non-behavioural set. This is, therefore, a form of testing models as hypotheses, if a suitable measure can be defined for deciding whether a model is behavioural or not. As more observations are made available, the parameter sets in each ensemble can be updated.

A simple extension of this method was proposed by Beven and Binley (1992) in the Generalised Likelihood Uncertainty Estimation (GLUE) methodology (see also Beven and Binley 2014). This allowed for each member of the behavioural ensemble be to be associated with a likelihood measure based on past calibration performance. Predictions made with the behavioural ensemble can then be likelihood weighted so that prediction quantiles can be estimated for any predicted variable. The plots of Fig. 32.1a, b and c are taken from applications of the GLUE methodology, where the measures of goodness-of-fit shown on the y-axes are transformed into likelihood weights for each of the models considered behavioural. This method preserves the effects of parameter interactions within the model structure in producing a behavioural simulation. It can also be extended to parameter sets within competing model structures, as long as the model outputs can be compared in the same way so that the likelihood values from different model structures are commensurate. It can therefore allow for equifinality in both model structures and parameter sets within those model structures, even if some parameters might have quite different meanings, or different effective values within different structures.

The GLUE methodology requires a number of decisions to be made.

1. Which model structures are to be evaluated?
2. What are the plausible parameter ranges or distributions to be sampled?
3. What sampling method will be used?
4. What are the criteria for differentiating between behavioural and non-behavioural parameter sets?
5. What likelihood measure or measures will be used for weighting the outputs of the behavioural parameter sets?
6. How will the likelihoods be updated as new observational data become available?

Similar decisions are required within a probabilistic identification methodology based on Bayes' equation. In a formal Bayesian framework, it is necessary to decide on prior distributions for parameters and their covariance; a method for sampling the model space (normally a form of Monte Carlo Markov Chain sampling); and a likelihood function based on the statistical characteristics of the model residuals. The updating step uses Bayes' equation to combine the prior probabilities and likelihoods to produce a posterior joint distribution for the parameters. Effectively, the Bayesian approach is a special case of the GLUE methodology where specific probabilistic assumptions can be made about the link between model residual structure and likelihood value, and where likelihoods are updated using Bayes' equation (e.g. Romanowicz et al. 1994, 1996). If the assumptions can be shown to be valid, it has the advantage that the model outputs and joint parameter distribution can be interpreted in a formal probabilistic way. It also avoids any decision about differentiating between behavioural and non-behavioural models. No model will be rejected; poor models will be given a very low likelihood but not zero (but will then have only a small or negligible contribution to the cumulative likelihood). Model structures can be compared within this framework, as long as the residual error model structures can be assumed to be the same so that the likelihoods are commensurate. There are then accepted ways of choosing one model structure over another using, for example ratios of integral posterior likelihoods for each model structure.

There is a variant of the Bayes' methodology that is used when it is difficult to define a formal likelihood function because of the complexity of the model residuals or other epistemic issues. This is Approximate Bayesian Computation (ABC) where models are evaluated in terms of some tolerance threshold for closeness to the evaluation data. This is analogous therefore to decisions about behavioural or non-behavioural models, except in that within ABC the tolerance level can be allowed to vary as part of the search algorithm, becoming smaller as the area of higher performance in the model space becomes better defined. The aim is to find an ensemble of models that lie in a high likelihood part of the model space. Since there is no formal likelihood definition, all models found within the tolerance limits are given equal weight. It has been shown for some simple cases that this approach can provide a good approximation to the formal Bayes' methodology.

GLUE is more general than either Bayes' or ABC approaches in that different ways of defining likelihoods, rejecting non-behavioural models, and combining likelihood weights can be used. It can be used, for example using fuzzy measures and fuzzy operators (see Beven 2009, p. 134, for example in different application areas), a framework that lends itself well to a behavioural/non-behavioural model differentiation for cases where the support for the fuzzy measures can be defined in terms of some limits of acceptability around the evaluation data. This can be for single observations, or for summary statistics derived from groups of observations. Defining such limits should allow for the uncertainties in both the input data and evaluation observations, and should ideally be done prior to running the model (Beven 2006, 2016). Behavioural models will then be those that satisfy the required limits of acceptability. Such an approach is very general, and allows for testing models as hypotheses. Those that do not provide simulated values within the limits of acceptability will be rejected (see also Chap. 6 by Beven and Lane in this volume).

GLUE can also include alternative representations of model residuals without the strong assumptions of a statistical likelihood formulation (e.g. the non-parametric representation of Beven and Smith 2015) but the treatment of residuals is often left implicit under the expectation that where a model under- or over-predicts in calibration it will under- or over-predict in a similar way under similar conditions in prediction. Where the behavioural model outputs have sufficient range to bracket the observations, then applying this approach to the full ensemble of behavioural models can provide a useful estimate of predictive uncertainty (but without the explicit probabilistic interpretation of formal Bayes' likelihoods). Where the ensemble of model outputs does not bracket the observations, it is in any case, a useful indication of a modelling problem, either as a result of model structural deficiencies or, as noted earlier, of inconsistencies with the data that might not be informative in model calibration. Such cases are perhaps more common than we like to think (e.g. Beven and Westerberg 2011; Beven et al. 2011b; Beven and Smith 2015).

32.6 Defining a Model Likelihood

The equifinality of model performance shown in Fig. 32.1 necessarily depends on the definition of a likelihood measure. The GLUE approach has been heavily criticised in the past (e.g. Montanari 2005; Mantovan and Todini 2006; Stedinger et al. 2008) because this is often left to the subjective judgement of the modeller rather than making use of a statistical definition of likelihood (though clearly, the latter is a choice that *could* be made within GLUE where deemed appropriate, see Beven et al. 2008; Beven 2009). The effect of using a statistical likelihood is to effectively stretch the likelihood surface as a result of the multiplicative combination of the probability estimates from individual model residuals. In general the greater the number of observations used in the evaluation, the greater the stretching will be. The more the observations can be considered independent, the greater the stretching will be. This is, in fact, one solution to the equifinality issue, since the greater the stretching of the likelihood surface, the smaller the region of apparent equifinality will be. Beven (2016) shows, again for a hydrological model evaluated against time series data, how models of very similar error variance can be assigned likelihoods of tens of orders of magnitude different in this way, even when allowing for correlation in the residuals. The stretching can be so extreme that it is normal practice to do the calculations, and report the results, as log-likelihoods to avoid rounding errors in the computations. The extreme case of this would be to only retain the maximum likelihood model as behavioural, treating the modelling uncertainties only in terms of the residuals for that model, with all other models given likelihoods of zero.

The question then is whether that stretching actually reflects a plausible belief in the relative merits of models with similar error variance. It follows mathematically from the assumptions about the model residuals (e.g. typically that they have a stationary Gaussian distribution of constant bias and variance and a simple correlation structure and that the contributions from individual residuals combine multiplicatively in accordance with Bayes' equation). These assumptions are based on treating the uncertainties in the modelling process as if they have simple aleatory characteristics. The assumptions can be checked against the characteristics of the actual model residuals. The claim to objectivity of the approach lies in this possibility of validation of the assumptions (though where this is done, if at all, it is usually only done for the model with the highest likelihood value).

But, for me at least, it is hard to accept that the resulting stretching of the likelihoods should represent my belief in the expected performance of different models, when if I plot out the time series of residuals for two models of similar error variance I find it difficult to really say whether one is better than another, even if they might have calculated statistical likelihoods that are orders of magnitude different. For me, this seems to be a problem in applying this form of statistical theory to nonlinear dynamic models in the inexact sciences when the errors and uncertainties may come from epistemic rather than aleatory sources. It seems counter to common sense in that it may over-constrain the prediction of future events when the maximum likelihood model (or more generally, the ensemble of high likelihood models) is not robust to the use of different periods of data. Including a wider range of predictive models (accepting equifinality) would then seem to be advantageous.

This question of plausibility may be interpreted in two ways. First, we might wish to take advantage of statistical theory as an approximation to a more realistic analysis. In doing so, we recognise the limitations and approximate nature of the necessary statistical assumptions in representing epistemic uncertainties, but because those uncertainties are epistemic we do not know how they should be handled more formally. If we had more information about them then we would have a better idea of what more sophisticated assumptions might be appropriate. If we can recognise important sources of epistemic error (such as the rainfall inputs over a catchment area in a hydrological model) then we might be able to invest more effort in trying to reduce those uncertainties and refine the approximate analysis.

This is the interpretation of critics of the subjectivity of GLUE likelihood choice (e.g. Mantovan and Todini 2006; Stedinger et al. 2008) who give preference to the explicit probabilistic predictions based on the statistical likelihood approach. The only problem with considering it as an approximation to a more realistic approach is, however, precisely in the unrealistic stretching of the likelihood surface that results, with a consequent over-conditioning of beliefs in the correctness of the model and its associated parameters. This seems to me to be a preference for mathematical formalism over common sense.

The second interpretation is to allow that the aleatory statistical assumptions overestimate the information content of the data being used to run and evaluate the model. We have suggestions of this in the identification of periods of hydrologically inconsistent data revealed in some applications of hydrological models (e.g. Beven and Smith 2015). This is a form of epistemic uncertainty, since we do not know what it is about those periods of data that results in the inconsistency in mass balance. Such periods should not, however, be used to evaluate a model based on a principle of mass balance as it might lead to incorrect inference. This suggests that what is needed is for the information content of the available data to be reflected in a definition of likelihood that does not result in extreme stretching of the relative likelihood of similarly acceptable models, i.e. which reflects the type of equifinality that is obvious in Fig. 32.1.

In GLUE we have a framework that allows the choice of any sensible likelihood measure, including the recognition of periods of disinformation (though that result in problems in validation and prediction, see next section). So, the question is how to define a likelihood measure that reflects the effects of epistemic uncertainties on the information content of the calibration data. This has proven difficult, for good epistemic reasons of course; we do not generally know enough about the data to be able to assess its real information content. Indeed the interpretation of the data itself often depends on a model with its own epistemic errors (e.g. the rating curve in estimating stream discharges, see Westerberg et al. 2011; Coxon et al. 2015; McMillan and Westerberg 2015; or in the processing and meaning of remote sensing images, see Franks and Beven 1999; Gagehan and Ehlers 2000; Lobell et al. 2003).

Recent studies within the GLUE methodology have been based on using limits of acceptability, defined prior to running any models and defined based on what is known about uncertainties in the input and evaluation data. Individual evaluation observations can be associated with their own limits of acceptability. The limits can also serve to determine whether a model is acceptable or not. It is often possible to make an estimate of the uncertainty associated with an observation itself. However, the main problem that arises with this approach is that whether a model produces a predicted value that falls within the observational uncertainty will also depend on the uncertainty (which may be both aleatory and epistemic) in the input variables, of which there may be several interacting sources. This could be handled by constructing many realisations of the inputs, but this would also require assumptions on which to base the realisations, which may not be easy to justify. It is a problem precisely because of that lack of knowledge so that any such assumptions would involve a high degree of subjectivity, especially if there is the potential for nonstationarity in the epistemic input uncertainties, as there often is.

A simple, but also subjective, solution is to use the limits of acceptability on the evaluation observations as a base, and to allow those to expand to a sensible degree to allow for the unknown input uncertainties. To make the limits of acceptability on different variables commensurate in this respect, they can be normalised to a common scale (e.g. -1 to $+1$ from minimum to maximum limits). How much expansion should be allowed? This clearly cannot be defined in anyway objectively but is related to what the modeller is prepared to accept as acceptable. It does raise the interesting possibility that the degree of expansion necessary to have any behavioural models might not be acceptable. This is discussed further in Beven and Lane (Chap. 6, this volume).

For the cases where, perhaps with expanded limits of acceptability, sampling of the model space results in some models that make predictions within the limits for all the evaluation observations, then we would expect relative belief in those models to be higher, the closer the predictions are to the observations. A simple weighting function within the limits of acceptability allows this to be taken into account. The individual weights can then be combined in different ways to produce a likelihood measure. This could include using the multiplicative Bayes' equation, but with large numbers of observations (say in a time series evaluation) this would result in a similar stretching of the likelihood surface. Thus an additive combination might be more appropriate, making the resulting likelihood measures more analogous to fuzzy possibilities than probabilities (see for example, Halpern 2005). Similar choices arise in combining the measures for different types of evaluation variable, or from different evaluation periods. We could still interpret the resulting weights (relative probabilities within the ensemble of behavioural models (and their associated residual distributions if necessary). Nearing et al. (2016), for example argue that probabilistic reasoning is the only consistent axiomatic framework for reasoning about uncertainty (see also O'Hagan and Oakley 2004, for a statistician's perspective). However, this will generally require an assumption that those probabilities are complete or bounded, which may not be the case given the approximate nature of the models and the epistemic uncertainties involved.

The subjective choices needed in this type of analysis result from common sense reasoning about the impacts of unknown epistemic uncertainties in the modelling process. They raise issues about how to interpret the resulting uncertainty estimates, or to convey the meaning of those estimates to stakeholders and users. Beven and Alcock (2012) have suggested that, since every uncertainty estimate is conditional on the assumptions on which it is based, one way of facilitating this communication is through the use of a condition tree, within which the choices and assumptions required to perform an analysis are listed and, as far as possible, justified. This also provides an audit trail for the analysis such that those assumptions can be checked (and perhaps understood more explicitly) by the potential users. Examples of such condition trees for some case studies of flood inundation mapping are given in Beven et al. (2014).

32.7 Equifinality and Model Validation in the Inexact Sciences

The previous sections have been primarily concerned with model evaluation based on data for some calibration period but, as noted earlier, many modelling studies include further validation checks on some "independent" data set for those models that survive calibration. The modeller is then often caught in a bind: between the desire to use as wide a range of conditions as possible in model calibration to refine the ensemble of models used in prediction, and the need to check model predictions on some validation data set, *before* making the predictions necessary for an application. The situation is made more difficult in the inexact sciences by the potential for epistemic errors in data sets, in ways that might include periods of inconsistent data or with sources of error that vary between different periods of data.

The type of ensemble framework for model evaluation suggested above provides a methodology for model validation despite these difficulties. It does so by allowing the initial estimates of likelihoods associated with the models in the behavioural set to be updated as new validation observations become available. This might also include some of the models that have previously been considered behavioural being rejected as a result of the new evaluations. In some cases, this might mean all the models being rejected (e.g. Page et al. 2007; Dean et al. 2009; Mitchell et al. 2011; Hollaway et al. 2018). In doing so, it is necessary to take account of potential uncertainties and inconsistencies in the data, particularly in avoiding false negatives in model rejection. Those that remain can be considered to be validated in the sense of surviving the further test. At each stage, the ensemble of behavioural models can be used to make predictions. This is a form of Bayesian updating, but not necessarily using a multiplicative Bayesian operator in doing the updating. This approach is quite general but puts more focus on a decision about what should be considered an acceptable or behavioural model, and what should be considered as invalidated as not useful for making predictions. This is explored in more detail in Beven and Lane (Chap. 6, this volume).

32.8 Discussion

In the decades since the concept of equifinality was introduced by von Bertalanffy (1951) and applied to dynamic models by Beven (1975, 1993, 2006), the idea of using ensembles of models in prediction has become much more accepted. Even in the field of earth system science and climate modelling, which remains strongly computationally constrained, limited ensemble of models are being used to inform future policy decisions, such as the Paris Agreement within the United Nations Framework Convention on Climate Change that went into effect after ratification by sufficient national governments in November 2016. The ensembles of predictions used are limited in terms of the number of future emissions scenarios used, the numbers of parameter sets run and the choices of process representations. They largely represent the collection of the best deterministic projections made by different global modelling groups around the world, with each model being "tuned" in different ways to historical data sets. These are generally given equal weight, and it is left up to the expert opinion of the modellers themselves as to whether they are adequately tuned to the past data. There is no reason, in principle, why such ensembles should not be weighted according to past performance (see Chap. 34 by Knutti et al. in this volume), even though many model configurations will have been rejected in the development and tuning process.

Larger ensembles of climate change projections have been generated and evaluated using rough limits of acceptability in reproducing past change (most recently using cloud computing resources, e.g. Evangelinos and Hill 2008; Frame et al. 2009; Fowler et al. 2010; Rowlands et al. 2012) but the models used have generally made use of coarser grids or simplifying assumptions to reduce the computational burden. The weight given to these studies has generally been much less because they do not provide the highest resolution projections. Even after conditioning on past data, they do, however, give some indication that the finer resolution models might be underestimating the potential range of future change.

In other areas of environmental science, it is computationally possible to do much more extensive explorations of model structures, parameter sets and boundary conditions. The ensembles of possible responses and potential for ensembles of acceptable models can be explored in much more detail in both model evaluation and predictions or projections. Given the epistemic uncertainties associated with environmental systems and other similar scientific domains, equifinality of model representations is a generic problem and should be considered explicit.

However, in most cases, the problem of how to deal with epistemic uncertainty in the driving variables and boundary conditions remains and has rarely been adequately addressed. It is very much easier to make the outputs conditional on an assumption that the input and boundary condition data are correct. Such an assumption is common in the application of formal statistical likelihoods. That is also one reason why it is so difficult to properly evaluate the real information content of data used for model calibration and validation. Consequently, how to define an appropriate likelihood measure, and how to decide whether models are really fit for purpose or should be given a likelihood of zero, remain contentious topics in the literature (see Beven and Lane, Chap. 6, this volume). It seems likely that the equifinality concept will persist into the future in the form of ensembles of (stochastic) model runs being used in estimating prediction uncertainties, but that much more research is needed into model limitations, information content of data and both quantitative and qualitative means of evaluating and validating models in the inexact sciences.

References

Bertalanffy, L. von. (1951). An outline of general systems theory. *British Journal for the Philosophy of Science, 1*, 134–165.

Bertalanffy, L. von. (1968). *General systems theory*. New York: Braziller.

Beven, K. J. (1975). *A deterministic spatially distributed model of catchment hydrology*. Unpublished Ph.D. thesis, University of East Anglia: Norwich, UK.

Beven, K. J. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources, 16,* 41–51.

Beven, K. J. (2002). Towards a coherent philosophy for environmental modelling. *Proceedings of the Royal Society of London A, 458,* 2465–2484.

Beven, K. J. (2006). A manifesto for the equifinality thesis. *J. Hydrology, 320,* 18–36.

Beven, K. J. (2009). *Environmental modelling: An uncertain future?*. London: Routledge.

Beven, K. J. (2012a). *Rainfall-runoff modelling: The primer* (2nd ed.). Chichester: Wiley-Blackwell.

Beven, K. J. (2012b). Causal models as multiple working hypotheses about environmental processes. *Comptes Rendus Geoscience, Académie de Sciences, Paris, 344,* 77–88. https://doi.org/10.1016/j.crte.2012.01.005.

Beven, K. J. (2016). EGU Leonardo lecture: Facets of Hydrology-epistemic error, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal, 61*(9), 1652–1665. https://doi.org/10.1080/02626667.2015.1031761.

Beven, K. J., & Kirkby, M. J. (1979). A physically-based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin, 24*(1), 43–69.

Beven, K. J., & Binley, A. M. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes, 6,* 279–298.

Beven, K. J., & Westerberg, I. (2011). On red herrings and real herrings: disinformation and information in hydrological inference. *Hydrological Processes, 25,* 1676–1680. https://doi.org/10.1002/hyp.7963.

Beven, K. J., & Alcock, R. (2012). Modelling everything everywhere: A new approach to decision making for water management under uncertainty. *Freshwater Biology, 56,* 124–132. https://doi.org/10.1111/j.1365-2427.2011.02592.x.

Beven, K., & Binley, A. (2014). GLUE: 20 years on. *Hydrological Processes, 28*(24), 5897–5918.

Beven, K. J., & Smith, P. J. (2015). Concepts of information content and likelihood in parameter calibration for hydrological simulation models. *ASCE Jornal of Hydrologic. Engineering*. https://doi.org/10.1061/(asce)he.1943-5584.0000991.

Beven, K. J., Smith, P. J., & Freer, J. (2008). So just why would a modeller choose to be incoherent? *Journal of Hydrology, 354,* 15–32.

Beven, K. J., Leedal, D. T., McCarthy, S. (2011a). Framework for assessing uncertainty in fluvial flood risk mapping, CIRIA report C721, 2014, at http://www.ciria.org/Resources/Free_publications/fluvial_flood_risk_mapping.aspx.

Beven, K., Smith, P. J., & Wood, A. (2011b). On the colour and spin of epistemic error (and what we might do about it). *Hydrology and Earth System Sciences*, *15*, 3123–3133. https://doi.org/10.5194/hess-15-3123-2011.

Beven, K. J., Leedal, D. T., & McCarthy, S. (2014). Framework for assessing uncertainty in fluvial flood risk mapping, CIRIA report C721. Available at http://www.ciria.org/Resources/Free_publications/fluvial_flood_risk_mapping.aspx.

Chorley, R. J. (1962). *Geomorphology and general systems theory*, U.S. Geological Survey, Prof. Paper 500-1B, Washington, DC.

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., & Smith, P. J. (2015). A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water Resources Research, 51*(7), 5531–5546.

Culling, W. E. H. (1957). Mulitcycle streams and the equilibrium theory of grade. *The Journal of Geology, 65*, 259–274.

Culling, W. E. H. (1987). Equifinality: Modern approaches to dynamical systems and their potential for geomorphological thought. *Transactions of the Institute of British Geographers, 13*, 345–360.

Dean, S., Freer, J. E., Beven, K. J., Wade, A. J., & Butterfield, D. (2009). Uncertainty assessment of a process-based integrated catchment model of phosphorus (INCA-P). *Stochastic Environmental Research and Risk Assessment, 2009*(23), 991–1010. https://doi.org/10.1007/s00477-008-0273-z.

Evangelinos, C., & Hill, C. (2008). Cloud computing for parallel scientific HPC applications: Feasibility of running coupled atmosphere-ocean climate models on Amazon's EC2. *Ratio*, *2*(2.40), 2–34.

Fowler, H. J., Cooley, D., Sain, S. R., & Thurston, M. (2010). Detecting change in UK extreme precipitation using results from the climateprediction. net BBC climate change experiment. *Extremes*, *13*(2), 241–267.

Frame, D. J., Aina, T., Christensen, C. M., Faull, N. E., Knight, S. H. E., Piani, C., et al. (2009). The climateprediction. net BBC climate change experiment: Design of the coupled model ensemble. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *367*(1890), 855–870.

Franks, S. W., & Beven, K. J. (1999). Conditioning a multiple patch SVAT model using uncertain time-space estimates of latent heat fluxes as inferred from remotely-sensed data. *Water Resources Research, 35*(9), 2751–2761.

Gahegan, M., & Ehlers, M. (2000). A framework for the modelling of uncertainty between remote sensing and geographic information systems. *ISPRS Journal of Photogrammetry and Remote Sensing, 55*(3), 176–188.

Gupta, H. V. & Kling, H. (2011). On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. *Water Resources Research*, *47*(10).

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology, 377*(1), 80–91.

Halpern, J. Y. (2005). *Reasoning about uncertainty*. Cambridge, MA: MIT Press.

Helmer, O., & Rescher, N. (1959). On an epistemology of the inexact sciences. *Management Science, 6*(1), 25–52.

Hollaway, M. J., Beven, K. J., Benskin, C. M. W. H., Collins, A. L., Evans, R., Falloon, P. D. et al. (2018). The challenges of modelling phosphorus in a headwater catchment: Applying a 'limits of acceptability' uncertainty framework to a water quality model, *Journal of Hydrology* (in press).

Hornberger, G. M., & Spear, R. C. (1981). An approach to the preliminary analysis of environmental systems. *Journal of Environmental Management, 12*, 7–18.

Klemes, V. (1986). Delettantism in hydrology: Transition or destiny? *Water Resources Research, 22,* S177–S188.

Lobell, D. B., Asner, G. P., Ortiz-Monasterio, J. I., & Benning, T. L. (2003). Remote sensing of regional crop production in the Yaqui Valley, Mexico: Estimates and uncertainties. *Agriculture, Ecosystems & Environment, 94*(2), 205–220.

Madsen, H. (2003). Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Advances in Water Resources, 26*(2), 205–216.

Montanari, A. (2005). Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research, 41*(8), W08406.

Mantovan, P., & Todini, E. (2006). Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. *Journal of Hydrology, 330*(1), 368–381.

Mitchell, S, Beven, K. J., Freer, J., & Law, B. (2011). Processes influencing model-data mismatch in drought-stressed, fire-disturbed, eddy flux sites. *JGR-Biosciences, 116*. https://doi.org/10.1029/2009jg001146.

McMillan, H. K., & Westerberg, I. K. (2015). Rating curve estimation under epistemic uncertainty. *Hydrological Processes, 29*(7), 1873–1882.

Mizukami, N., Rakovec, O., Newman, A., Clark, M., Wood, A., Gupta, H., et al. (2018). On the choice of calibration metrics for "high flow" estimation using hydrologic models. *Hydrology and Earth system Science Discussions*. https://www.hydrol-earth-syst-sci-discuss.net/hess-2018-391/.

Nash, J. E., & Sutcliffe, J. S. (1970). River-flow forecasting through conceptual models. 1. A discussion of principles. *Journal of Hydrology, 10,* 282–290.

Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., & Weijs, S. V. (2016). A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal, 61*(9), 1666–1678.

O'Hagan, A., & Oakley, A. E. (2004). Probability is perfect but we can't elicit it perfectly. *Reliability Engineering and System Safety, 85,* 239–248.

Page, T., Beven, K. J., & Freer, J. (2007). Modelling the chloride signal at the Plynlimon catchments, wales using a modified dynamic TOPMODEL. *Hydrological Processes, 21,* 292–307.

Pappenberger, F., Frodsham, K., Beven, K. J., Romanovicz, R., & Matgen, P. (2007). Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations. *Hydrology and Earth System Sciences, 11*(2), 739–752.

Pokhrel, P., Yilmaz, K. K., & Gupta, H. V. (2012). Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. *Journal of Hydrology, 418,* 49–60.

Refsgaard, J. C., & Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. *Water Resources Research, 32*(7), 2189–2202.

Rose, K. A., Smith, E. P., Gardner, R. H., Brenkert, A. L., & Bartell, S. M. (1991). Parameter sensitivities, Monte Carlo filtering, and model forecasting under uncertainty. *Journal of Forecasting, 10*(1–2), 117–133.

Romanowicz, R., Beven, K. J., & Tawn, J. (1994). Evaluation of predictive uncertainty in non-linear hydrological models using a Bayesian approach. In V. Barnett & K. F. Turkman (Eds.), *Statistics for the environment II. Water related issues* (pp. 297–317). Wiley.

Romanowicz, R., Beven, K. J., & Tawn, J. (1996). Bayesian calibration of flood inundation models. In M. G. Anderson, D. E. Walling, & P. D. Bates, (Eds.) *Floodplain Processes* (pp. 333–360).

Reusser, D. E., Blume, T., Schaefli, B., & Zehe, E. (2009). Analysing the temporal dynamics of model performance for hydrological models. *Hydrology and earth system sciences*, *13*(EPFL-ARTICLE-162488), 999–1018.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*(5).

Rowlands, D. J., Frame, D. J., Ackerley, D., Aina, T., Booth, B. B., Christensen, C., et al. (2012). Broad range of 2050 warming from an observationally constrained large climate model ensemble. *Nature Geoscience, 5*(4), 256–260.

Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes, 21*(15), 2075–2080.

Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, *46*(10).

Smith, P., Beven, K. J., & Tawn, J. A. (2008). Informal likelihood measures in model assessment: Theoretic development and investigation. *Advances in Water Resources, 31*(8), 1087–1100.

Sorooshian, S., & Gupta, H. V. (1995). Model calibration. In V. P. Singh (Ed.), *Computer models of watershed hydrology*. Highlands Ranch CO: Water Resource Publications.

Spear, R. C., & Hornberger, G. M. (1980). Eutrophication in peel inlet—II. Identification of critical uncertainties via generalized sensitivity analysis. *Water Research, 14*(1), 43–49.

Stedinger, J. R., Vogel, R. M., Lee, S. U., & Batchelder, R. (2008). Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research, 44*(12), W00806.

Thompson, T. D. (1961). *Numerical weather analysis and prediction*. New York: Macmillan.

Van Straten, G. T., & Keesman, K. J. (1991). Uncertainty propagation and speculation in projective forecasts of environmental change: A lake-eutrophication example. *Journal of Forecasting, 10*(1–2), 163–190.

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., & Sorooshian, S. (2003). Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research, 39*(8), W01214.

Westerberg, I., Guerrero, J. L., Seibert, J., Beven, K. J., & Halldin, S. (2011). Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes, 25*(4), 603–613.

Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1998). Multi-objective global optimization for hydrologic models. *Journal of Hydrology, 204*(1–4), 83–97.

Zhang, D., Beven, K. J., & Mermoud, A. (2006). A comparison of nonlinear least square and GLUE for model calibration and uncertainty estimation for pesticide transport in soils. *Advances in Water Resources, 29,* 1924–1933.

Chapter 33
Validation and Over-Parameterization—Experiences from Hydrological Modeling

Jan Seibert, Maria Staudinger and H. J. (Ilja) van Meerveld

Abstract Models that simulate environmental processes by quantifying fluxes and states vary largely in their complexity and number of parameters. Most models suffer from over-parameterization, meaning that the available information does not allow identification of all model parameters. Over-parameterization is a serious problem in environmental modeling, as it might imply that a model works well, but could do so for the wrong reasons. This can lead to unreliable results when the model is used to make predictions. Model testing, or model validation, is therefore crucial. Usually, in more complex models more, internal variables are explicitly simulated, and, thus, there are more opportunities for model testing against observations than is the case for simple models. Increasing model complexity, however, comes at the cost of more parameters, and therefore the risk for over-parameterization increases as well. In this chapter, we discuss different ways to validate models, which simulate hydrological processes at the catchment scale, and the balance between model testability and over-parameterization.

Keywords Model validation · Catchment modeling · Model complexity · Parameter identification

33.1 Introduction

Computer models that simulate environmental processes by quantifying fluxes and states vary largely in their complexity. In general, they have been developed for two main purposes: (1) for hypothesis testing and (2) to make predictions (Beven 1989). The time horizon for the predictions can vary largely. For example, for hydrological models, these include short-term forecasts, such as the streamflow in the coming hours or days, and long-term forecasts, such as the potential impacts of climate change on water resources. Like most models in environmental sciences,

J. Seibert (✉) · M. Staudinger · H. J. (Ilja) van Meerveld
Department of Geography, University of Zurich, Zurich CH 8057, Switzerland
e-mail: jan.seibert@geo.uzh.ch

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_33

hydrological models are over-parameterized, which means that even though a model appears to work well, it could do so for the wrong reasons. This can lead to unreliable predictions. Thus, if a model is used as a predictive tool, it needs to be tested extensively to ensure that it produces accurate output for the changed (future) setting in which we want to apply it (Grayson et al. 1992; Klemeš 1983, 1986a). Validation, as defined in this book, includes any activity that increases our confidence in model simulations. We use the term in a narrower sense and refer to validation as model testing using the model parameters that were determined for a calibration data set and then remain unchanged. Most hydrological models require at least some calibration because the parameters are not measurable or cannot be measured at the right scale (see Sect. 33.2.2). In catchment-scale hydrological modeling (but also other types of environmental modeling), often split sample tests are used for validation, where one part of the available data is used for model calibration and another part is reserved for testing. The aim of the testing is to ensure that the model also works for independent periods, for different catchments or for independent variables. In this chapter, we first discuss what is meant by over-parameterisation and related terms, and then focus on different ways in which hydrological models can be tested, or validated, after calibration to improve our trust in the model.

33.1.1 Over-Parameterization Terminology

Environmental studies are usually limited by the amount of available information. This restricts the possibilities to model environmental systems in their full complexity and heterogeneity. Hence, of necessity, the models aggregate the natural variability to a large degree. This leads to the fundamental problem that there is not just one possible model representation but different models can represent the available data equally well. Several terms are used to describe this situation; they are all caused by insufficient information but have slightly different meanings. *Underdetermination* is used in the philosophy of sciences to describe situations where there is insufficient evidence to decide between different theories or, with regard to environmental modeling, between different descriptions of a system (Laudan 1990). In environmental modeling, and especially in hydrological modeling, largely influenced by the work of Keith Beven (Beven 1993, 2006, see also Chap. 32 by Beven in this volume), the term equifinality is used to describe this situation. Two models or model parameterizations are said to be equifinal if they perform equally well in representing the available observations (note that the term *equifinality* is used differently in other fields. In geomorphology, for instance, the term is used to describe the situation where different processes lead to similar landforms). *Over-parameterization* describes the situation when more parameters are used in a model than can be identified based on the available information. In mathematical terms, over-parameterization refers to situations such as when a polynomial with more than n free parameters is fitted to n data pairs. In environmental modeling, the term is used in a slightly different meaning, which can be illustrated using an example with six data points (Fig. 33.1). Based on the information content of the data one can fit a fifth-degree polynomial function to the data. While all lower degree polynomial functions show a similar general trend, only the fifth-degree polynomial function fits all data points exactly. In environmental modeling, one would never try to fit each data point exactly because each data point (observation) is affected by observation uncertainty and the model is only an approximation of the much more complex reality. Hence, no environmental model is expected to reproduce each observation perfectly. The term over-parameterization is, thus, used in environmental modeling to describe the situation where a model contains more free parameters than are identifiable with some confidence. While in its mathematical meaning, it can be clearly decided whether a model is over-parameterized or not, this boundary is fuzzy for environmental models. Looking at the example with six data points (Fig. 33.1), it is not obvious which polynomial or, more general, which function, to choose. Different models, like the different polynomials in the example, usually do not vary much within the range of observations that are available for model calibration (situations like the behavior shown by the blue line of the fifth-degree polynomial between data points 1 and 2 or data points 5 and 6 in Fig. 33.1 are highly unusual for natural systems). It is more common that models differ largely in their simulations outside the calibration conditions. Usually one would choose a simple function that does not vary in a highly nonlinear way beyond the observed data points. Looking at Fig. 33.1, one might say that the first and second degree polynomials (red and orange curves) are most reliable for the extrapolation, but in practice, it is often not that easy to decide which model to trust. In some cases, a model that just continues the trend for the observed data may be less reliable than a model that predicts large changes beyond a certain point. Consider, for instance, a bucket with a small hole in the bottom, in which you pour one liter of water every minute. As long as the bucket is not completely filled, the outflow will slowly increase (due to the increasing water level, i.e., pressure). However, as soon as the bucket is filled, each addition of water will immediately cause the bucket to overflow, leading to a much larger increase in the flow rate. In this case, there is a clear (storage) threshold beyond which the outflow increases drastically. While this can be observed and simulated easily for the example of a leaky bucket, identifying and parameterizing these types of thresholds is more difficult for natural systems.

Obviously, there is an overlap between the terms underdetermination, equifinality, and over-parameterization, as well as other terms such as over-fitting. The subtle differences can be discussed using the example with six data points (Fig. 33.1) as well. The data do not allow one to decide for one single model or function (underdetermination). Furthermore, for each of the models (functions) there might be different parameterizations that represent the observed data points almost equally well, especially when considering the observation uncertainties (equifinality). Finally, with an increasing number of free parameters (here the degree of the polynomial function) the range of possible model behaviors increases, especially outside the observation range, which implies that over-parameterization might have a large effect on the predictions. For predictions (especially outside the observation range), one might want to restrict the number of parameters, although each additional parameter leads to a (slightly) better fit to the observations.

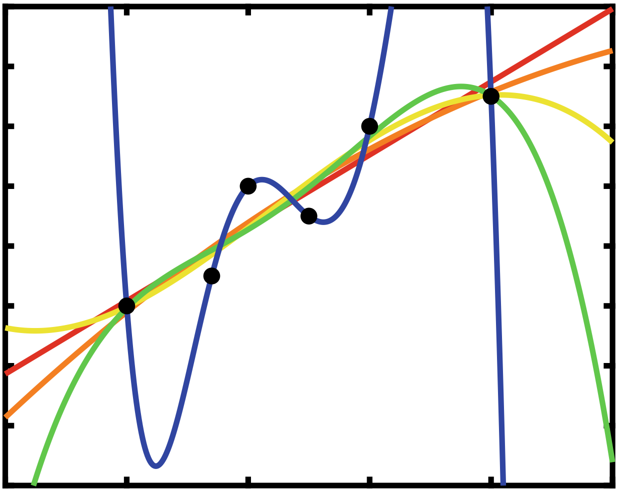


Fig. 33.1 Polynomial functions of different degree (1 (red) to 5 (blue)) fitted to six data points (black circles). While the higher degree polynomial functions reproduce the data points better, and for the fifth-degree function even perfectly, the lower degree functions are more robust between and outside the data points. See text for a further discussion

33.1.2 Main Types of Hydrological Models

In this chapter, we focus on hydrological models that simulate streamflow responses to rainfall and snowmelt events and the corresponding changes in the amount of water stored in the catchment. These models represent hydrological processes (see Textbox: Short description of catchment hydrology) in different ways and vary in complexity and range from simple spatially lumped models to complex, spatially explicit models. The number of parameters ranges from three to six for parsimonious models, up to more than a hundred for complex models. The code representing the models similarly varies from a few lines to many pages, and the time needed to simulate, for instance, one year of streamflow, varies from less than a second to several hours (even with today's computers). The overall model complexity consists of process complexity, i.e., the number of hydrological processes that the model explicitly represents, and spatial complexity, i.e., the degree of spatial discretization and connectivity (Clark et al. 2016). While these two types of complexity are often connected, they do not necessarily have to be connected. A model can have a simple process representation but a detailed spatial discretization, which can be motivated by the availability of spatial information for factors that affect the hydrological response, such as topography (e.g., Grabs et al. 2009). The opposite, i.e., a complex process representation with a rough spatial discretization usually makes little sense for catchment-scale hydrological studies because the physical realism of the more detailed process representation diminishes when averaged over larger heterogeneous areas (Beven 1989; Kirchner 2006a).

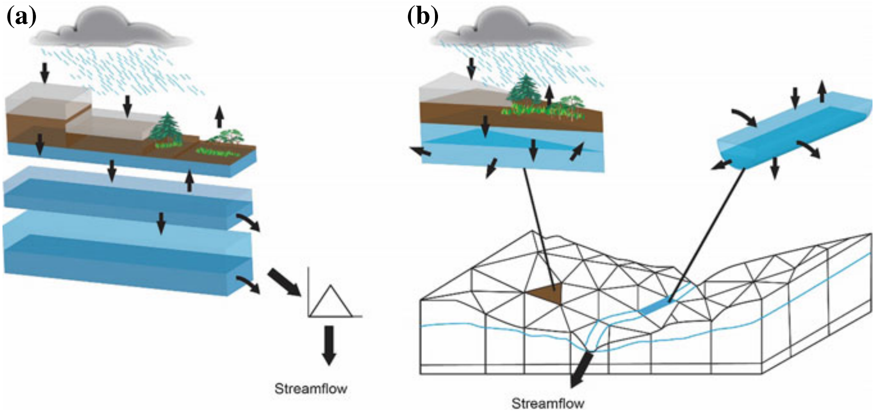


Fig. 33.2 Schematic overview of the two main hydrological catchment model types: **a** conceptual (bucket-type) model, **b** spatially explicit physically based model

Within this variety of catchment-scale hydrological models, there are three main model families: (1) purely empirical models (also called black-box models), (2) bucket-type models, which represent the fluxes in a more conceptual way (also called conceptual models), and (3) physically based models. Note that the term physically based model has also been used for bucket-type models in the literature on hydrological modeling. Here, however, we use the term in the strict meaning of a model that represents the fluxes based on equations that explicitly describe the small-scale physics, such as the Darcy-Richards equation.

In black-box models, the empirical relationships are based on observed input and output data without any attempt to represent hydrological processes. These empirical relationships can be derived, for instance, from regression equations or artificial neural networks. This model family provides very limited opportunities for model validation because there are no additional internal variables and therefore we will not consider the black-box models in the remainder of this chapter.

Bucket-type models represent a catchment in a lumped way, i.e., they represent an entire catchment by a few boxes that can store and drain water (Fig. 33.2a). Examples for bucket-type models are the HBV model (Lindström 1997), the Variable Infiltration Capacity model (ARNO/VIC) (Liang et al. 1994), and the GR4 J model (Perrin et al. 2003). Parameter values in bucket-type models are, in general, not directly observable as they aggregate hydrological processes over the entire catchment or large parts of it. The values are, thus, effective values at the catchment scale; in most cases, there are no clear aggregation schemes to derive these effective values from small-scale measurements. Consequently, model parameter values for bucket-type models are determined indirectly by model calibration. In the case of ungauged catchments (i.e., catchments for which no calibration data are available) regionalisation of parameter values is needed to estimate the parameter values for the studied catchment (Viviroli et al. 2009).

Physically based models usually represent the catchment in a more spatially explicit way and simulate fluxes for smaller elements (e.g., grid cells), which together represent the entire catchment (Fig. 33.2b). Examples of spatially explicit physically based models are InHM (VanderKwaak and Loague 2001; Smerdon et al. 2007; Mirus et al. 2011), CATHY (Bixio et al. 2002; Camporese et al. 2010), tRIBS (Ivanov et al. 2004), HydroGeoSphere (Jones et al. 2006, 2008; Brunner and Simmons 2012), and MIKE SHE (Hansen et al. 2013). While these models in theory can represent all hydrological processes in a physically based way, in reality often simplified representations are used for some of the processes. When hydrologists or hydrogeologists, for instance, apply the same model to the same catchment, the actual model application can look rather different (Staudinger et al., in review). One assumption of spatially explicit physically based models is that the physical processes within a catchment can be represented in a deterministic way and that the final catchment response is the combination of the processes in the individual areas. Further, it is assumed that the spatial variability in hydrological processes and responses in the catchment can be characterized by spatially varying values of the model parameters.

Despite the term physically based, these models are not always as physically based as one might think at first glance. Many physically based models require information on soil physical parameters, such as porosity, hydraulic conductivity, and the relation between soil moisture and matric potential (i.e., pF curve) (see Textbox: Short description of catchment hydrology). While these variables can be measured for soil cores, it is often difficult (or impossible) to use this information directly in hydrological models because soil cores do not capture the large heterogeneity in the soil and hydrological processes are highly nonlinear. As a result of these spatial heterogeneities, the hydraulic conductivity often increases significantly with scale (Schulze-Makuch et al. 1999; Martinez-Landa and Carrera 2005): the saturated hydraulic conductivity of small soil cores is often at least an order of magnitude smaller than the saturated hydraulic conductivity of large soil cores, which in turn is at least an order of magnitude smaller than the effective hydraulic conductivity obtained from fitting hydrological models (Brooks et al. 2004). This is largely due to the presence of macropores, which can transmit water at much higher rates than matrix flow (Beven and Germann 1982), but are generally insufficiently included in small soil cores. Comparable to soil macropores, but at a larger scale, karstic systems with flow pathways through caves similarly challenge the use of Darcy's law to describe groundwater flow (e.g., Hartmann et al. 2014). Since the Darcy-Richards equation that is used in many physically based models does not consider flow through macropores or fracture flow, often effective parameters for soil hydraulic characteristics need to be used. As these effective parameters cannot be obtained from soil cores or field data, they need to be obtained through model calibration. The usual procedure is to define a limited number of soil classes, estimate parameter values for these classes and then assign a class to each grid cell in the model. While this is a pragmatic solution, the physical realism is compromised by this approach and the predictive capability of the model is reduced because the parameter values cannot be determined a priori (Beven 1983; Grayson et al. 1992; James and Burges 1982). It might even be questionable, whether there is an effective parameter value at the scale of the grid cell, as the functional behavior might differ more fundamentally than is represented by a single effective parameter (Kirchner 2006b; Weiler 2017). For instance, regarding the application of the Richards equation, Weiler (2017) argues:

> "It is really difficult to see how such a misuse of 'physics' is justified" and "Although hundreds of papers have been published showing that the capillary flow hypothesis of the Richards equations does not work in many field soils, hydrology as a field is still unwilling to reject it." (p. 16).

33.1.3 Peculiarities of Hydrological Models

33.1.3.1 Model Development

Many hydrological models have been developed over several decades and the original purpose of the model may have gotten lost along the way. Often various people or groups have contributed to the development of a model, and different routines have been added during the years of continued model development for specific questions or case-specific data availability. As a result, many models now have a "shanty town appearance" (Clark et al. 2017) and for more complex models it can become challenging to fully understand the model code, including the assumptions and limitations behind the computations. This can lead to models being applied beyond their original scope and hence beyond their capabilities (Grayson et al. 1992). While motivated by individual situations, often these incremental model developments lead to predictive models that are over-parameterized. Also, internal variables of the models may have names that suggest that they are a physical variable but do not necessarily represent this variable at a measurable scale, but rather (if at all) at some effective, integrated scale. This means that validation against these variables is not straightforward (or even impossible).

33.1.3.2 Dominance of Streamflow Data for Model Testing

One peculiarity of catchment-scale hydrological modeling is the existence of one dominant variable that is used for model calibration and validation, namely streamflow. While most catchment models also simulate other variables, streamflow time series are the first and most important information used to evaluate model performance. In climate modeling, for instance, the situation is quite different; there are many potential variables to evaluate climate models against, and there is much more variation in which one(s) modelers use for model evaluation (Hourdin et al. 2016).

Streamflow is the component of the water cycle that is easiest to measure at the catchment scale due to its aggregated nature, i.e., streamflow observed at one location provides information about the entire upstream catchment. Other variables, such as groundwater levels, soil moisture or evapotranspiration vary largely in space, so that it is difficult to derive a single representative value at the catchment scale. Obviously, streamflow is also a very important variable for many practical issues, such as flood management and water supply. As a result, streamflow serves as the main variable in model calibration and validation, even though it might not be the most important hydrologic variable in all situations and is often also not the largest flux leaving a catchment (evapotranspiration is often larger). The focus on streamflow also implies the risk that certain model parts (routines), which are less crucial for streamflow simulations, are not tested in enough detail and quick fixes (kluges) might remain in the model code (see discussion on holism versus modularity, and on kluging in Chap. 39 by Lenhard in this volume).

Streamflow is measured at the outlet of a catchment and integrates the different hydrological processes that contribute to the streamflow response and their spatial variability. Hydrological processes at the small scale are often highly nonlinear, spatially variable and affected by thresholds (Seyfried and Wilcox 1995). However, streamflow responses at the catchment scale are in most cases rather smooth. For modeling, this is both a blessing and a curse. On the one hand, streamflow at the catchment scale can be modeled relatively well without explicitly considering all small-scale heterogeneities. On the other hand, even very simplistic models without any physical realism can produce relatively good fits between the observed and simulated streamflow (Grayson et al. 1992; Seibert and McDonnell 2002; Kirchner 2006a). This, in turn, implies that a good model fit for streamflow is by no means an indication for a correct model and there is a high risk of being right for the wrong reasons (Klemeš 1997; Loague and VanderKwaak 2004a; Kirchner 2006a).

The clear focus on one variable might have contributed to two further aspects of hydrological modeling: (1) there is a large body of literature on model calibration approaches, including the consideration of uncertainties (for a summary see, for instance, Chap. 7 ("Parameter Estimation and Predictive Uncertainty") of Beven 2012) and (2) there is a relatively broad consensus about the calibration criteria. Most studies evaluate model performance using the model efficiency (Nash and Sutcliffe 1970), which is the scaled sum of squared errors, although recently other criteria, such as the Kling-Gupta efficiency (Gupta et al. 2009; Pool et al. 2018), have gained popularity. In other environmental sciences, e.g., climate modeling, the situation is quite different: calibration is seen as a rather 'ugly' business and is usually called tuning. There are only few papers describing and evaluating the tuning (or calibration) approaches in climate modeling and different research groups use different evaluation criteria (Hourdin et al. 2016).

33.2 Types of Validation in Hydrological Modeling

33.2.1 Validation Based on Independent Time Periods

Over-parameterization may lead to a model that works fine for the period for which it has been calibrated but performs poorly for an independent period. In hydrological modeling, however, this is hardly ever observed because generally the two time periods used for model calibration and validation are not that different. Hence, models tend to perform approximately similar for consecutive calibration and validation periods (classic split sample test; Klemeš 1986b).

The differential split sample test assesses if the model performs well for a particular transition and is, thus, a more powerful test of model performance (Klemeš 1986a). For this test, the period used for validation is chosen such that it is in some aspect different from the data used for model development and calibration. For instance, a model can be calibrated on data from years with only small floods and then be validated based on data from years with larger floods (Seibert 2003; Coron et al. 2012; Dakhlaoui et al. 2017). For instance, to determine the hydrological impacts of climate change, the periods should be split into a drier and a wetter period and the one that is assumed to represent the future conditions is used for validation (Klemeš 1986a). This test is generally applicable and not model type dependent.

33.2.2 Validation Based on Independent Catchments

Besides testing for different time periods, we can also validate the model at a different location. The proxy-basin test (Klemeš 1986a) is a basic test for the geographical transferability of a model within a region with similar hydro-climatological, geological and land use characteristics. The model is calibrated for two gauged catchments from a specific region and then reciprocally validated. Only if the validation is satisfactory for both catchments, the model can be used for a third (ungauged) catchment in that region (Refsgaard and Knudsen 1996; Motovilov et al. 1999). This test is so far not used regularly, in part due to the uniqueness of each catchment (Beven 2000), which makes it difficult to find similar catchments. Usually, even neighboring catchments are very different in their response to rainfall and snowmelt and, therefore, their model parameterizations might vary largely. The validation based on independent catchments is especially applicable to the more complex physically based models because in these models differences in catchment characteristics, such as different soil depths or land use, can be explicitly considered (if the data are available for both catchments).

33.2.3 Validation Based on Independent Variables

Validation is also possible by evaluating the model simulations against other variables than those used in calibration. This can either be a particular aspect of streamflow that was not directly used in calibration, streamflow at different locations in the catchment or different variables, such as groundwater levels, stream chemistry or snow cover. The validation on independent variables is principally possible for all model types, but for simple models the observations have to be transferred into a quantity that is comparable to the model simulations.

33.2.3.1 Stream Flow Signatures

When a model is calibrated based on observed streamflow data using criteria such as the model efficiency (Nash and Sutcliffe 1970), this does not imply that all aspects of streamflow are simulated correctly. For example, it is well known that the model efficiency puts more emphasis on high flows (Krause and Boyle 2005; Schaefli and Gupta 2007) and consequently the streamflow simulations might be rather poor when evaluated against other streamflow-based criteria, such as low flows, recession shapes, flow duration curve characteristics, etc. (Vis et al. 2015). Therefore, also aspects of the streamflow time series or the variability of streamflow, so-called signatures (Gupta et al. 2008; Yilmaz et al. 2008; Euser et al. 2013) or streamflow characteristics (Vis et al. 2015; Pool et al. 2017), can be used for model validation.

33.2.3.2 Spatial Variation of Streamflow

Within a catchment, there is usually a large spatial variation in streamflow due to, for instance, differences in topography, geology or vegetation (Karlsen et al. 2016), which can be used for model validation. A model can be calibrated using streamflow data for the catchment outlet and then validated using streamflow data from subcatchments (e.g., Ambroise et al. 1995; Refsgaard 1997; Uhlenbrook and Leibundgut 2002). In this case, however, the validation is not fully independent as the model has been calibrated against 'part of the streamflow' because streamflow along a stream network is obviously not independent (i.e., the downstream streamflow used for calibration includes (parts of) the streamflow for an upstream sub-catchment) (Seibert 2001a). In such situations it is necessary to use an appropriate benchmark, such as the specific runoff time series (Seibert 2001b).

33.2.3.3 Other Hydrological Variables

Models can also be validated using variables that were not used in model calibration. Usually, a model simulates one (or a few) main variables (as mentioned for hydrological models this is in most cases streamflow), but there are also other model variables that describe different fluxes and states inside the model. Evaluating the simulation of these variables against observations during model validation is somewhat similar to the differential split sample test described above, but instead of testing the model for a different time-period, the simulations are validated for a different variable. Comparison against internal variables can be done in a validation type approach, i.e., after the model has been calibrated, but also during the calibration. This latter procedure, also called multi-criteria calibration (Seibert 2000; Vaché et al. 2004), aims to reduce the parameter uncertainty during calibration. The idea of multi-criteria calibration is that if the simulation aims for good fits for several variables, the parameter space will be more constrained (Seibert and McDonnell 2002). However, the value of the additional variables for model calibration and validation differs. In an application of a catchment-scale water and salinity model, for instance, it was found that groundwater level data did not help to constrain model parameters but salinity data did (Kuczera and Mroczkowski 1998).

One of the issues with model validation based on hydrological variables other than streamflow is that the scale of the measurements is typically very small (particularly compared to the size of the grid cell or the entire catchment) and the spatial variability is large so that effective values need to be used. Another issue is that some of these variables represent properties below the surface and are difficult to observe and measure. In general, point measurements need to be transformed so that they can be used in model validation. When spatially lumped model simulations are compared with point measurements, such as groundwater level, soil moisture or snow depth measurements, there are two options: (1) the model simulations are compared to spatially averaged values (e.g., the mean value of several groundwater levels or snow depth observations) or (2) the simulations are compared to relative values or dynamics, rather than the exact values. In the first case, the challenge is to derive a meaningful average, which usually includes determining weights for different observations based on their representativeness. In the second case, only part of the information is used for model validation.

Some studies have attempted to validate bucket-type models against internal variables, such as for instance groundwater dynamics or the isotopic composition of streamflow (Seibert and McDonnell 2002; Fenicia et al. 2008). These studies have shown that this information is useful for choosing plausible model structures. However, mainly due to the challenges in comparing point observations and simulated catchment-scale values, bucket-type models are not routinely validated against internal variables.

For spatially explicit models, validation against field observations is (in theory) more straightforward, but this type of validation is still not as widespread and detailed as one may expect. One of the main challenges is the lack of suitable data for internal model testing (Loague and VanderKwaak 2004b; Loague and Ebel 2016). Stephenson and Freeze (1974) validated a spatially distributed model against streamflow data and information on water tables and vertical hydraulic gradients from piezometers, pressure heads from tensiometers and soil moisture. Despite the large number of measurements that were available, the test failed due to a lack of adequate data, due to limitations in the model, due to uncertainties in initial and boundary conditions and due to computer limitations. Now, more than 40 years later, the computer limitations are less of an issue, but there are still relatively few studies that validate a model against other variables (Parkin et al. 1996; Refsgaard 1997; Loague and VanderKwaak 2002, 2004a; Bathurst et al. 2004). Almost all detailed validation studies that are reported in the literature were done for a handful of relatively small intensively studied catchments, such as Coos Bay (Ebel et al. 2007), Tarawarra (Western et al. 1999a) and R-5 (Loague and VanderKwaak 2002).

A direct comparison with observed data is not always possible for spatially explicit models either. For example, a model might be based on a grid cell representation of a catchment with a resolution of 250 m by 250 m. Within such a grid cell, variables like soil moisture might vary considerably and the simulated value only represents the mean behavior of that grid cell. In other words, even here, some averaging of the observations or disaggregation of the simulations is required before model simulations can be compared to observations.

Snow

In snow-dominated catchments, snow observations can be beneficial for model validation because the hydrological response of these catchments is closely linked to snow accumulation and melt (Parajka and Blöschl 2008; Magnusson et al. 2014, 2015; Finger et al. 2015; Griessinger et al. 2016). Usually, the spatial extent of the snow cover or the snow water equivalent (i.e., the amount of water if the snow would melt) is used for model validation. These data can be derived from point observations (Magnusson et al. 2014), digital photogrammetry (Bühler et al. 2015) or satellite remote sensing based on passive microwave and visible spectrum imagery (e.g., Andreadis and Lettenmaier 2006; Durand and Margulis 2006; Dong et al. 2007). In catchments with glaciers, glacier mass balance information can be used in addition to the snow observation for multi-criteria model validation (Koboltschnig et al. 2008; Finger et al. 2015).

Below-Ground Storage

If we could easily measure the amount of water that is stored below the ground at the catchment scale, streamflow would probably not have been the single most used variable for model testing. New data on the amount of water stored in the subsurface, e.g., from gravimeters, may be useful for model validation but their use for catchment-scale hydrological model validation has been limited so far. Global hydrological models have been validated using remotely sensed storage dynamics (Werth et al. 2009; Milzow et al. 2011) but this information is too coarse to be used for catchment-scale hydrological models.

Soil moisture and groundwater level data can provide information on the dynamics (and spatial variability) of the amount of water that is stored in the catchment. However, usually these are point observations and the challenge is to translate the point data into something that can be compared to model simulations. In the case of lumped models, some catchment average has to be computed from the observations. Besides a direct comparison of absolute groundwater levels, also relative groundwater levels or groundwater level dynamics can be used for model validation (Refsgaard 1997). For spatially explicit models, observations and simulations can either be compared point by point or their frequency distributions can be compared. Tromp-van Meerveld and Weiler (2008) demonstrated the value of the latter approach to compare observed and simulated groundwater levels for a hillslope hydrological model. Additionally, patterns of saturated areas (i.e., areas where the groundwater level is at or near the surface) can be used for model validation (e.g., Grayson and Blöschl 2001; Blazkova et al. 2002; Glaser et al. 2016).

For lumped models (or spatially distributed models with large grid sizes relative to the spatial variability), the measured groundwater levels need to be aggregated. This aggregation is usually not straightforward because groundwater levels vary spatially; they are very different in the riparian zone and in hillslopes (Seibert et al. 2003; Detty and McGuire 2010; Haught and van Meerveld 2011; van Meerveld et al. 2015). Information on shallow groundwater levels is often only available for a few points in the catchment and information on deep groundwater storage is usually absent. Groundwater wells are frequently only installed in areas where shallow groundwater is expected to be present regularly (e.g., riparian areas) and not on upper hillslopes or areas with very shallow soils, so that it is difficult to determine the change in the total amount of groundwater stored in the catchment. Groundwater levels may also drop below the depth of the well so that the actual groundwater level is not known (this is particularly the case for hillslope and shallow soil areas).

Spatial information on soil moisture has been used effectively in model validation (Western et al. 1999b; Mirus et al. 2009) but there are very few detailed soil moisture data sets that can be used for model evaluation. Soil moisture measurements often represent only a very small area (~0.1–1 dm^3; Robinson et al. (2008)), while soil moisture varies largely in space (in all three dimensions). Remotely sensed soil moisture data can, in theory, be used for model validation but these measurements only reflect the soil moisture content of the topsoil and often integrate over a relatively large area, which is particularly problematic in hilly and mountainous terrain where soil moisture varies greatly with topography. Geophysical measurements (e.g., electromagnetic induction (Kachanoski et al. 1990; Sheets and Hendrickx 1995)) and mobile cosmic ray neutron probe measurements (Rivera Villarreyes et al. 2011) can provide information on soil moisture storage over larger areas. The pattern of vegetation or certain types of vegetation can also give an indication of the moisture state (e.g., waterlogged conditions for wetland plants) and can, therefore, be used to represent the spatial variability in soil moisture.

Stream Chemistry

Stream chemistry data has not only been used in the calibration and validation of hydrological models to reduce parameter uncertainty (Mroczkowski et al. 1997; Kuczera and Mroczkowski 1998), but also to compare and test different model structures (Vaché and McDonnell 2006; Fenicia et al. 2008; Birkel et al. 2011; Davies et al. 2011; McMillan et al. 2012; Hartmann et al. 2013; Stadnyk et al. 2013). Conservative tracers, such as water isotopes (oxygen-18, deuterium) and chloride (which is contained in the precipitation), are particularly useful for model testing (Birkel and Soulsby 2015). Stream chemistry data can be used for model validation either directly by simulating time series of the streamflow isotopic composition or concentrations or indirectly by using the fractions of new water for events (signatures) (de Grosbois et al. 1988; Mroczkowski et al. 1997; Kuczera and Mroczkowski 1998; Uhlenbrook and Leibundgut 2002; McGuire et al. 2007; Stadnyk et al. 2013; Birkel and Soulsby 2015; van Huijgevoort et al. 2016). Recent stream chemistry data sets are both longer (Kirchner and Neal 2013) and in many cases now also available at a higher resolution (van Huijgevoort et al. 2016; von Freyberg et al. 2018) than in the past, which opens new opportunities for model validation.

In order to use stream chemistry data for model validation, the model has to include a representation of mixing in the catchment. For the simulation of streamflow, it is sufficient to simulate the hydraulic response, whereas for the simulation of a tracer response the mixing of different water sources with water that was already stored in the catchment before a rainfall or snowmelt event has to be considered. Even for a very simple approach (i.e., complete mixing), this usually requires additional parameters (Barnes and Bonell 1996; Birkel and Soulsby 2015), although Remondi et al. (2018) showed that a spatially explicit physically based model can represent the observed concentrations in the stream reasonably well without additional model parameters to describe the mixing. While some studies have shown that the use of stream chemistry data improved parameter identifiability (de Grosbois et al. 1988; Kuczera and Mroczkowski 1998; McGuire et al. 2007; Stadnyk et al. 2013) and that it was thus worth the additional parameters (e.g., Iorgulescu et al. 2005), other studies have shown that these additional parameters may not be identifiable (Seibert et al. 2003; Hrachowitz et al. 2013). The tracer signal is in some cases so damped that it contains almost no additional information for model calibration (Dunn et al. 2008a). This led to the cautious statement by Turner and Barnes (1998, p. 725):

> One question that can be legitimately raised in the application of environmental isotope or hydrogeochemical data to catchment studies is whether or not the complementary use of the additional data simply adds to the problem of over-parameterization.

33.3 Conclusions—All Models Are Wrong, but Which Are Useful?

Hydrological models usually contain more than the two to four parameters that can be identified by the information contained in the streamflow data that are used for model calibration. We usually develop models with more parameters because we want to represent the dominant hydrological processes (see Box 1) in the model. As a result, models can contain a dozen or more parameters. Some systems are also not simple enough to be described by simple models with only very few parameters (Ebel and Loague 2006).

Over-parameterization leads to the risk that the hydrological model can produce nice fits to the streamflow observations but does so for the wrong reasons, i.e., internally, the model does not represent the hydrological processes, and the model hence cannot be used for predictions. We discussed different ways to validate hydrological models and show that multi-criteria model validation can be useful. However model validation is limited by both the complexity of the model and the ability to measure the data that are necessary for model validation. In addition, there is a balance between model parsimony and options for model validation, which needs to be considered and evaluated for each case individually. Model parsimony reduces the

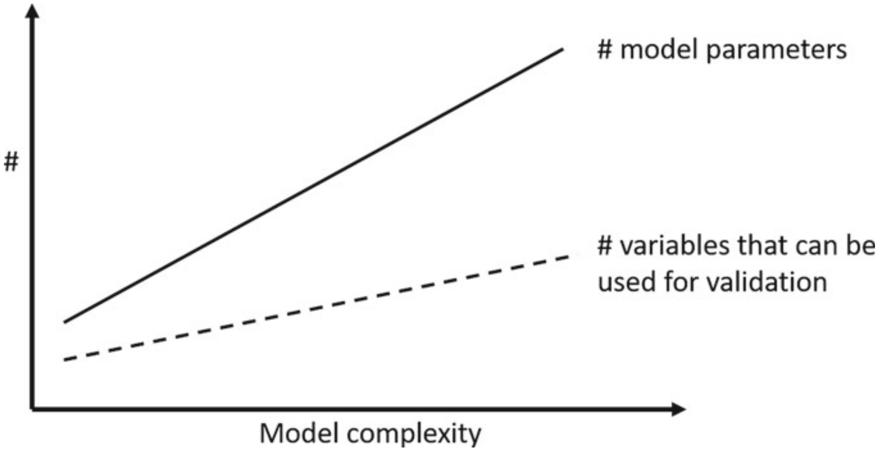


Fig. 33.3 As model complexity increases, the number of model parameters and the number of variables that can be used for validation increase (note that the rate at which both increase with model complexity depends on the model, parameters and the information content of the data)

risk for over-parameterization, while model complexity increases the possibilities to evaluate different simulated states and fluxes. Simple models with few parameters might avoid over-parameterization, but also provide few opportunities for model testing based on other variables than the main model output (i.e., streamflow). In order to have more options for (internal) model testing, we need models that are more detailed and these have, as a direct consequence, more parameters (Fig. 33.3).

Logically, if more complex models are used then there is a need for richer data-sets (Ebel and Loague 2006). Stephenson and Freeze stated already in 1974 (Stephenson and Freeze 1974) that massive data collection campaigns would be needed for the application of more complex models. Despite all progress in observational hydrology, much remains to be done to achieve a better link between field observations and modeling (Seibert and McDonnell 2002; e.g., Clark et al. 2017). Especially, it seems to be important that modelers evaluate and communicate better which data are most informative and, thus, which data should be collected to improve model validation and to reduce the risk for over-parameterization. Hence, a possible solution to better validate (particularly spatially explicit) models is to launch targeted measurement campaigns designed to validate models (Ebel and Loague 2006; Beven 2012; Zehe et al. 2014; Clark et al. 2016) and to use the model output to further refine the measurement campaign (Dunn et al. 2008b; Kikuchi 2017; Leaf 2017). A more fruitful communication between modelers and field hydrologists has been called for since decades (e.g., Dunne 1983) but is still a desideratum today. In addition, there is a need to find ways to measure hydrological variables across larger areas (e.g., catchment storage instead of ground water level or soil moisture measurements at a point), so that they can be more easily compared to simulated variables. Already in 1986, Klemeš expressed the need for new measurement methods for improved model testing, a quest that is still valid today:

> […] new measurement methods that would yield areal distributions, or at least reliable areal totals or averages of hydrologic variables such as precipitation, evapotranspiration, and soil moisture would be a much better investment for hydrology than the continuous pursuit of a perfect massage that would squeeze the nonexistent information out of the few poor anaemic point measurements. (Klemeš 1986b, p. 187S)

Textbox: Short Description of Catchment Hydrology

A catchment is the area that contributes to streamflow at a certain point and is usually determined based on the surface topography (Fig. 33.4). Precipitation can fall as snow or rainfall. A fraction of the precipitation is intercepted by the vegetation and evaporates from the leaves or needles without ever reaching the soil surface. While rainwater that is not intercepted by the vegetation infiltrates directly into the soil, snow can accumulate on the surface until it melts, resulting in a significant time lag between snowfall and infiltration. If the soils are saturated, i.e., the pores between the soil particles are completely filled with water, or if the precipitation intensity is higher than the infiltration capacity, part of the rain or meltwater does not infiltrate but flows over the soil surface to the stream as surface runoff. The infiltrated water is stored in the soil as soil moisture storage or percolates through the soil to the groundwater. Groundwater eventually flows to the stream and becomes streamflow, but flow times for groundwater can vary largely. Some of the water stored in the soil (and groundwater) is taken up by plants and transpired. To describe the sum of all forms of evaporation (from the canopy and soil surface) and transpiration, the term evapotranspiration is used. In the unsaturated zone, the soil pores are not completely filled with water but partly filled with air; water is held in the pores by capillary and adhesive forces. The amount of water that a soil can hold against gravity, also called the field capacity, depends largely on the size distribution of the pores. In a sand soil with large pores, only little water can be held against gravity, whereas field capacity is much higher in loamy or silty soils with smaller pores. However, plants can only extract water that is held not too tightly in the soil. The so-called wilting point describes the amount of water in the soil that cannot be extracted by the plants. The soil moisture retention (pF) curve describes the amount of water in the soil and how strongly it is held back (matric potential). Flow velocities and water fluxes in the subsurface can be computed based on the hydraulic conductivity and pressure gradients (Darcy and Darcy-Richardson equation). The hydraulic conductivity describes how quickly water can flow through the soil and is highly dependent on the moisture content of the soil. Water flows faster through large pores than small pores. Therefore, a large portion of the water may flow through macropores (i.e., large pores or macropores such as cracks, old roots, or animal burrows), particularly when the soil is wet and these macro pores are filled with water.

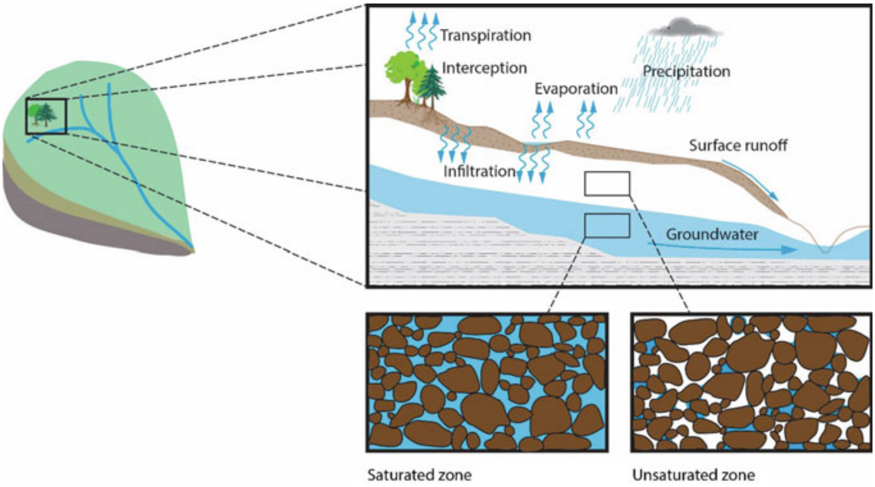


Fig. 33.4 Overview of the main hydrological processes in a catchment

Over longer time periods (i.e., at least one year), the inputs and outputs are balanced so that the sum of streamflow and evapotranspiration are equal to the precipitation. If shorter periods are considered, storage changes must be considered in the water balance. The storage term includes all water that is stored in snow and ice, soil- or groundwater and lakes. The changes in storage are important as streamflow responses to rainfall and snowmelt are largely dependent on the storage conditions. Hydrological models simulate how precipitation eventually leaves the catchment as streamflow, including the above-described processes in more or less detail. While there are different types of hydrological models as discussed in Sect. 33.2.2, all catchment-scale hydrological models include one or several differential equations to describe the relation between storage and streamflow.

References

Ambroise, B., Perrin, J. L., & Reutenauer, D. (1995). Multicriterion validation of a semidistributed conceptual model of the water cycle in the Fecht Catchment (Vosges Massif, France). *Water Resources Research, 31*(6), 1467–1481. https://doi.org/10.1029/94WR03293.

Andreadis, K. M., & Lettenmaier, D. P. (2006). Assimilating remotely sensed snow observations into a macroscale hydrology model. *Advances in Water Resources, 29*(6), 872–886. https://doi.org/10.1016/j.advwatres.2005.08.004.

Barnes, C. J., & Bonell, M. (1996). Application of the Unit hydrograph techniques to solute transport in catchments. *Hydrological Processes, 10*(6), 793–802. https://doi.org/10.1002/(SICI)1099-1085(199606)10:6%3c793:AID-HYP372%3e3.0.CO;2-K.

Bathurst, J., Ewen, J, Parkin, G, O'Connell, P., & Cooper, J. (2004). Validation of catchment models for predicting land-use and climate change impacts. 3. Blind validation for internal and outlet responses. *Journal of Hydrology, 287*(1–4), 74–94. https://doi.org/10.1016/j.jhydrol.2003.09.021.

Beven, K. (1983). Surface water hydrology—runoff generation and basin structure. *Reviews of Geophysics and Space Physics, 21*(3), 721–730. https://doi.org/10.1029/RG021i003p00721.

Beven, K. (1989). Changing ideas in hydrology—the case of physically-based models. *Journal of Hydrology, 105*(1–2), 157–172. https://doi.org/10.1016/0022-1694(89)90101-7.

Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources, 16,* 41–51.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology, 320,* 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007.

Beven, K. (2012). Rainfall-runoff modelling: The primer (2nd Ed.).

Beven, K., & Germann, P. (1982). Macropores and water flow in soils. *Water Resources Research, 18*(5), 1311–1325. https://doi.org/10.1029/WR018i005p01311.

Beven, K. J. (2000). Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences, 4*(2), 203–213. https://doi.org/10.5194/hess-4-203-2000.

Birkel, C., & Soulsby, C. (2015). Advancing tracer-aided rainfall-runoff modelling: A review of progress, problems and unrealised potential. *Hydrological Processes, 29*(25), 5227–5240. https://doi.org/10.1002/hyp.10594.

Birkel, C., Soulsby, C., & Tetzlaff, D. (2011). Modelling catchment-scale water storage dynamics: Reconciling dynamic storage with tracer-inferred passive storage. *Hydrological Processes, 25*(25), 3924–3936. https://doi.org/10.1002/hyp.8201.

Bixio, A., et al. (2002). Modeling groundwater-surface water interactions including effects of morphogenetic depressions in the Chernobyl exclusion zone. *Environmental Geology, 42*(2–3), 162–177. https://doi.org/10.1007/s00254-001-0486-7.

Blazkova, S., Beven, K., Tacheci, P., & Kulasova, A. (2002). Testing the distributed water table predictions of TOPMODEL (allowing for uncertainty in model calibration): The death of TOPMODEL? *Water Resources Research, 38*(11), 39-1–39–11. https://doi.org/10.1029/2001wr000912.

Brooks, E. S., Boll, J., & McDaniel, P. A. (2004). A hillslope-scale experiment to measure lateral saturated hydraulic conductivity. *Water Resources Research, 40*(W04208), 1–10. https://doi.org/10.1029/2003WR002858.

Brunner, P., & Simmons, C. T. (2012). Hydrogeosphere: A fully integrated, physically based hydrological model. *Ground Water, 50*(2), 170–176. https://doi.org/10.1111/j.1745-6584.2011.00882.x.

Bühler, Y., et al. (2015). Snow depth mapping in high-alpine catchments using digital photogrammetry. *Cryosphere, 9*(1), 229–243. https://doi.org/10.5194/tc-9-229-2015.

Camporese, M., Paniconi, C., Putti, M., & Orlandini, S. (2010). Surface-subsurface flow modeling with path-based runoff routing, boundary condition-based coupling, and assimilation of multisource observation data. *Water Resources Research, 46*(2). https://doi.org/10.1029/2008wr007536.

Clark, M. P., et al. (2017). The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrology and Earth System Sciences, 21*(7), 3427–3440. https://doi.org/10.5194/hess-21-3427-2017.

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., et al. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research, 52*(3), 2350–2365. https://doi.org/10.1002/2015WR017910.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., et al. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research, 48*(5), 1–17. https://doi.org/10.1029/2011WR011721.

Dakhlaoui, H., Ruelland, D., Tramblay, Y., & Bargaoui, Z. (2017). Evaluating the robustness of conceptual rainfall-runoff models under climate variability in Northern Tunisia. *Journal of Hydrology, 550,* 201–217. https://doi.org/10.1016/j.jhydrol.2017.04.032.

Davies, J., Beven, K., Nyberg, L., & Rodhe, A. (2011). A discrete particle representation of hillslope hydrology: Hypothesis testing in reproducing a tracer experiment at Gårdsjön, Sweden. *Hydrological Processes, 25*(23), 3602–3612. https://doi.org/10.1002/hyp.8085.

Detty, J. M., & McGuire, K. J. (2010). Topographic controls on shallow groundwater dynamics: Implications of hydrologic connectivity between hillslopes and riparian zones in a till mantled catchment. *Hydrological Processes, 24*(16), 2222–2236. https://doi.org/10.1002/hyp.7656.

Dong, J., Walker, J. P., Houser, P. R., & Sun, C. (2007). Scanning multichannel microwave radiometer snow water equivalent assimilation. *Journal of Geophysical Research, 112*(D7), D07108. https://doi.org/10.1029/2006JD007209.

Dunn, S. M., Bacon, J. R., Soulsby, C., Tetzlaff, D., Stutter, M. I., Waldron, S., et al. (2008a). Interpretation of homogeneity in d18O signatures of stream water in a nested sub-catchment system in North-East Scotland. *Hydrological Processes, 22,* 4767–4782.

Dunn, S. M., Freer, J., Weiler, M., Kirkby, M. J., Seibert, J., Quinn, P. F., et al. (2008b). Conceptualization in catchment modelling: Simply learning? *Hydrological Processes, 22*(13), 2389–2393. https://doi.org/10.1002/hyp.7070.

Dunne, T. (1983). Relation of field studies and modeling in the prediction of storm runoff. *Journal of Hydrology, 65*(1–3), 25–48. https://doi.org/10.1016/0022-1694(83)90209-3.

Durand, M., & Margulis, S. A. (2006). Feasibility test of multifrequency radiometric data assimilation to estimate snow water equivalent. *Journal of Hydrometeorology, 7*(3), 443–457. https://doi.org/10.1175/JHM502.1.

Ebel, B. A., & Loague, K. (2006). Physics-based hydrologic-response simulation: Seeing through the fog of equifinality. *Hydrological Processes, 20*(13), 2887–2900. https://doi.org/10.1002/hyp.6388.

Ebel, B. A., Loague, K., Vanderkwaak, J. E., Dietrich, W. E., Montgomery, D. R., Torres, R., et al. (2007). Near-surface hydrologic response for a steep, unchanneled catchment near Coos Bay, Oregon: 2. Physics-based simulations. *American Journal of Science, 307*(4), 709–748. https://doi.org/10.2475/04.2007.03.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences, 17*(5), 1893–1912. https://doi.org/10.5194/hess-17-1893-2013.

Fenicia, F., McDonnell, J. J., & Savenije, H. H. G. (2008). Learning from model improvement: On the contribution of complementary data to process understanding. *Water Resources Research, 44* (December 2007), 1–13. https://doi.org/10.1029/2007wr006386.

Finger, D., Vis, M. J. P., Huss, M., & Seibert, J. (2015). The value of multiple data set calibration versus model complexity for improving the performance of hydrological models in mountain catchments. *Water Resources Research, 51*(1), 1939–1958. https://doi.org/10.1002/2014WR016259.

Glaser, B., Klaus, J., Frei, S., Frentress, J., Pfister, L., & Hopp, L. (2016). On the value of surface saturated area dynamics mapped with thermal infrared imagery for modeling the hillslope-riparian-stream continuum. *Water Resources Research, 52*(10), 8317–8342. https://doi.org/10.1002/2015WR018414.

Grabs, T., Seibert, J., Bishop, K., & Laudon, H. (2009). Modeling spatial patterns of saturated areas: A comparison of the topographic wetness index and a dynamic distributed model. *Journal of Hydrology, 373*(1–2), 15–23. https://doi.org/10.1016/j.jhydrol.2009.03.031.

Grayson, R. B., & Blöschl, G. (eds.) (2001). *Spatial patterns in catchment hydrology: Observations and modelling*. CUP Archive.

Grayson, R. B., Moore, I. D., & McMahon, T. A. (1992). Physically based hydrologic modeling 2. Is the concept realistic. *Water Resources Research, 26*(10), 2659–2666.

Griessinger, N., Seibert, J., Magnusson, J., & Jonas, T. (2016). Assessing the benefit of snow data assimilation for runoff modeling in Alpine catchments. *Hydrology and Earth System Sciences, 20*(9), 3895–3905. https://doi.org/10.5194/hess-20-3895-2016.

de Grosbois, E., Hooper, R. P., & Christophersen, N. (1988). A multisignal automatic calibration methodology for hydrochemical models: A case study of the Birkenes Model. *Water Resources Research, 24*(8), 1299–1307. https://doi.org/10.1029/WR024i008p01299.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology, 377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes, 22,* 3802–3813. https://doi.org/10.1002/hyp.6989.

Hansen, A. L., Refsgaard, J. C., Christensen, B. S. B., & Jensen, K. H. (2013). Importance of including small-scale tile drain discharge in the calibration of a coupled groundwater-surface water catchment model. *Water Resources Research, 49*(1), 585–603. https://doi.org/10.1029/2011WR011783.

Hartmann, A., Goldscheider, N., Wagener, T., Lange, J., & Weiler, M. (2014). Karst water resources in a changing world: Review of hydrological modeling approaches. *Reviews of Geophysics, 52*(3), 218–242. https://doi.org/10.1002/2013RG000443.

Hartmann, A., et al. (2013). Process-based karst modelling to relate hydrodynamic and hydrochemical characteristics to system properties. *Hydrology and Earth System Sciences, 17*(8), 3305–3321. https://doi.org/10.5194/hess-17-3305-2013.

Haught, D. R. W., & van Meerveld, H. J. (2011). Spatial variation in transient water table responses: Differences between an upper and lower hillslope zone. *Hydrological Processes, 25*(25), 3866–3877. https://doi.org/10.1002/hyp.8354.

Hourdin, F, et al. (2016). The art and science of climate model tuning. *Bulletin of the American Meteorological Society (MARCH)*: BAMS-D-15-00135.1. https://doi.org/10.1175/bams-d-15-00135.1.

Hrachowitz, M., Savenije, H., Bogaard, T. A., Tetzlaff, D., & Soulsby, C. (2013). What can flux tracking teach us about water age distribution patterns and their temporal dynamics? *Hydrology and Earth System Sciences, 17*(2), 533–564. https://doi.org/10.5194/hess-17-533-2013.

Iorgulescu, I., Beven, K. J., & Musy, A. (2005). Data-based modelling of runoff and chemical tracer concentrations in the Haute-Mentue research catchment (Switzerland). *Hydrological Processes, 19*(13), 2557–2573. https://doi.org/10.1002/hyp.5731.

Ivanov, V. Y., Vivoni, E. R., Bras, R. L., & Entekhabi, D. (2004). Catchment hydrologic response with a fully distributed triangulated irregular network model. *Water Resources Research, 40*(11). https://doi.org/10.1029/2004wr003218.

James, L. D., & Burges, S. J. (1982). Selection, calibration, and testing of hydrologic models, Hydrologic Modeling of Small Watersheds (C Haan, H Johnson, and D Brakensiek, eds). American Society of Agricultural Engineers: St. Joseph, Mich.

Jones, J. P., Sudicky, E. A., Brookfield, A. E., & Park, Y. -J. (2006). An assessment of the tracer-based approach to quantifying groundwater contributions to streamflow. *Water Resources Research, 42*(2). https://doi.org/10.1029/2005wr004130.

Jones, J. P., Sudicky, E. A., & McLaren, R. G. (2008). Application of a fully-integrated surface-subsurface flow model at the watershed-scale: A case study. *Water Resources Research, 44*(3). https://doi.org/10.1029/2006wr005603.

Kachanoski, R. G., De Jong, E., & Van, Wesenbeeck I. J. (1990). Field scale patterns of soil water storage from non-contacting measurements of bulk electrical conductivity. *Canadian Journal of Soil Science, 70*(3), 537–542. https://doi.org/10.4141/cjss90-056.

Karlsen, R. H., Seibert, J., Grabs, T., Laudon, H., Blomkvist, P., & Bishop, K. (2016). The assumption of uniform specific discharge: Unsafe at any time? *Hydrological Processes, 30*(21), 3978–3988. https://doi.org/10.1002/hyp.10877.

Kikuchi, C. (2017). Toward increased use of data worth analyses in groundwater studies. *Groundwater, 55*(5), 670–673. https://doi.org/10.1111/gwat.12562.

Kirchner, J. W. (2006a). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research, 42*(3), 1–5. https://doi.org/10.1029/2005WR004362.

Kirchner, J. W. (2006b). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research, 42*(3), W03S04 https://doi.org/10.1029/2005wr004362.

Kirchner, J. W., & Neal, C. (2013). Universal fractal scaling in stream chemistry and its implications for solute transport and water quality trend detection. *Proceedings of the National Academy of Sciences, 110*(30), 12213–12218. https://doi.org/10.1073/pnas.1304328110.

Klemeš, V. (1983). Conceptualization and scale in hydology. *Journal of Hydrology, 65,* 1–23.

Klemeš, V. (1986a). Operational testing of hydrological simulation models. *Hydrological Sciences Journal, 31*(1), 13–24. https://doi.org/10.1080/02626668609491024.

Klemeš, V. (1986b). Dilettantism in hydrology: Transition or destiny? *Water Resources Research, 22*(9 S), 177S–188S. https://doi.org/10.1029/wr022i09sp0177s.

Klemeš, V. (1997). Guest editorial: Of carts and horses in hydrologic modeling. *Journal of Hydrologic Engineering, 2*(2), 43–49. https://doi.org/10.1061/(ASCE)1084-0699(1997)2:2(43).

Koboltschnig, G. R., Schöner, W., Zappa, M., Kroisleitner, C., & Holzmann, H. (2008). Runoff modelling of the glacierized Alpine Upper Salzach Basin (Austria): Multi-criteria result validation. *Hydrological Processes, 22*(19), 3950–3964. https://doi.org/10.1002/hyp.7112.

Krause, P., & Boyle, D. P. (2005). Advances in geosciences comparison of different efficiency criteria for hydrological model assessment. *Advances In Geosciences, 5*(89), 89–97. https://doi.org/10.5194/adgeo-5-89-2005.

Kuczera, G., & Mroczkowski, M. (1998). Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. *Water Resources Research, 34*(6), 1481–1489. https://doi.org/10.1029/98WR00496.

Laudan, L. (1990). Demystifying underdetermination. In C. W. Savage (Ed.), *Scientific theories. Minnesota studies in the philosophy of science* (pp. 267–297). Minneapolis: University of Minnesota Press. https://doi.org/10.1080/03634528709378635.

Leaf, A. T. (2017). Using models to identify the best data: An example from Northern Wisconsin. *Groundwater, 55*(5), 641–645. https://doi.org/10.1111/gwat.12561.

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research, 99*(D7), 14415. https://doi.org/10.1029/94JD00483.

Lindström, G. (1997). A simple automatic calibration routine for the HBV model. *Nordic Hydrology, 28*(3), 153–168.

Loague, K., & Ebel, B. A. (2016). Finite-element modelling of physics-based hillslope hydrology, Keith Beven, and beyond. *Hydrological Processes, 30*(14), 2432–2437. https://doi.org/10.1002/hyp.10762.

Loague, K., & VanderKwaak, J. E. (2002). Simulating hydrological response for the R-5 catchment: Comparison of two models and the impact of the roads. *Hydrological Processes, 16*(5), 1015–1032. https://doi.org/10.1002/hyp.316.

Loague, K., & VanderKwaak, J. E. (2004a). Physics-based hydrologic response simulation: Platinum bridge, 1958 Edsel, or useful tool. *Hydrological Processes, 18*(15), 2949–2956. https://doi.org/10.1002/hyp.5737.

Loague, K., & VanderKwaak, J. E. (2004b). Physics-based hydrologic response simulation: Platinum bridge, 1958 Edsel, or useful tool. *Hydrological Processes, 18*(15), 2949–2956. https://doi.org/10.1002/hyp.5737.

Magnusson, J., Gustafsson, D., Hüsler, F., & Jonas, T. (2014). Assimilation of point SWE data into a distributed snow cover model comparing two contrasting methods. *Water Resources Research, 50*(10), 7816–7835. https://doi.org/10.1002/2014WR015302.

Magnusson, J., Wever, N., Essery, R., Helbig, N., Winstral, A., & Jonas, T. (2015). Evaluating snow models with varying process representations for hydrological applications. *Water Resources Research, 51*(4), 2707–2723. https://doi.org/10.1002/2014WR016498.

Martinez-Landa, L., & Carrera, J. (2005). An analysis of hydraulic conductivity scale effects in granite (Full-scale Engineered Barrier Experiment (FEBEX), Grimsel, Switzerland). *Water Resources Research, 41*(3), 1–13. https://doi.org/10.1029/2004WR003458.

McGuire, K. J., Weiler, M., & McDonnell, J. J. (2007). Integrating tracer experiments with modeling to assess runoff processes and water transit times. *Advances in Water Resources, 30*(4), 824–837. https://doi.org/10.1016/j.advwatres.2006.07.004.

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes, 26*(26), 4078–4111. https://doi.org/10.1002/hyp.9384.

Milzow, C., Krogh, P. E., & Bauer-Gottwein, P. (2011). Combining satellite radar altimetry, SAR surface soil moisture and GRACE total storage changes for hydrological model calibration in a large poorly gauged catchment. *Hydrology and Earth System Sciences, 15*(6), 1729–1743. https://doi.org/10.5194/hess-15-1729-2011.

Mirus, B. B., Ebel, B. A., Heppner, C. S., Loague, K. (2011). Assessing the detail needed to capture rainfall-runoff dynamics with physics-based hydrologic response simulation. *Water Resources Research 47*(3). https://doi.org/10.1029/2010wr009906.

Mirus, B. B., Loague, K., VanderKwaak, J. E., Kampf, S. K., & Burges, S. J. (2009). A hypothetical reality of Tarrawarra-like hydrologic response. *Hydrological Processes, 23*(7), 1093–1103. https://doi.org/10.1002/hyp.7241.

Motovilov, Y. G., Gottschalk, L., Engeland, K., & Rodhe, A. (1999). Validation of a distributed hydrological model against spatial observations. *Agricultural and Forest Meteorology, 99,* 257–277.

Mroczkowski, M., Raper, P. G., & Kuczera, G. (1997). The quest for more powerful validation of conceptual catchment models. *Water Resources Research, 33*(10), 2325–2335. https://doi.org/10.1029/97WR01922.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I-a discussion of principles. *Journal of Hydrology, 10,* 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Parajka, J., & Blöschl, G. (2008). The value of MODIS snow cover data in validating and calibrating conceptual hydrologic models. *Journal of Hydrology, 358*(3–4), 240–258. https://doi.org/10.1016/j.jhydrol.2008.06.006.

Parkin, G., O'Donnell, G., Ewen, J., Bathurst, J. C., O'Connell, P. E., & Lavabre, J. (1996). Validation of catchment models for predicting land-use and climate change impacts. 2. Case study for a Mediterranean catchment. *Journal of Hydrology, 175*(1–4), 595–613. https://doi.org/10.1016/s0022-1694(96)80027-8.

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology, 279*(1–4), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7.

Pool, S., Vis, M. J. P., Knight, R. R., & Seibert, J. (2017). Streamflow characteristics from modeled runoff time series – importance of calibration criteria selection. *Hydrology and Earth System Sciences, 21*(11), 5443–5457.

Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: Towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal, 63*, 13–14, 1941–1953. https://doi.org/10.1080/02626667.2018.1552002.

Refsgaard, J. C. (1997). Parameterisation, calibration and validation of distributed hydrological models. *Journal of Hydrology, 198,* 69–97.

Refsgaard, J. C., & Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. *Water Resources Research, 32*(7), 2189–2202. https://doi.org/10.1029/96WR00896.

Remondi, F., Kirchner, J. W., Burlando, P., & Fatichi, S. (2018). Water flux tracking with a distributed hydrological model to quantify controls on the spatiotemporal variability of transit time distributions. *Water Resources Research,* 3081–3099. https://doi.org/10.1002/2017wr021689.

Rivera Villarreyes, C. A., Baroni, G., & Oswald, S. E. (2011). Integral quantification of seasonal soil moisture changes in farmland by cosmic-ray neutrons. *Hydrology and Earth System Sciences, 15*(12), 3843–3859. https://doi.org/10.5194/hess-15-3843-2011.

Robinson, D. A., Campbell, C. S., Hopmans, J. W., Hornbuckle, B. K., Jones, S. B., Knight, R., et al. (2008). Soil moisture measurement for ecological and hydrological watershed-scale observatories: A review. *Vadose Zone Journal, 7*(1), 358. https://doi.org/10.2136/vzj2007.0143.

Schaefli, B., & Gupta, H. V. (2007). Do nash values have value? *Hydrological Processes, 21,* 2075–2080. https://doi.org/10.1002/hyp.

Schulze-Makuch, D., Carlson, D. A., Cherkauer, D. S., & Malik, P. (1999). Scale dependency of hydraulic conductivity in heterogeneous media. *Groundwater, 37*(6), 904–919. https://doi.org/10.1111/j.1745-6584.1999.tb01190.x.

Seibert, J. (2000). Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Science, 4*(2), 215–224. https://doi.org/10.5194/hess-4-215-2000.

Seibert, J. (2001a). Comment on "On the calibration and verification of two-dimensional, distributed, Hortonian, continuous watershed models" by Sharika U. S. Senarath et al. *Water Resources Research 37*(12), 3393–3395. https://doi.org/10.1029/2000wr000017.

Seibert, J. (2001b). On the need for benchmarks in hydrological modelling. *Hydrological Processes, 15*(6), 1063–1064. https://doi.org/10.1002/hyp.446.

Seibert, J. (2003). Reliability of model predictions outside calibration conditions. *Hydrology Research, 34*(5), 477–492.

Seibert, J., & McDonnell, J. J. (2002). On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research, 38*(11), 23-1–23–14 https://doi.org/10.1029/2001wr000978.

Seibert, J., Rodhe, A., & Bishop, K. (2003). Simulating interactions between saturated and unsaturated storage in a conceptual runoff model. *Hydrological Processes, 17*(2), 379–390. https://doi.org/10.1002/hyp.1130.

Seyfried, M. S., & Wilcox, B. P. (1995). Scale and the nature of spatial variability: Field examples having implications for hydrologic modeling. *Water Resources Research, 31*(1), 173–184. https://doi.org/10.1029/94WR02025.

Sheets, R., & Hendrickx, J. M. H. (1995). Noninvasive soil water content measurement using electromagnetic induction. *Water Resources Research, 31*(10), 2401–2409.

Smerdon, B. D., Mendoza, C. A., & Devito, K. J. (2007). Simulations of fully coupled lake-groundwater exchange in a subhumid climate with an integrated hydrologic model. *Water Resources Research, 43*(1). https://doi.org/10.1029/2006wr005137.

Stadnyk, T. A., Delavau, C., Kouwen, N., & Edwards, T. W. D. (2013). Towards hydrological model calibration and validation: Simulation of stable water isotopes using the isoWATFLOOD model. *Hydrological Processes, 27*(25), 3791–3810. https://doi.org/10.1002/hyp.9695.

Staudinger, M., Stoelzle, M., Cochand, F., Seibert, J., Weiler, M., & Hunkeler, D. (in review). Your work is my boundary condition! Challenges and approaches for a closer collaboration between hydrologists and hydrogeologists. Revised version resubmitted to Journal of Hydrology.

Stephenson, G. R., & Freeze, R. A. (1974). Mathematical simulation of subsurface flow contributions to snowmelt runoff, Reynolds Creek Watershed Idaho. *Water Resources Research, 10*(2), 284–294. https://doi.org/10.1029/WR010i002p00284.

Tromp-van Meerveld, I., & Weiler, M. (2008). Hillslope dynamics modeled with increasing complexity. *Journal of Hydrology, 361*(1–2), 24–40. https://doi.org/10.1016/j.jhydrol.2008.07.019.

Turner, J. V., & Barnes, C. J. (1998). Modeling of isotope and hydrogeochemical responses in catchment hydrology. In C. Kendall & J. Mcdonnell (Eds.), *Isotope Tracers in Catchment Hydrology* (723–760). Elsevier.

Uhlenbrook, S., & Leibundgut, C. (2002). Process-oriented catchment modelling and multiple-response validation. *Hydrological Processes, 16*(2), 423–440. https://doi.org/10.1002/hyp.330.

Vaché, K. B., & McDonnell, J. J. (2006). A process-based rejectionist framework for evaluating catchment runoff model structure. *Water Resources Research, 42*(2), 1–15. https://doi.org/10.1029/2005WR004247.

Vaché, K. B., McDonnell, J. J., & Bolte, J. (2004). On the use of multiple criteria for a posteriori model rejection: Soft data to characterize model performance. *Geophysical Research Letters, 31*(21), 1–4. https://doi.org/10.1029/2004GL021577.

van Huijgevoort, M. H. J., Tetzlaff, D., Sutanudjaja, E. H., & Soulsby, C. (2016). Using high resolution tracer data to constrain water storage, flux and age estimates in a spatially distributed rainfall-runoff model. *Hydrological Processes, 30*(25), 4761–4778. https://doi.org/10.1002/hyp.10902.

Van Meerveld, H. J., Seibert, J., & Peters, N. E. (2015). Hhillslope–riparian-stream connectivity and flow directions at the panola mountain research watershed. *Hydrological Processes, 29*, 3556–3574. https://doi.org/10.1002/hyp.10508.

VanderKwaak, J. E., & Loague, K. (2001). Hydrologic-response simulations for the R-5 catchment with a comprehensive physics-based model. *Water Resources Research, 37*(4), 999–1013. https://doi.org/10.1029/2000WR900272.

Vis, M., Knight, R., Pool, S., Wolfe, W., & Seibert, J. (2015). Model calibration criteria for estimating ecological flow characteristics. *Water (Switzerland), 7*(5), 2358–2381. https://doi.org/10.3390/w7052358.

Viviroli, D., Mittelbach, H., Gurtz, J., & Weingartner, R. (2009). Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland—Part II: Parameter regionalisation and flood estimation results. *Journal of Hydrology, 377*(1–2), 208–225. https://doi.org/10.1016/j.jhydrol.2009.08.022.

von Freyberg, J., Studer, B., Rinderer, M., Kirchner, J. W. (2018). Studying catchment storm response using event- and pre-event-water volumes as fractions of precipitation rather than discharge. *Hydrology and Earth System Sciences, 22*, 5847–5865. https://doi.org/10.5194/hess-22-5847-2018.

Weiler, M. (2017). Macropores and preferential flow—a love-hate relationship. *Hydrological Processes, 31*(1), 15–19. https://doi.org/10.1002/hyp.11074.

Werth, S., Güntner, A., Petrovic, S., & Schmidt, R. (2009). Integration of GRACE mass variations into a global hydrological model. *Earth and Planetary Science Letters, 277*(1–2), 166–173. https://doi.org/10.1016/j.epsl.2008.10.021.

Western, A. W., Grayson, R. B., & Green, T. R. (1999a). The tarrawarra project: High resolution spatial measurement, modelling and analysis of soil moisture and hydrological response. *Hydrological Processes, 13*(5), 633–652. isi:000079622700002.

Western, A. W., Grayson, R. B., & Green, T. R. (1999b). The Tarrawarra project: High resolution spatial measurement, modelling and analysis of soil moisture and hydrological response. *Hydrological Processes, 13*(5), 633–652.

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research, 44*(9), 1–18. https://doi.org/10.1029/2007WR006716.

Zehe, E., et al. (2014). HESS Opinions: From response units to functional units: A thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments. *Hydrology and Earth System Sciences, 18*(11), 4635–4655. https://doi.org/10.5194/hess-18-4635-2014.

Chapter 34
Uncertainty Quantification Using Multiple Models—Prospects and Challenges

Reto Knutti, Christoph Baumberger and Gertrude Hirsch Hadorn

Abstract Model evaluation for long-term climate predictions must be done on quantities other than the actual prediction, and a comprehensive uncertainty quantification is impossible. An ad hoc alternative is provided by coordinated model intercomparisons which typically use a "one model one vote" approach. The problem with such an approach is that it treats all models as independent and equally plausible. Reweighting all models of the ensemble for performance and dependence seems like an obvious way to improve on model democracy, yet there are open questions on what constitutes a "good" model, how to define dependency, how to interpret robustness, and how to incorporate background knowledge. Understanding those issues have the potential to increase confidence in model predictions in modeling efforts outside of climate science where similar challenges exist.

Keywords Ensemble modeling · Idealization · Model independence · Robustness · Structural model uncertainty · Uncertainty quantification

34.1 Introduction

Whether conceptual, analytical, or numerical, a model is usually an idealization, i.e., a simplified representation of a target system. A model represents certain elements or processes in order to reproduce or understand the characteristic behavior of a system, to test a hypothesis, or to predict target system quantities of interest that cannot be measured. Often, there are practical limitations that determine the complexity of a model, like the availability of data, computational cost, or even the lack of understanding of some processes that are deemed relevant. What is part of a model and what is not, and how it is represented, is driven by the purpose of the model, i.e., the research question in hand. Therefore, there is not only one possible model of one target, but there are many. The benefit of picking another model, or success of

R. Knutti (✉) · C. Baumberger · G. Hirsch Hadorn
ETH Zurich, Zürich, Switzerland
e-mail: reto.knutti@env.ethz.ch

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_34

changing the model (or lack thereof) can usually be quantified in terms of prediction skill. Thus, while an infinite number of model structures, boundary conditions, and parameter sets is possible in principle, in practice the decisions on how to further develop a model and whether to accept or reject a proposed change can often be made on a pragmatic basis: a change is likely to be implemented if it is more firmly rooted in theory and if it improves the skill, explanatory power, or usefulness of the model without compromising other desirable properties like efficiency. Improving the model may still be very challenging. But if the model can be evaluated by repeatedly testing its predictions (as, e.g., in the case of weather prediction models), this provides a clear feedback that guides model development. We distinguish model evaluation or validation as the determination of whether a model represents reality well enough for a particular purpose from verification as the determination of whether the output of a simulation approximates the true solutions to the differential equations of the original model. In what follows, we restrict ourselves to computer simulation models. Our focus is on model evaluation rather than on verification.

Model evaluation for long-term climate predictions cannot be based on repeated confirmation of the predictions against observation-based data. Moreover, model evaluation requires uncertainty estimation, ideally in quantitative terms. However, a comprehensive uncertainty quantification, which requires testing different assumptions in a model (i.e., variations in the structure, the processes included), exploring the uncertainty in parameter choices, and quantifying the effect of boundary conditions and datasets, is effectively impossible (see Sect. 34.2; for methods of uncertainty quantification in engineering contexts where repeated confirmation is possible; see Chap. 22 by Dougherty, Dalton and Dehghannasiri in this volume). As an ad hoc alternative, the climate modeling community has therefore started to establish coordinated model intercomparisons. The resulting ensembles of different models can be used to explore uncertainties either by testing the robustness of projections or as a basis for statistical methods that estimate the uncertainty about future climate change. A model projection is usually called robust if it is simulated by most models in the ensemble (although that does not imply that it is accurate). The notion of robustness is more generally used in the sciences to characterize the invariance of a result under multiple independent determinations, be these multiple different modeling approaches or, e.g., diverse experimental devices and measurement practices (Woodward 2006; Wimsatt 2012).

Here we use climate modeling to illustrate a few major (and possibly unique) challenges of determining the robustness of simulation results and estimating their uncertainty (for a general view on validation in climate science see also Chap. 30 by Rood in this volume). These challenges include definitions of core concepts, requirements for ensembles, and metrics for robustness that would support inferences from the robustness of projections, e.g., to warranted confidence in the projections. The challenges are interesting from both a philosophical and a practical point of view. Understanding these issues and finding smarter ways to deal with the resulting plurality of models has the potential to increase the value of models for climate as well as for other environmental areas, and potentially beyond. Eventually, this may increase the confidence we can have in such models as epistemic tools and provide scientists with a clearer explanation of what they are doing, and stronger arguments when it does or does not work.

We first discuss some peculiarities of climate modeling which make a comprehensive uncertainty quantification impossible (Sect. 34.2). We then distinguish between different sources of uncertainty in predicting climate change in order to better understand the motivation of using model ensembles as a means of estimating uncertainties in climate predictions (Sect. 34.3). The usual "one model one vote" approach problematically assumes that all models are independent and equally plausible (Sect. 34.4). As a way to improve on this model democracy, we suggest reweighting all models of an ensemble for performance and dependence (Sect. 34.5), and illustrate the idea for the case of Arctic sea ice (Sect. 34.6). We discuss some open issues, such as whether better agreement with observation reduces uncertainties in predictions, how to define model dependence, and how to incorporate background knowledge in the suggested weighting scheme (Sect. 34.7), and close with a short conclusion (Sect. 34.8).

34.2 Challenges for Uncertainty Quantification in Climate Modeling

Climate and Earth system models of various complexity are used to simulate the statistics of weather and how these will change in the future as a result of the emission of greenhouse gases like carbon dioxide and other radiatively active species (Claussen et al. 2002; Knutti 2008; Flato 2011). The problem of using such models for simulations has several peculiarities.

The first peculiarity relates to the system's many dimensions: simulating the weather in principle requires resolving the atmosphere, ocean, ice, and land surface of the Earth, because of the many processes and timescales that affect weather. From the condensation of water on a tiny aerosol (on spatial scales of micrometers and timescales of fractions of a second) to the large-scale ocean circulations and melting of ice sheets (extending over thousands of kilometers and thousands of years), the processes involved occur over at least twelve orders of magnitudes in both time and space. And from soil microbes that potentially affect the growth of a tree and its effect on the local carbon and water cycle to complex chemistry affecting cloud formation, from subglacial hydrology to volcanoes affecting the radiative balance in the stratosphere, from our technological progress in developing renewable energy sources to policy instruments that affect the rate of decarbonization, the list of (potentially) relevant processes that affect future climate is extremely long. The challenge consists of nothing less than simulating the whole Earth including human behavior, which by construction is impossible; and even if it were possible, it would not be reasonable. Due to the interactions of the many aspects in the climate system, an increase in complexity typically decreases the analytic understanding of a model

(Lenhard and Winsberg 2010). However, deciding on what to include and exclude, and how to simplify, is tricky.

The second peculiarity, partly a consequence of the first, is that it is prohibitively expensive to build a new model for each research question. The expertise and effort required imply that a big institution typically builds only one or two (often similar) versions of a model every few years. The same model is then used to study literally hundreds of different questions. Thus, rather than a specific purpose guiding model construction, we observe that it is the model, once it is built, that determines what purposes it can be used for. The third peculiarity, also a consequence of the first, is the computational cost and volume of data involved. A climate simulation typically takes days to months running on hundreds to thousands of processors of a supercomputer, which makes it prohibitively expensive to systematically optimize the dozens of parameters it has, or try hundreds of ideas before converging on a new model. Development is therefore strongly guided by experts' understanding of what could work, based on background knowledge and experience of what ideas have worked in similar situations or in other models in the past (Held 2005).

The fourth peculiarity is that a direct confirmation of the actual prediction is often impossible. To confirm the prediction of climate in the year 2100, one would have to wait for nearly a century, and even then a single confirmation would not be sufficient given the chaotic component of atmospheric variability. The development cycle of a model is usually much shorter than the typical timescales for confirmation. Model evaluation for long-term predictions therefore must be done on quantities other than the actual long-range prediction, e.g., observations of current climate (Gleckler et al. 2008; Knutti 2008; Flato 2011; Schaller et al. 2011), its variability, past changes, or paleoclimate data (Harrison et al. 2015). The question then becomes which quantities matter most for what question (see Sect. 34.5).

The mentioned peculiarities make it practically impossible to test many different assumptions in a model (i.e., variations in the structure and the processes included), different choices for parameters, and to quantify the effects of boundary conditions and datasets in a systematic way. However, such a systematic assessment would be required for comprehensive quantification of uncertainties. Coordinated model intercomparisons offer an ad hoc work-around to this problem. Such efforts were started by the climate modeling community about two decades ago. They require whoever is willing to contribute to perform standardized simulations and provide the results to others for analysis. The resulting ensembles of different models are often referred to as "ensembles of opportunity", since they group together existing models and are not designed to span an uncertainty range (Knutti 2010a; Knutti et al. 2010; Eyring et al. 2016).

34.3 Uncertainty Quantification Using Model Ensembles

To better understand the motivation of using ensembles of different models, it is useful to characterize the sources of uncertainty in predicting climate change. Three sources of uncertainty can be distinguished: natural variability (both internal to the system and externally forced from changes in solar irradiance and volcanic eruptions), scenario uncertainty and model uncertainty. Natural internal variability is an inherent property resulting from the chaotic nature of the ocean–atmosphere system. We cannot predict the weather more than about a week in advance, because tiny uncertainties in the initial conditions grow as we run the model forward in time. The system is sensitive to its initial conditions, much like a Lorenz system with multiple attractors. That does *not* imply that the system is fundamentally unpredictable; the models indicate that some aspects like the temperature difference between winter and summer or the long-term trend resulting from increased CO_2 in the atmosphere are predictable, although bifurcations may exist in parts of the system, e.g., the Atlantic meridional overturning circulation (Lenton et al. 2008). Climate, the distribution of all weather states, therefore is very likely predictable, but the individual sequence of weather events, is not (Deser et al. 2012). This uncertainty, often referred to as ontic uncertainty because it is due to the chaotic nature of the target system, is largely accounted for by making statements about the climate averaged over 20 or more years. Hence, it is not fundamentally impossible to deal with this variability, but it is challenging because we can only evaluate the model in a probabilistic sense (i.e., by comparing distributions), and single events are of little value for judging the adequacy of a model. The second source of uncertainty, scenario uncertainty, results from uncertainty in emissions of anthropogenic forcings like CO_2, methane, -SO_2, or ozone. These are driven by technological progress, climate policy, values in society, wars, etc., all of which are difficult to predict because they are based on human behavior. This is also an ontic uncertainty, due to inherent properties of in this case socio-techno-economic systems. This uncertainty is often accounted for by considering projections (as opposed to predictions), defined as the response of climate conditional on a predefined scenario of societal development (along with emissions, land use change, etc.) (Vuuren et al. 2011).

This leaves us with model uncertainty, which is an epistemic uncertainty, i.e., a lack of knowledge about whether the model is an appropriate representation of the target system in question. A model is a representation of reality that is necessarily simplified in important ways. First, some processes in the climate system are not fully understood, e.g., changes in complex ecosystems. Second, some are rather well understood but are so complex or small-scale that their effect has to be parameterized in a simple way as a function of available large-scale properties (Gent et al. 1995; McFarlane 2011), e.g., ocean mixing processes and transports occurring on scales smaller than the resolution of the model (typically 100 km). The corresponding parameters (e.g., an equivalent diffusivity) must be calibrated to match large-scale observations and have no analog measurable equivalent quantity in reality. Third, numerical approximations and finite resolution lead to small errors when integrating the equations. In principle, this could be improved by larger computers, but, in practice, every doubling of horizontal resolution requires about ten times more computing capacity, so it will take many decades before the relevant scales (tens to hundreds of meters) can be resolved in global simulations (Schneider et al. 2017). In addition, boundary conditions (like the bathymetry of the ocean or the structure and properties of the soil) at every location are not fully known.

As a consequence of all of the above, it is often said that climate models are uncertain, but this is a misconception. Strictly, a model, once it is specified in the form of equations or code, is perfectly certain, in the sense that applying the equations twice will give the exact same results, and the effect of any change in the equations can be quantified precisely. The uncertainty comes from the model being a representation of a target in the real world, which requires specification and inference steps, in deciding what to include in the model, and how to interpret the results of the models for the real world. Of course, every climate model is false, by construction, in the sense that it is an idealized representation of a real and open system (Oreskes et al. 1994). Not only does the model ignore some climate processes but it also distorts the represented processes in different ways in order to make them mathematically and computationally tractable. The question is not whether the model is true but whether it is “true enough” (Elgin 2017), i.e., how well it represents the real system, and how useful or adequate it is for learning about a particular aspect of the real system.

This last point, the adequacy of a model, motivates the pluralism in climate modeling: because of the complexity of the system, the computational cost, and the lack of direct confirmation of prediction, there is no single agreed-on “best” model. Scientists inevitably have to make choices in what to include, how to parameterize unresolved processes, and how to manage the tradeoff between complexity, resolution, the number of simulations and number of years to simulate. Since there is disagreement on how to make these choices, to some extent even for a given purpose, there is no consensus on which one is the “best” model. So while multiple models could be seen as ontologically incompatible (strictly speaking, they make conflicting assumptions about the real world), and one could argue that scientists have to assess how well they are supported by data, the community seems happy with the model pluralism. The models are seen as complementary in the sense that they are all plausible (although not necessarily equally plausible) representations of the real system given the incomplete knowledge, data, and computational constraints; they are used pragmatically to investigate uncertainties (Parker 2006, 2010, 2013).

The diversity of models across an ensemble provides one avenue to try to estimate the consequences of model uncertainty by testing the robustness of results. For example, there are several ways one can parameterize atmospheric small-scale convection, and it is helpful to test whether the model behavior depends on the structure of that parameterization and the parameter values. Robustness in a qualitative sense is often invoked as a premise in an argument to the effect that a model result can be trusted (see Parker 2011, for a critical discussion). Robustness analysis goes a step further, using robust results to confirm certain parts of a model. Robustness analysis was developed as a modeling strategy in population ecology (Levins 1966). It has been generalized and systematized (Weisberg 2006; Wimsatt 2012), and also been applied to climate science (Lloyd, 2009, 2010). Robustness analysis uses a robust result as confirmatory evidence for more general relations of a model, which are then called "robust theorems". Robust theorems have the form: "Ceteris paribus, if [common core (causal) structure] obtains, then [robust property] will obtain" (Weisberg 2006). For instance, if all models that share a core causal structure but use a variety of simplifications show that higher CO_2 concentrations lead to substantial warming, then that result is unlikely to be just a consequence of particular choices made in a model. This robust result is then used to formulate the robust theorem: "Ceteris paribus, if [Greenhouse gases relate in law-like interaction with the energy budget of the earth] obtains, then [increased global mean temperature] will obtain" (Lloyd, 2009, 2010). But there are of course limits to such an argument: there are cases where all models are known to be robustly wrong in the same way because they all ignore a process (e.g., ice sheet dynamics) or parameterize it in a similar way. In order to avoid being misled by the robustness of results that is, in fact, pseudo-robustness (Wimsatt 2012), models must be sufficiently diverse in the relevant regards. There is considerable controversy on how to specify this requirement. A typical way is to specify "diversity" as "independence" (Wimsatt 2012) and to elaborate on a formal account for explicating this concept, for instance in a Bayesian framework (Fitelson 2001; Lloyd 2010; Stegenga and Menon 2017). However, these approaches are not uncontested (Schupbach 2016), and their appropriate specification remains a challenge.

In our discussion, we focus on determining the robustness of simulation results used to estimate the uncertainty in long-term climate predictions, which needs to be distinguished from robustness analysis used to confirm certain parts of a model. For brevity, we will focus on the most interesting and challenging case of multiple structurally different models in the Coupled Model Intercomparison Projects CMIP (Eyring et al. 2016), noting that similar ideas can, of course, be applied to what is often called perturbed physics ensembles, a model run with a variety of parameter sets (Stainforth et al. 2005). Many issues are similar, except that a single model structure can only capture so much of the range of behavior: no parameter set of one model will ever behave (in all respects) like a structurally different model that resolves other processes, although parameter calibration can compensate for some missing aspects of processes.

34.4 Problems with Model Democracy

Ensembles of opportunity like CMIP are often used for uncertainty quantification in a naïve way: the average of all models is taken as a best estimate, and the spread of the models is reported as the uncertainty of the projection. This "one model one vote" or "model democracy" (Knutti 2010), often used based on a lack of more convincing or generally agreed-upon alternatives, makes several assumptions which are rarely explicitly stated and even less frequently defended by actual evidence. First, model democracy treats all models as reasonably independent, and second, it assumes that all models are about equally plausible. Third, it assumes that the range of model projections represents what we believe is the uncertainty in the projection. In a weather forecast, the equivalent would be a probabilistic projection that is neither too broad nor overconfident, so that for many trials, observed outcomes would fall within the estimated 5–95% confidence intervals in about 90% of the trials.

Unfortunately, none of the assumptions made by model democracy is strictly fulfilled by present-day model ensembles. On the first point of dependence: many models use ideas, parts of the code, or even whole components (e.g., the sea ice model) from other models. The sheer complexity and cost lead groups to merge their efforts in jointly developing or using components of other groups (Bellouin et al. 2011). New models are almost never developed from scratch but are based on earlier models (Edwards 2011). As a consequence, some models are not providing much additional information, and multiple replications of a model may strongly bias the result toward that particular model (Annan and Hargreaves 2011; Masson and Knutti 2011a; Pennell and Reichler 2011; Knutti et al. 2013). How to actually define model dependence is not straightforward (Annan and Hargreaves 2016). The models are of course dependent in the sense that they all describe the same system, but that is not the point: they are also similarly biased with regard to how they represent reality because they share structural limitations or simplify things in the same way, and therefore their projections will likely be biased in the same way. If two models share several parts, the success of one model in simulation results has implications for the probability of the other model's success. This leaves us with the question of how to explicate an appropriate notion of dependence and specify a metric to determine model dependence (see Sects. 34.6 and 34.7).

On the second assumption, some models clearly perform better than others in some metrics (for an introduction and overview on relevant metrics, see Chap. 18 by Saam in this volume), i.e., simulation results are closer to observations of reality, with differences of up to a factor of two (Knutti et al. 2013). Reductions in the biases by 20–30% from one model intercomparison to the next imply that some models are about a decade of model development ahead of others in terms of how well they reproduce the observations. No model is clearly far superior to all others, consistent with the idea of pluralism where all models are seen as plausible representations of reality given some practical boundary conditions; but some are more plausible in certain respects than others. Some models perform well on certain metrics while others perform well on others (Gleckler et al. 2008), which reflects different modeling groups' focus in terms of development and calibration. But a model that performs well on one metric also tends to perform well on many others for at least two reasons: the climate system is coupled, so a correct representation of rainfall, for example, requires humidity (and therefore temperature), the dynamics (weather patterns), and clouds to be well represented. The other fact is a practical one: some centers simply have more resources (people and computing power) and experience than others, and their models tend to do well on many criteria.

On the third assumption, the spread of model projections does not necessarily represent what we believe is the uncertainty in the prediction. The spread of the ensemble may be too big if the ensemble contains demonstrably unrealistic members that can be rejected upfront based on physical understanding or disagreement with observations (see Chap. 6 by Beven and Lane in this volume). A model of Venus or Mars, for example, is unlikely to provide a useful projection of climate for the Earth and should thus be excluded from the respective ensemble. The model spread can also be too small if all models are missing the same relevant thing and are biased in the same way. In many cases, we do not know whether the spread tends to be too large or too small, and that likely depends on the variable, the timescale and the spatial scale (Masson and Knutti 2011b).

A further complication is the question whether the ensemble of models is centered around the truth (the so-called "truth plus error" paradigm, in which every model simulation approximates the observations of reality with a random error), or whether the observations of reality and the models are drawn from the same distribution (the "indistinguishable" paradigm, in which truth is not necessarily in the center). The former implies that predictions would get ever more certain as more models are added (in much the same way as the estimate of the average fall speed of a rock gets more and more precise as we continue to measure the time for the rock to reach the ground, if the measurement errors are random) which is certainly not the case. But the average of all models often does perform better than any individual model, suggesting some truth-centeredness at least for the observations available. This interpretation however can also change from the past to the future. For projections, the indistinguishable paradigm appears to be the more plausible interpretation in most cases (i.e., reality has about the same likelihood to approximately match any of the model realizations, and it is not necessarily in the center of the distribution) (Annan and Hargreaves 2010; Sanderson and Knutti 2012).

34.5 Beyond Model Democracy

Reweighting the ensemble for performance and dependence seems like an obvious way to improve on model democracy: poor and duplicated models would be down weighted and models whose performances agree well with observations and are relatively independently developed would constitute stronger evidence. Yet the discussions around such methods have been controversial so far. One argument against weighting is the sensitivity of the results to the chosen metric and possible overconfidence: if we weight by something that is unrelated to the quantity of interest or dominated by variability, then there is a possibility that the result gets worse rather than better (Weigel et al. 2010), and we may not know whether it does get worse. However, this is really only an issue when the number of models is very small. For a large number of models, it would essentially converge to random weights which should not affect the results. Sometimes, there are also political sensitivities: it can be difficult to dismiss models from certain centers or countries in a coordinated modeling effort. The other main argument raised against model weighting is that there are many ways to do it and the lack of direct confirmation prevents us from testing which approach is optimal. Indeed we can define an infinite number of model performance metrics (measuring the agreement with data in some way, e.g., a root mean square difference to observations, or a spectrum, or conservation of properties), and arguing which performance metric is relevant for the quality of a model is challenging (Knutti et al. 2010b). Unlike in weather forecasting, for example, we cannot quantify skill by repeated confirmation. Many broad brush metrics (e.g., patterns of temperature or rainfall) in fact appear to be only weakly correlated to large-scale projections like global temperature across a set of models (Jun et al. 2008; Knutti et al. 2010a). The reasons for the lack of relationships can be a large structural uncertainty in the models, lack of observed trend due to large variability, or lack of observations. Another hypothesis is that most of the observed data have already been used in model development and evaluation, such that the current set of models can already be interpreted as a posterior conditional on the observations; as a consequence, using the same observations again would not add anything (Sanderson and Knutti 2012).

The argument of model weighting gets more convincing, we would argue, if we assess model quality in relation to a particular purpose (Parker 2009). The question of which model is "best" is ill-posed unless we agree on the task the model is used for. The answer depends on the task we are trying to accomplish, in much the same way as which car people would say is best depends on whether they try to go really fast, or drive off-road, or move furniture. Defining weights for predicting a certain variable X is easier both politically and scientifically. Politically because one model will get more weight for predicting X, and another one will get more weight for predicting a different variable Y, which is only natural as some groups focus their development more on X and others more on Y. Scientifically, it is easier to select processes and quantities that are relevant for predicting X: one can refer to background knowledge, i.e., knowledge of various kinds that are accepted in the scientific community about the factors that determine X. Such insight can come from process understanding, trends emerging from natural variability, detection, and attribution, or from so-called emergent constraints, which typically are strong relationships between an observable quantity and a prediction. Observing the former can provide a constraint on the latter. For example, the strength of the albedo feedback on a seasonal timescale is related to the albedo feedback on decadal timescales (Hall and Qu 2006); hence, e.g., models that lose Arctic sea ice faster in the past tend to lose it faster in the future (Boé et al. 2009; Mahlstein and Knutti 2012; Overland and Wang 2013; Notz and Stroeve 2016; Knutti et al. 2017). Not all such relationships are robust across a wide range of models and there is a danger of spurious correlation when testing a large number of predictors (Masson and Knutti 2013; Caldwell et al. 2014). But despite all difficulties, when relationships across models are well understood in terms of the underlying processes, they can provide guidance on which quantities to use for model weighting.

As an alternative to attaching weights to models, emergent constraints can also be used to define a relationship between the observable and the projection (usually through some form of regression across models). This relationship can then be used to estimate an observationally constrained projection that is relatively independent of the set of underlying models (Boé et al. 2009; Mahlstein and Knutti 2012; Cox et al. 2013). Other options are interpolations in a low-dimensional model space (Sanderson et al. 2015b) or Bayesian methods (Tebaldi et al. 2004). They all vary in their statistical methods but share the idea of deviating from model democracy by using observed evidence. Also, strategies that combine dynamic models with other types of models using data-driven methods (Mazzocchi and Pasini 2017) need to use observational data, which are unavailable for long-term predictions. Data-driven approaches are genetically independent from dynamic models and are using different modeling schemes and methodological approaches. They may fit observations better given enough degrees of freedom, but may still be biased when it comes to out-of-sample prediction.

34.6 Illustration of Model Weighting for Arctic Sea Ice

We illustrate the idea of combining projections from multiple models here for Arctic sea ice, by weighting models both for their performance relative to observations and for model dependence. The method is relatively straightforward in the sense that a single number is defined as a weight for each simulation (although the choices that need to be made are not trivial, as discussed below), and it has been used in various contexts (Sanderson et al. 2015a, b; Knutti et al. 2017; Sanderson et al. 2017). The example is taken from an earlier study by Knutti et al. (2017), and is chosen because the processes are relatively well understood, and the added value of using observations is immediately obvious: to estimate when the Arctic will likely be ice-free, the model should have about the right sea ice extent today, and about the right trend over the past decades. Sea ice loss in the past and the future is correlated across models (Boé et al. 2009; Mahlstein and Knutti 2012; Overland and Wang 2013; Notz and Stroeve 2016; Knutti et al. 2017), which is plausibly explained by some models having a stronger sea ice albedo feedback than others. Observed sea ice trends are therefore an obvious constraint. There are of course other methods to weight models (Abramowitz and Gupta 2008; Waugh and Eyring 2008; Boé et al. 2009; Massonnet et al. 2012; Abramowitz and Bishop 2015), but the method outlined here may be the most straightforward one to illustrate the concepts.

For M models in the ensemble, the weight w_i for model i is defined as

$$w_i = e^{-\frac{D_i^2}{\sigma_D^2}} \Bigg/ \left(1 + \sum_{j \neq i}^{M} e^{-\frac{S_{ij}^2}{\sigma_S^2}}\right) \tag{34.1}$$

The numerator weighs a simulation by the distance metric D_i of model i to observations (performance), while the denominator effectively takes into account how many times parts of a model are replicated based on S_{ij}, the distance metric between model i and model j, which informs about the dependence of the models in the ensemble. Both D_i and S_{ij} are evaluated here as root mean square differences of a series of variables, but different choices for the metric and the functional form of the weighting can be defended. The weights are scaled such that their sum over the whole ensemble equals one. The constants σ_D and σ_S determine how strongly the model performance and dependence (“similarity”) are weighted (see below). This weighting scheme fulfills two basic requirements: a model that is infinitely far from observations and does in no way represent the real Earth (very large D_i) gets zero weight. For a model with no close neighbors, the denominator equals one and has no effect. However, duplicating an otherwise independent model ($S_{ij} = 0$) leads to a denominator being equal to two: as a consequence the two duplicates each get half of the weight, and the result is unaffected by the duplication. Because initial condition members (multiple simulations of the same model with slightly different starting conditions) are very similar, they are effectively treated as near replicates, and all available simulations can be used in a straightforward way even if the number of initial condition ensemble members varies strongly between models.

The metrics D_i and S_{ij} give equal weight to the climatological mean hemispheric mean September Arctic sea ice extent (1980–2013), and its trend over the same period, gridded climatological mean surface air temperature for each month, and climatological mean gridded interannual variability of monthly surface air temperature, but the sensitivity of the results to the choice of variable is illustrated in the results.

The choice of σ_D and σ_S determines how close a model’s simulation results need to be to observations to be considered “good” (performance), and how close two models need to be in order to be considered “similar” (dependence), respectively. The choice of these parameters is not straightforward. A very small σ_D, for example, may lead to the total weight being concentrated on just one or two models, at the expense of the results’ robustness. A very large value, on the other hand, will result in the weighting having almost no effect. One way to inform the choice of these parameters is to use perfect model tests, i.e., sequentially treating one of the models as reality and using the others to predict its future. Confirmation is possible, in this case, and allows optimizing the parameters for maximum skill while ensuring that the predictions are not overconfident. However, if models are similarly wrong then the perfect model tests might suggest that the method works well even if it does not in the real world. As such, perfect model tests are a necessary but not sufficient step for informing the choice of these parameters and to demonstrate the skill of the proposed method. A more in-depth discussion is provided by Knutti et al. (2017).

Weighting models can be done straightforwardly based on Eq. (34.1), but a number of choices with regards to variables, regions, time periods, and parameters are important. Hence, the results’ sensitivity toward these assumptions needs to be tested, and background knowledge is required to judge which choices are plausible. If clear constraints exist from observations, then the weighting makes the models more consistent with the past and narrows the model spread of the projection. In the following, we present an example of how taking the observations into account can improve the projections relative to a model democracy case.

Figure 34.1 shows the simulated September Arctic temperature and sea ice extent for all available fully coupled climate models (i.e., structurally different models as well as multiple initial condition members). Colors from gray to yellow to red indicate increasingly higher weights. The blue line indicates observations. Weighting is not based on the time series only, but on how well the models simulated the whole Arctic climate (see figure caption for details). Figure 34.1c indicates that the observationally weighted projection range (red) is substantially narrower than the raw range, i.e., the model democracy case (gray), and agreement with the observed trends is better. Note that we would not expect perfect agreement, as the observations represent one single realization whereas the weighted model average is closer to a forced response with much weaker variability. In the case of Arctic sea ice, there is evidence that part of the strong ice loss might be due to natural variability (Kay et al. 2011; Swart et al. 2015; Screen and Francis 2016). This would be consistent with the observed decline in sea ice being steeper than the weighted model average.

34.7 Discussion and Open Issues

As we argued in recent articles (Knutti 2010; Knutti et al. 2017), model democracy is increasingly hard to justify for climate model projections. Biases in some models and variables are so large that they cannot be ignored; in the example of Arctic sea ice discussed above, a model without sea ice in the present day or one with more sea ice by 2100 than observed today would be challenging to deal with. Simple bias correction methods that consider anomalies from a reference state will not work well or at all in such cases, as the change will depend on the reference state (if no sea ice is left, the change will also be zero). So if there are observations or other sources of information that can inform, or even better narrow the range of plausible projections, it would be strange not to use them.

In our view, there are essentially three points that need to be considered: performance as measured by agreement with observed data, model dependence, and background knowledge. In the case of dynamical models (as opposed to statistical models that are fitted), good agreement with a variety of observations provides strong evidence that the models are doing the relevant things correctly, but is not a formal proof of course (Baumberger et al. 2017). While confidence in the results should be larger when they are obtained by models that reproduce relevant aspects of current climate more accurately, performance alone provides insufficient support for long-term predictions. Furthermore, if the processes likely relevant for specific projections are sufficiently well understood and captured in the models, the coherence of models with this background knowledge provides an additional reason that increases our confidence in a projection (Baumberger et al. 2017). Given the complexity of the system, a model never agrees with all the data, but that is not required. The question is whether the model provides insight that we would not have otherwise. But how do we deal with a situation where improving the model based on process understanding, either through a more physical representation of a process, through increased resolution, or by explicitly resolving a process that has been prescribed or ignored before, leads to poorer agreement with data? Such situations are not uncommon, and can result from observation biases or from compensating errors in the models. From an understanding point of view, we might trust the new model more than the old one, and further development might improve the agreement again. Yet in an operational setting where users depend on predictions, a lower skill is hard to justify. Even in a

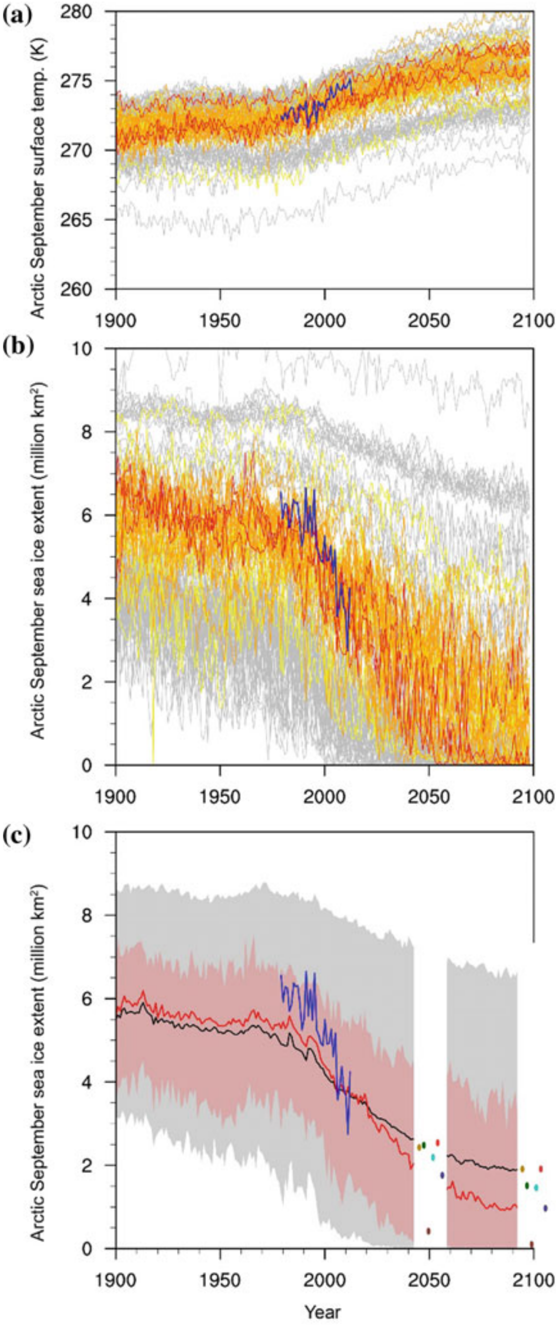


Fig. 34.1 **a** Arctic (60–90°N) September surface air temperature, **b** Arctic September sea ice extent in all CMIP3/5 simulations. Gray, yellow, orange and red indicates those that get <0.5%, >0.5%, >1%, and >5% weight, respectively, from weighting with Eq. (34.1). Observations (NCEP) are shown in blue. **c** Mean and 5–95% range for no weighting (black line, gray band) and weighting (red line and band). Colored dots near 2050 and 2100 show 2046–2055 and 2090–2099 average sea ice extent using (from left to right) the following metrics: (1) none (unweighted), (2) climatological mean (1980–2013) September sea ice extent, (3) September sea ice extent trend 1980–2013, (4) climatology of monthly surface temperature (1980–2013), (5) interannual variability of monthly surface temperature, (6) all 2–5. Figure reproduced from Knutti et al. (2017)

research context there is a tendency for a “dog and pony show”: the argument that it “looks good” is easier to sell than the fact that the underlying processes are more realistically described. This, of course, raises interesting discussion about the value of fit, and calibration (“tuning”) (Baumberger et al. 2017; Knutti 2018).

It is important to keep in mind that better agreement with observations will not necessarily reduce uncertainties in projections (Knutti and Sedláček 2012). But even in cases where it does not (Sanderson et al. 2017), we should not conclude that the effort was useless. This inability to further constrain the model range can arise either because the spread was not sufficient to begin with, or because the ensemble was already weighted due to good models being replicated a lot (Sanderson et al. 2017), or because the observations are not long enough or of sufficient quality or have too much variability to provide a constraint, or because the quantity of interest is inherently unpredictable, or because we have already used most of the information in the model development, evaluation and calibration. But in any case, we would not know until we have actually done the exercise. If the posterior after weighting is similar to the prior, then we have not reduced the spread, but we can be confident that the projection is reasonably consistent (in both magnitude and spread) with the observations we have on mean and trends. The raw model spread is just a range across models and cannot be interpreted as an uncertainty. It is an ad hoc measure of spread reflecting the ensemble design, or lack thereof, whereas the weighted results can be interpreted as an incomplete measure of uncertainty given all observations we have. The numbers may be similar, but the interpretation of the range is very different, and we should have more confidence in the latter.

Stronger constraints will come in the future (and have already in the past) from better observing systems specifically designed for climate change (early observations were mostly taken for weather prediction where long-term stability of a system was less of a concern), and from anthropogenic trends. Often past trends are more strongly related to future trends in a model than the mean state is related to future trends. But past forced trends may have been amplified or masked by natural variability, in particular over shorter periods (Deser et al. 2012; Fischer and Knutti 2016; Saffioti et al. 2016; Medhaug et al. 2017). Given the strong limits of available observations and computational capacity, model development and evaluation will, therefore, be a continuous process, and uncertainty estimates of projections will continue to change, as is the case in most other research areas. The lack of direct confirmation and the reliance on multiple potentially strongly dependent models however is somewhat unique to climate projections.

Model performance is an issue that any model developer always considers. In contrast, the issue of model dependence has gotten far less attention in the climate community. It is something that only becomes apparent after the various institutions have finalized their models. Only the most recent intercomparisons provided clear evidence that this problem can no longer be ignored, and there is less of a consensus on how to deal with it. It is likely to get more pronounced as model development gets increasingly complex and expensive. People sharing ideas or code, or developing code in a collaborative way is perfectly fine, but its impact on projections has to be considered in the interpretation of the results. In Sect. 34.6, we proposed a straightforward way how to include model dependence as a term in the weighting.

An open issue is a proper mathematical definition of model dependence that can actually be implemented in practice (Annan and Hargreaves 2016). Models' resembling each other by sharing certain parts or features is an indication for them being related, but once the simulation results of two models converge to observations, the simulation results of the various models will also get closer and closer to each other without the models necessarily being dependent. Furthermore, models that are independent from others may be irrelevant for the hypothesis in question. Because of these basic problems, it has also been questioned whether a concept of dependence is appropriate to explicate the diversity of models or other methods for reliably determining the robustness of their results (Schupbach 2016). More pragmatic concerns with applying a formal concept of probabilistic dependence in the case of climate models are for example how to find out which processes are represented in which way in different models. In most cases of using ensembles for determining the robustness of simulation results, these concerns are not an issue right now, because the distance of models' results to observations is typically far bigger than the distance between two strongly related models. Dependence and performance are treated independently in the example in Sect. 34.6, but further work may come up with different or more sophisticated alternatives.

Another open issue is an adequate selection process for ensemble members that avoids both pseudo-robustness resulting from excluding relevant plausible but diverging models (too narrow spread of results, e.g., because few centers in CMIP try to develop models with extreme behavior) and lack of robustness resulting from including irrelevant models (too broad spread of results). Which models are relevant depends on the hypothesis (purpose) for which the ensemble is used, which needs to be assessed by reference to relevant background knowledge about the problem in question and experiences with modeling practices. While this is a question that cannot be answered in general, making the considerations on the relevance of models explicit in each case would be a general requirement on using ensembles to determine the robustness of predictions. Scientists, e.g., often implicitly consider background knowledge when selecting an ensemble, but these considerations should be made explicit.

Background knowledge is important for considering whether to exclude or downweight models which violate basic physical principles (such as conservation of water or energy), or which lack representations of processes or feedbacks that are known to play an important role for future climate. In general terms: If the models within an ensemble differ strongly in how coherent they are with background knowledge, and if it is likely that there is a correlation between how well a model is based on process understanding and the model's adequacy for long-term projections, then the coherence with background knowledge should be considered in weighting the models for estimating uncertainties in such projections. It is important to see why coherence with background knowledge cannot be built into the dimension of performance: If two models reproduce equally well observed mean climate and trends but we know from background theories that only the first represents certain feedbacks (e.g., greenhouse gas emissions from thawing permafrost) which significantly influence future climate, then the first model should be given more weight than the second. On the one hand, one could think of coherence with relevant background knowledge as a consideration additional to determining the robustness of results (Parker 2013), e.g., for determining which models to include in an ensemble in the first place. On the other hand, one might think about integrating coherence with the relevant background knowledge as a further term in the weighting. In a Bayesian framework, the first option affects the prior, which is based on the whole ensemble; the second option affects the posterior, which depends on the weighing of the models. However, there is still considerable work to do in order to find a qualitative or a quantitative way to consider coherence with relevant background knowledge. It needs, e.g., to be determined how to deal with the intransparency of what exactly is in the models, and with limitations in the state of knowledge. Moreover, a procedure for assessing this coherence, e.g., something like eliciting expert judgments, needs to be established. Accounting for coherence with relevant background knowledge is a challenging task, but it needs to be addressed in order to improve the epistemic significance of robust results.

34.8 Conclusion

We have used climate modeling to illustrate a few major (and possibly unique) challenges of determining the robustness of simulation results for long-term predictions and of estimating their uncertainty. We have proposed to weight the models of an ensemble in order to avoid biases that result when all models are treated equally. We have proposed a somewhat ad hoc scheme that considers dependence and performance of the models, yet there are challenges that need further work. These include how to quantitatively account for coherence with background knowledge as a further important requirement on ensembles, as well as definitions of core concepts and metrics in order to provide a quantitative determination of the robustness of simulation results. Such an explicit and systematic approach to robustness of results is required to support inferences from the robustness of projections and to establish confidence in the projections. These challenges are interesting from both a philosophical and a practical point of view. Improving our understanding of these issues and finding better ways to deal with the plurality of models has the potential to increase the value of models not just for climate but other environmental areas, and potentially beyond, where determining the robustness of results is a strategy to assess confidence in results. Eventually, this may provide scientists with a clearer explanation of what they are doing in modeling, and stronger arguments about when modeling as an epistemic tool does or does not work.

Acknowledgements We thank Claus Beisbart, Keith Beven, Benedikt Knüsel, Wendy Parker, and Marius Wälchli for discussions and comments which have improved the quality of this manuscript. This project was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 641816 (CRESCENDO). We acknowledge the World Climate Research

Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led the development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

Abramowitz, G., & Gupta, H. (2008). Toward a model space and model independence metric. *Geophysical Research Letters, 35*(5), 1–4. https://doi.org/10.1029/2007GL032834.

Abramowitz, G., & Bishop, C. H. (2015). Climate model dependence and the ensemble dependence transformation of CMIP projections. *Journal of Climate, 28,* 2332–2348. https://doi.org/10.1175/JCLI-D-14-00364.1.

Annan, J. D., & Hargreaves, J. C. (2010). Reliability of the CMIP3 ensemble. *Geophysical Research Letters, 37*(2), 1–5. https://doi.org/10.1029/2009GL041994.

Annan, J. D., & Hargreaves, J. C. (2011). Understanding the CMIP3 multimodel ensemble. *Journal of Climate, 24*(16), 4529–4538. https://doi.org/10.1175/2011JCLI3873.1.

Annan, J., & Hargreaves, J. (2016). On the meaning of independence in climate science. *Earth System Dynamics Discussions*, 1–17. https://doi.org/10.5194/esd-2016-34.

Baumberger, C., Knutti, R., & Hirsch Hadorn, G. (2017). Building confidence in climate model projections: An analysis of inferences from fit. *Wiley Interdisciplinary Reviews: Climate Change, 8*(3), e454. https://doi.org/10.1002/wcc.454.

Bellouin, N., et al. (2011). The HadGEM2 family of Met Office Unified Model climate configurations. *Geoscientific Model Development*, *4*(3), 723–757. https://doi.org/10.5194/gmd-4-723-2011.

Boé, J., Hall, A., & Qu, X. (2009). September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nature Geoscience*, *2*(5), 341–343. (Nature Publishing Group). https://doi.org/10.1038/ngeo467.

Caldwell, P. M., et al. (2014). Statistical significance of climate sensitivity predictors obtained by data mining. *Geophysical Research Letters, 41*(5), 1803–1808. https://doi.org/10.1002/2014GL059205.

Claussen, M., 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. https://doi.org/10.1007/s00382-001-0200-1.

Cox, P. M., et al. (2013). Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature*, *494*(7437), 341–344. (Nature Publishing Group). https://doi.org/10.1038/nature11882.

Deser, C., et al. (2012). Communication of the role of natural variability in future North American climate. *Nature Climate Change, 2*(11), 775–779. https://doi.org/10.1038/nclimate1562.

Edwards, P. N. (2011). History of climate modeling. *Wiley Interdisciplinary Reviews: Climate Change, 2*(1), 128–139. https://doi.org/10.1002/wcc.95.

Elgin, C. Z. (2017). *True enough*. Project MUSE: The MIT Press.

Eyring, V., et al. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016.

Fischer, E. M., & Knutti, R. (2016) Observed heavy precipitation increase confirms theory and early models. *Nature Climate Change*, *6*(11), 986–991. (Nature Publishing Group). https://doi.org/10.1038/nclimate3110.

Fitelson, B. (2001). A Bayesian account of independent evidence with applications. *Philosophy of Science, 68*(S3), S123–S140. https://doi.org/10.1086/392903.

Flato, G. M. (2011). Earth system models: An overview. *Wiley Interdisciplinary Reviews: Climate Change, 2*(6), 783–800. https://doi.org/10.1002/wcc.148.

Gent, P. R., et al. (1995). Parameterizing eddy-induced tracer transports in ocean circulation models. *Journal of Physical Oceanography*, *25*(4), 463–474. (American Meteorological Society).

Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research, 113*(D6), 1–20. https://doi.org/10.1029/2007JD008972.

Hall, A., & Qu, X. (2006). Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophysical Research Letters, 33*(3), L03502. https://doi.org/10.1029/2005GL025127.

Harrison, S. P., et al. (2015). Evaluation of CMIP5 palaeo-simulations to improve climate projections. *Nature Climate Change, 5*(8), 735–743. https://doi.org/10.1038/nclimate2649.

Held, I. M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society, 86*(11), 1609. https://doi.org/10.1175/BAMS-86-11-1609.

Jun, M., Knutti, R., & Nychka, D. W. (2008). Spatial analysis to quantify numerical model bias and dependence. *Journal of the American Statistical Association, 103*(483), 934–947. https://doi.org/10.1198/016214507000001265.

Kay, J. E., Holland, M. M., & Jahn, A. (2011). Inter-annual to multi-decadal Arctic sea ice extent trends in a warming world. *Geophysical Research Letters, 38*(15), 2–7. https://doi.org/10.1029/2011GL048008.

Knutti, R. (2008). Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 366*(1885), 4647–4664. https://doi.org/10.1098/rsta.2008.0169.

Knutti, R. (2010). The end of model democracy? *Climatic Change*, *102*(3–4), 395–404. https://doi.org/10.1007/s10584-010-9800-2.

Knutti, R. (2018). Climate model confirmation: From philosophy to predicting climate in the real world. In *Climate modelling* (pp. 325–359). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65058-6_11.

Knutti, R., & Sedláček, J. (2012). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change, 3*(4), 369–373. (Nature Publishing Group). https://doi.org/10.1038/nclimate1716.

Knutti, R., Furrer, R., et al. (2010a). Challenges in combining projections from multiple climate models. *Journal of Climate, 23*(10), 2739–2758. https://doi.org/10.1175/2009JCLI3361.1.

Knutti, R., Abramowitz, G., et al. (2010b). Good practice guidance paper on assessing and combining multi model climate projections, meeting report of the intergovernmental panel on climate change expert meeting on assessing and combining multi model climate projections. In T. F. Stocker, et al. (Eds.), *IPCC working group I technical support unit*. Switzerland: University of Bern, Bern.

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters, 40*(6), 1194–1199. https://doi.org/10.1002/grl.50256.

Knutti, R., et al. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters, 44*(4), 1–10. https://doi.org/10.1002/2016GL072012.

Lenhard, J., & Winsberg, E. (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. (Elsevier). https://doi.org/10.1016/j.shpsb.2010.07.001.

Lenton, T. M., et al. (2008). Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0705414105.

Levins, R. (1966). The strategy of model building in population biology. *American Naturalist*, 421–431. https://doi.org/10.2307/27836590.

Lloyd, E. 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.

Lloyd, E. (2010). Confirmation and robustness of climate models. *Philosophy of Science*, *77*(5), 971–984. Retrieved July 7, 2014, from http://www.jstor.org/stable/10.1086/657427.

Mahlstein, I., & Knutti, R. (2012). September Arctic sea ice predicted to disappear near 2 °C global warming above present. *Journal of Geophysical Research, 117*(D6), 1–11. https://doi.org/10.1029/2011JD016709.

Masson, D., & Knutti, R. (2011a). Climate model genealogy. *Geophysical Research Letters, 38*(8), L08703. https://doi.org/10.1029/2011GL046864.

Masson, D., & Knutti, R. (2011b). Spatial-scale dependence of climate model performance in the CMIP3 ensemble. *Journal of Climate, 24*(11), 2680–2692. https://doi.org/10.1175/2011JCLI3513.1.

Masson, D., & Knutti, R. (2013). Predictor screening, calibration, and observational constraints in climate model ensembles: An illustration using climate sensitivity. *Journal of Climate, 26*(3), 887–898. https://doi.org/10.1175/JCLI-D-11-00540.1.

Massonnet, F., et al. (2012). Constraining projections of summer Arctic sea ice. *The Cryosphere*, *6*(6), 1383–1394. https://doi.org/10.5194/tc-6-1383-2012.

Mazzocchi, F., & Pasini, A. (2017). Climate model pluralism beyond dynamical ensembles. *Wiley Interdisciplinary Reviews: Climate Change, 8*(6), e477. https://doi.org/10.1002/wcc.477.

McFarlane, N. (2011). Parameterizations: Representing key processes in climate models without resolving them. *Wiley Interdisciplinary Reviews: Climate Change, 2*(4), 482–497. https://doi.org/10.1002/wcc.122.

Medhaug, I., et al. (2017). Reconciling controversies about the "global warming hiatus". *Nature*, *545*(7652), 41–47. (Nature Publishing Group). https://doi.org/10.1038/nature22315.

Notz, D., & Stroeve, J. (2016). Observed Arctic sea-ice loss directly follows anthropogenic CO_2 emission. *Science, 354*(6313), 747–750. https://doi.org/10.1126/science.aag2345.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*(5147), 641. AAAS. Retrieved June 4, 2014, from http://www.sciencemag.org/cgi/content/abstract/sci;263/5147/641.

Overland, J. E., & Wang, M. (2013). When will the summer Arctic be nearly sea ice free? *Geophysical Research Letters, 40*(10), 2097–2101. https://doi.org/10.1002/grl.50316.

Parker, W. S. (2006). Understanding pluralism in climate modeling. *Foundations of Science*, *11*(4), 349–368. (Springer). http://www.springerlink.com/index/138424X1082M7277.pdf.

Parker, W. S. (2009). Confirmation and adequacy-for-purpose in climate modelling. *Aristotelian Society Supplementary, 83*(1), 233–249. https://doi.org/10.1111/j.1467-8349.2009.00180.x.

Parker, W. S. (2010). Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *41*(3), 263–272. (Elsevier). https://doi.org/10.1016/j.shpsb.2010.07.006.

Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science, 78*(4), 579–600. https://doi.org/10.1086/661566.

Parker, W. S. (2013). Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change, 4*(3), 213–223. https://doi.org/10.1002/wcc.220.

Pennell, C., & Reichler, T. (2011). On the effective number of climate models. *Journal of Climate, 24*(9), 2358–2367. https://doi.org/10.1175/2010JCLI3814.1.

Saffioti, C., et al. (2016). Reconciling observed and modeled temperature and precipitation trends over Europe by adjusting for circulation variability. *Geophysical Research Letters, 43*(15), 8189–8198. https://doi.org/10.1002/2016GL069802.

Sanderson, B. M., & Knutti, R. (2012). On the interpretation of constrained climate model ensembles. *Geophysical Research Letters, 39*(16), L16708. https://doi.org/10.1029/2012GL052665.

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015a). A representative democracy to reduce interdependency in a multimodel ensemble. *Journal of Climate, 28*(13), 5171–5194. https://doi.org/10.1175/JCLI-D-14-00362.1.

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015b). Addressing interdependency in a multimodel ensemble by interpolation of model properties. *Journal of Climate, 28*(13), 5150–5170. https://doi.org/10.1175/JCLI-D-14-00361.1.

Sanderson, B. M., Wehner, M., & Knutti, R. (2017). Skill and independence weighting for multi-model assessments. *Geoscientific Model Development*, *10*(6), 2379–2395. https://doi.org/10.5194/gmd-10-2379-2017.

Schaller, N., et al. (2011). Analyzing precipitation projections: A comparison of different approaches to climate model evaluation. *Journal of Geophysical Research, 116*(D10), 1–14. https://doi.org/10.1029/2010JD014963.

Schneider, T., et al. (2017). Climate goals and computing the future of clouds. *Nature Climate Change, 7*(1), 3–5. (Nature Publishing Group). https://doi.org/10.1038/nclimate3190.

Schupbach, J. N. (2016). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science*, *69*(February), axw008. https://doi.org/10.1093/bjps/axw008.

Screen, J. A., & Francis, J. A. (2016). Contribution of sea-ice loss to Arctic amplification is regulated by Pacific Ocean decadal variability. *Nature Climate Change, 6*(9), 856–860. https://doi.org/10.1038/nclimate3011.

Stainforth, D. A., 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.

Stegenga, J., & Menon, T. (2017). Robustness and independent evidence. *84*(July), 414–435. http://www.journals.uchicago.edu/doi/10.1086/692141.

Swart, N. C., et al. (2015). Influence of internal variability on Arctic sea-ice trends. *Nature Climate Change, 5*(2), 86–89. (Nature Publishing Group). https://doi.org/10.1038/nclimate2483.

Tebaldi, C., et al. (2004). Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophysical Research Letters, 31*(24), 1–5. https://doi.org/10.1029/2004GL021276.

Vuuren, D. P., et al. (2011). The representative concentration pathways: An overview. *Climatic Change, 109*(1–2), 5–31. https://doi.org/10.1007/s10584-011-0148-z.

Waugh, D. W., & Eyring, V. (2008). Quantitative performance metrics for stratospheric-resolving chemistry-climate models. *Atmospheric Chemistry and Physics*, *8*(18), 5699–5713. https://doi.org/10.5194/acp-8-5699-2008.

Weigel, A. P., et al. (2010). Risks of model weighting in multimodel climate projections. *Journal of Climate, 23*(15), 4175–4191. https://doi.org/10.1175/2010JCLI3594.1.

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science, 73*(December), 730–742.

Wimsatt, W. C. (2012). Robustness, reliability, and overdetermination (1981). In L. Soler et al. (Eds.), *Characterizing the robustness of science: After the practice turn in philosophy of science* (pp. 61–87). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-2759-5_2.

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology, 13*(2), 219–240. https://doi.org/10.1080/13501780600733376.

Chapter 35
Challenges to Simulation Validation in the Social Sciences. A Critical Rationalist Perspective

Michael Mäs

Abstract I reflect on challenges to the validation of theoretical models from the perspective of a critical rationalist seeking to develop true explanations of empirical phenomena. I illustrate my arguments with examples from the rich literature on social-influence models, a field that has profited from contributions from various disciplines such as physics, and mathematics. While this field is characterized by a large number of competing formal models, it has been criticized for having failed to generate reliable explanations and predictions, because of a lack of empirical research validating models. I list five challenges to model validation in the social sciences: First, social-scientific theories are based on many obscure concepts. Second, many social-scientific concepts are latent. Third, the representation of time is unclear in many models. Forth, in most social settings, various processes influence dynamics in parallel. Fifth, context dependencies limit the development of general models.

Keywords Model validation · Social-influence · Social networks

35.1 Introduction

In 1964, Robert Abelson formulated a research question that many still consider to be one of the most challenging and persistent puzzles of the social sciences (Bonacich and Philip 2012; Mason et al. 2007; Hegselmann and Krause 2002; Mäs et al. 2010; Flache et al. 2017). Abelson developed formal models of social-influence dynamics in networks where nodes are described by a continuous attribute, often called an opinion, that is open to influence from network neighbors. Being puzzled by the finding that social influence generates convergence cascades that inevitably lead to perfect opinion consensus in connected networks, Abelson wondered "what on earth one must assume in order to generate the bimodal outcome of community cleavage studies." (1964, p. 153). In the past decades, hundreds of articles extended Abelson's

M. Mäs (✉)
Department of Sociology/ICS, University of Groningen, Groningen, The Netherlands
e-mail: m.maes@rug.nl

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_35

work and provided various competing answers to his question, creating an impressive literature that profited from contributions from disciplines as diverse as sociology, physics, computer science, philosophy, economics, communication science, political science, mathematics, and complexity science. Despite this impressive scholarly attention, a recent review of the literature concluded that "the literature still cannot offer reliable explanations and predictions for real-life influence dynamics" (Flache et al. 2017), criticizing that researchers have accumulated an overwhelming number of models but failed to develop empirical insights into which of many model candidates provides the most accurate description of the social-influence dynamics in a given real-life setting.

The literature on social-influence dynamics in social networks ideal-typically illustrates the intricacy of validation in sociological research programs. On the one hand, there is a rich arsenal of rigorous theoretical models analyzed with both analytical tools and computer simulation. Researchers also invest heavily into validating their models, confronting model assumptions and predictions with empirical data gathered the field and the laboratory. On the other hand, progress appears to be limited, both in terms of an emerging scholarly consensus on a standard model, and the ability to make accurate predictions about empirical phenomena.

Here, I reflect on fundamental methodological challenges that social scientists encounter in the process of model validation, arguing that these challenges might be responsible for the relatively limited progress in social-scientific modeling and model validation. The challenges that I list in this chapter are also well known in other disciplines but I argue that they create particularly problematic road blocks in the social sciences.

To illustrate my arguments, I repeatedly refer to the literature on social-influence dynamics in networks that emerged from Abelson's work. This field serves as an ideal-typical example, because many contributors follow the methodological guidelines of critical rationalism that dominate also in the natural sciences. That is, they propose general theories and rigorously deduce hypotheses that they critically confront with empirical data in order to gradually improve their models. What is more, social-influence models have also been studied by researchers coming from disciplines that seem to be characterized by faster progress, such as physics, mathematics, and computer science. This supports the notion that progress in the validation of social-scientific models is limited not because researchers apply problematic approaches but because validation is particularly challenging in the social sciences.

The present chapter is written from a critical rationalist perspective. Critical rationalism holds that in order to adequately explain an empirical phenomenon, scientists need to develop theories (a synonym is "model") that are based on general, law-like statements and demonstrate that statements about the observed phenomenon follow in a logically valid way from the statements of the theory (Hempel and Oppenheim 1948; Popper 1959; Nagel 1979). That is, it needs to be demonstrated that the statement that is being explained must be true if all implicit and explicit assumptions of the theory are true. When explanations are complex, human intuition tends to be too error-prone to demonstrate their logical validity, making it necessary to apply rigorous methods. Computers simulation is one of the available methods.

To run a computer simulation, the assumptions of a theory are translated into computer code. Next, the researcher uses the computer to derive and analyze the implications of the implemented assumptions. Compared to methods from formal logic, and mathematical methods, computer simulation tends to provide the researcher with more flexibility in the choice of modeling assumptions, making simulation the method of choice whenever analytical solutions are unavailable. While analytical methods come with the great advantage of providing formal proofs, they often force the researcher to make restrictive theoretical assumptions about, for instance, the behavior of humans, their perception of the environment, and the process of individual decision-making. Simulation is a powerful method to rigorously study the implications of theories that are based on more relaxed assumptions, for instance with the aim of developing more realistic theories, or the aim of answering what-if questions, or the aim of studying whether important predictions of a theory change when an assumption has been altered.

Hence, whether a researcher relies on computer simulation or an alternative method depends very much on the purpose of the theoretical analysis. The literature on social-influence models, for instance, provides numerous examples of models that were studied with both computer simulation and analytical tools (Hegselmann and Krause 2002; Lorenz 2005; Castellano et al. 2009). Accordingly, I do not restrict myself to challenges to the validation of theories studied with computer simulation but discuss the validation of theoretical models in general. However, I illustrate my arguments with a class of formalized social-scientific theories that are typically analyzed with computer simulation.

In the present chapter, the term "validation" describes the process of confronting a theory with empirical evidence with the ultimate aim of developing a sound explanation of the empirical phenomenon. That is, a critical rationalist seeks to develop theories that are not only "logically valid" but also based only on statements that are either true or theoretically innocent in that they do not affect important logical implications of the theory. Accordingly, one would consider a simulation model a "valid" explanation of a given phenomenon when all implemented assumptions are either true or innocent (Gilbert and Troitzsch 1999).[1] Theories can be empirically tested either directly, by putting their core assumptions to the empirical test, or indirectly, by empirically testing model implications that have been derived from the theory. While computer simulation and other formal methods do not allow to empirically test theories, they often play an important role in the process of indirect empirical testing, as they facilitate the process of identifying implications of theories that contradict alternative theories or intuition, theoretical predictions that are the preferred candidates for empirical tests.

It should be noted that the definition of the term validation used here is very restrictive in that it can only be applied to modeling efforts aimed at developing

[1] The concept of validity has been used in many different ways. I use it here in accordance with many contributions to the social-simulation literature (David 2009), defining a model as valid if it is based on true assumptions. In the field of logic, one would call such a theory "sound" rather than "valid," because in this literature explanations are considered valid when their assumptions logically imply the explanandum.

true explanations (Ahrweiler and Gilbert 2015; Gilbert and Ahrweiler 2005). Some of the most famous simulation work in the social sciences, however, has not been conducted with this aim. Schelling's seminal model of residential segregation, for instance, is frequently used to demonstrate that even cities with unrealistically tolerant populations can fall apart into ethnically homogenous districts (Hegselmann 2017; Schelling 1971; Sakoda 1971). Today, Schelling's model serves as an ideal-typical example of unintended consequences arising from the social interaction. A model that is based on more realistic assumptions about when and why individuals move to another neighborhood would not serve this purpose better. Thus, validity, as it is defined here, is not a quality criterion for this work on Schelling's model.

Already, the founding fathers of critical rationalism have acknowledged fundamental philosophical problems that also limit their approach (Popper 1959). For instance, it is argued that an adequate theory requires at least one general, law-like assumption, a statement about the truth that can never be proven true simply because it is general and empirical observation is specific. As a consequence, one can never be sure that one has identified a sound explanation, even when one has repeatedly failed to prove that it is not sound. Accordingly, one can also never be sure whether one has developed a valid simulation model. Likewise, a fundamental challenge to validation is that empirical testing always requires general assumptions about the measurement of reality, statements that cannot be proven true either. Nevertheless, the here adopted methodological approach is the dominant approach in many scientific disciplines and has led to important advances also in the social sciences, as the example of social-influence modeling demonstrates. Furthermore, the here discussed challenges to model validation, do also apply to alternative methodological approaches that seek to explain empirical phenomena based on general theories and consider the empirical testing of theories and their predictions a critical step toward arriving at true explanations. However, this chapter does not address challenges relating to the simulation and validation of models which are based on the interpretive and constructivist traditions in the social sciences (for a discussion of some alternative purposes of simulation modeling and competing methodological approaches in the social sciences see Ahrweiler and Gilbert 2015; Gilbert and Ahrweiler 2005; for a discussion of the validation of socioecological simulation models and a post-positivist understanding of the concept of validity see Chap. 17 by Saam in this volume).

In the following section, I provide a short overview over core contributions to the literature on social-influence models and approaches to validating these models. Subsequently, I reflect on five challenges to model validation. Finally, I discuss implications, formulating four recommendations for future modeling efforts in the social sciences.

35.2 Illustrative Example: Models of Social Influence

Models of social influence are concerned with dynamics of the distributions of opinions, beliefs, and behavior in social groups, organizations, and societies. Typical empirically observed dynamics that modelers seek to explain are the emergence of a shared consensus, the formation and stability of subgroups with different opinions, and the process of polarization where subgroups develop, for instance, increasingly different opinions.

While social-influence models are concerned with collective phenomena, they are based on assumptions about the behavior of individuals. Individuals are modeled as network nodes *i* that exert social influence on each other's attributes $x_{i,t}$. Already early contributions like Abelson's work represented these attributes $x_{i,t}$ on a continuous scale ranging from zero to one ($0 \leq x_{\mathrm{i,t}} \leq 1$) and referred to them as individuals' "opinion" on a given issue.

In this general framework, social influence is typically modeled as *weighted averaging* (Friedkin and Johnsen 2011). That is, when a node *i's* opinion is updated, the new opinion value $x_{i,t+1}$ is a function of *i*'s previous opinion $x_{i,t}$ and the weighted average of the opinions held by *i's* network contacts. Formally,

$$x_{i,t+1} = x_{i,t} + \gamma \cdot \sum\nolimits_{j=1}^{K} w_{ij,t}(x_{j,t} - x_{j,t}). \tag{35.1}$$

Parameter γ controls how open nodes are to social influence ($0 < \gamma \leq 1$). The influence weights $w_{ij,t}$ describe the social influence between all pairs of network nodes. When two network nodes *i* and *j* are not connected by a network link, the influence weight $w_{ij,t}$ adopts a value of zero. Small positive values, however, imply that node *j* exerts weak influence on *i*. Higher values correspond to stronger social influence.

The influence weights $w_{ij,t}$ play a critical role in the dynamics that influence models to generate and are, therefore, also in the focus of previous efforts to validate influence models. Early models assumed that weights remain unchanged over time and that they either adopt a value of zero, when two nodes *i* and *j* do not influence each other, or a positive value to represent that agents grow more similar due to social influence. Models that assume fixed and nonnegative weights predict that any population will inevitably arrive at a perfect consensus unless the social networks consist of two or more unconnected components. The contradiction between this model prediction and outcomes of empirical studies in small communities, which often found opinion differences to be stable and sometimes even increasing, motivated Robert Abelson to formulate the research puzzle quoted in the introduction of this paper.

A powerful approach to solve Abelson's puzzle is to drop his assumption that weights are fixed. Figure 35.1 shows three weight functions that make different assumptions about how weights might depend on the opinion difference between nodes. Figure 35.2 shows the collective opinion dynamics that these weight functions generate in complete networks of 101 nodes. All three dynamics departed from a uniform opinion distribution. The modeled opinion varies between values of zero

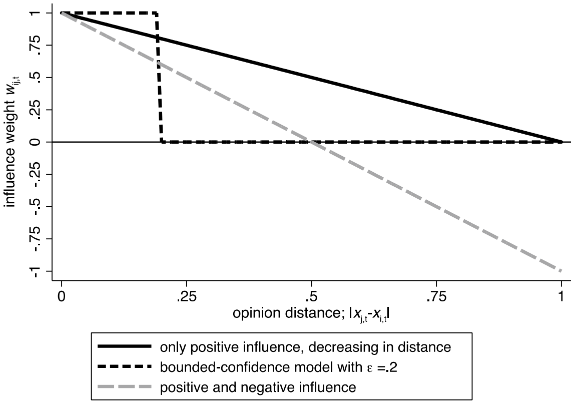


Fig. 35.1 Three prominent weights functions

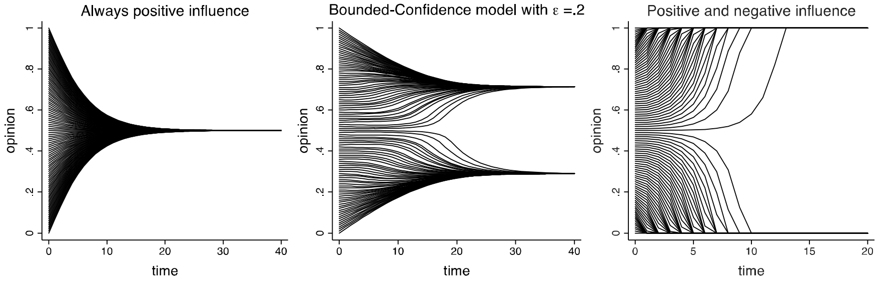


Fig. 35.2 Opinion dynamics predicted by social influence models with three different weight functions. Lines show the opinion trajectory of the 101 agents in each population. Modeling details are described by Flache et al. (2017)

and one ($0 \leq w_{ij,t} \leq 1$). The bold weight function in Fig. 35.1 assumes that influence weights decrease when *i* and *j* disagree which implements the notion that individuals feel attracted by similar others are, therefore, more open to influence. However, since weights remain positive unless nodes disagree maximally, model dynamics result in the same outcome as observed by Abelson, perfect opinion consensus (see left-hand-side panel of Fig. 35.2).

A prominent approach to prevent opinion convergence is to assume that individuals reject any influence from network neighbors they disagree with (Hegselmann and Krause 2002; Deffuant et al. 2005; Lorenz 2007). Formally, modelers included that the influence weights $w_{ij,t}$ drop to zero when the opinion distance between *i* and *j* exceed a threshold ε, as the short-dashed line in Fig. 35.1 illustrates. When this threshold is sufficiently small, these so-called "models of bounded confidence" can explain the emergence of several internally homogenous but mutually distinct subgroups (see center panel of Fig. 35.2). These subgroups can remain stable if members of opposite groups hold opinions that differ too much and therefore fail to influence each other. These models, thus, provide an explanation for the emergence and stability of multiple subgroups with different opinions. However, without additional assumptions, these models fail to explain the processes of opinion polarization where the opinion differences between subgroups increase over time (Hegselmann and Krause 2002).

A further extension was the assumption of negative-influence weights, which implements the notion that individuals may dislike certain network neighbors and therefore seek to increase opinion differences to them, an assumption that has been defended with prominent sociological theories of social differentiation and social psychological theories of intergroup relations (Macy et al. 2003; Mark 2003; Salzarulo 2006; Takács et al. 2016). In particular, it has been assumed that network nodes do not only reject influence from nodes that hold distant opinions but tend to increase opinion differences to these agents (see the long-dashed line in Fig. 35.1). When this negative form of social influence is sufficiently strong, networks can fall apart into multiple subgroups with increasingly different opinions, as the right-hand-side panel of Fig. 35.2 shows. This polarization process continues until agents adopted opinions at opposite poles of the opinion scale.

As assumptions about the weight function have critical impact on model predictions, efforts to validate social-influence models focused on the weight functions. There have been two main approaches to validating weight functions, a micro-approach, and a macro-approach. Following the *micro-approach*, empirical studies in the field and the laboratory measured individuals' opinions before and after exposure to information about the opinions of others, measuring or experimentally manipulating opinion differences to the source of influence (Takács et al. 2016; Clemm von Hohenberg et al. 2017; Marsden and Friedkin 1993; Friedkin and Johnsen 2011; Liu and Srivastava 2015). Next, it is statistically estimated how the opinion difference between the individual and the source of influence affected the size and the direction of the opinion shift after the influence event.

The macro-approach to validating social-influence models takes advantage of the relationship between microassumptions about influence weights and emerging macropatterns of opinion convergence, fragmentation into multiple groups, and opinion polarization (Brousmiche et al. 2016; Chattoe-Brown 2014; Jan Lorenz 2017; Clemm von Hohenberg et al. 2017; Friedkin and Johnsen 2011; Mäs and Flache 2013). Analyzing, for instance, data from large-scale representative surveys or controlled laboratory experiment with repeated opinion measurements, scholars described the observed collective opinion dynamics in terms of shifts in opinion variance, shifts in the opinion average, and changes in the degree of opinion polarization. Next, they drew conclusions about which assumptions about influence weights are necessary to explain the observed macro-dynamics. For instance, observing that opinions on initially highly controversial issues converged would lead one to conclude that negative influence played a minor role in these dynamics.

Despite all efforts to validate models of social-influence dynamics, a recent review concluded that the "empirical literature on social influence is still too limited to validate social-influence models at the level of precision needed for empirically informed choices between model alternatives." (Flache et al. 2017). The same review admitted that this limitation results from methodological challenges that researchers face, arguing that "the assumptions of social-influence models, the dynamics that they generate, and also many of their predictions are notoriously difficult to put to the test." In the remainder of the present paper, I reflect on these difficulties, arguing that they reflect general methodological challenges that social scientists encounter and using social-influence models as an illustration.

35.3 Challenges to Model Validation

35.3.1 Obscure Concepts

Most social sciences are very young. Emile Durkheim, who was the first professor in the discipline of sociology, for instance, received a teaching position at the University of Bordeaux in 1887, less than 150 years ago. As a consequence, many social-scientific concepts are still vaguely defined and disciplines have not developed consensus on the meaning of large parts of their terminology. This is a serious problem for model validation.

The field of social-influence dynamics serves as an example, as even the term "opinion" remains obscure in large parts of the literature. Today, most contributions refer to the characteristics that are socially influenced as individuals' "opinions" without including a definition of the concept. A prominent exception is the cultural-dissemination model by Robert Axelrod, who developed an influence model of "cultural attributes," defining the term culture as "the set of individual attributes that are subject to social influence" (Axelrod 1997, p. 204). On the one hand, this very broad definition comes with the advantage that Axelrod's model can be applied to any social context where individuals exert some form of influence on each other. For instance, the cultural-dissemination model can be applied to all forms of behavior (e.g., church attendance, signing a petition) and all forms of cognitions (e.g., believing in god, dissatisfaction with the government) that are open to influence. This is an important advantage, because models are easier to empirically falsify when they make predictions about many different contexts. On the other hand, theories suggest that the social influence of opinions works in very different ways than influence of behavior. A common and empirically supported assumption, for instance, is that individuals participate in political protest, such as the Arab Spring or the revolution in East Germany, more likely when they learn that many other members of their group also become active (Opp and Gern 1993; Opp 2009). One possible mechanism explaining this instance of social influence on a behavior is that individuals expect less repression by the state when they join a bigger movement. This reasoning should be independent of characteristics of the other protestors—only their number affects whether state repression is more of less likely. In contrast, the influence of opinions is explained in very different ways. Social-identity theory and self-categorization theory (Tajfel and Turner 1986; Brewer 1991; Salzarulo 2006), for instance, predict that individuals seek to intensify opinion differences to members of psychologically salient outgroups, trying to maximize their social distinctiveness. Such negative forms of social influence are not plausible in the context of individuals avoiding state repression. This simple example illustrates that social influence on behavioral dimensions is often traced back to very different psychological mechanisms than influence on opinions, and that these mechanisms may imply very different assumptions about how individuals influence each other. Thus, a researcher interested in behavior would often found her model on very different assumptions than a researcher focusing on opinions, which suggests that Axelrod's definition is too general.

However, even if one restricts the term opinion to individuals' cognitions, more clarification is needed. This is because there are fundamental differences between different forms of cognitions. One the one hand, there are so-called "beliefs," statements about the world that are either true or false (e.g., the statement "It will be raining today." is either true or false). On other hand, there are "evaluations" or "preferences", statements that are by definition not true or false. For instance, the statement "I hate rain." reflects that a person has negative feelings toward an object, which cannot be described by the attributes true or false.[2] The problem is that theories of belief influence often differ from the theory of influence of evaluations. Belief influence, for instance, plays an important role in economic theories of stock markets where traders are often imperfectly informed about the value of a stock (Bikhchandani et al. 1992). A critical assumption, however, is that the price of a good is informative of its value, as it reflects other traders' beliefs about the good's value. Thus, when the price of a good increases, traders might infer that other traders have different information and, as a consequence, different beliefs about the value of the stock. Price changes are reliable signals, because traders invested their own money. As a consequence, rational traders will be positively influenced by the behavior of others and also invest. In fact, no matter how much some traders might dislike other traders, it is not wise to be negatively influenced by their beliefs, as this would likely result in a financial loss. In contrast, negative influence might play an important role in the context of political preferences (Baldassarri and Bearman 2007). As already mentioned above, there are psychological theories that predict that individuals develop increasingly positive views on a political candidate when they learn that members of the opposite social category dislike that person.

Obscure concepts are a problem for model validation, because empirical studies might generate very inconsistent findings when they are testing the same model in fundamentally different contexts. Depending on whether one empirically studies social influence on behavior, beliefs, or evaluations, different models of influence might be supported by the data, leaving the researcher uncertain about which model is the best.

[2] To be sure, the statement "I feel that I hate rain" can be true (if I do hate rain) or false (If I do not hate rain). Nevertheless, an individual's evaluation of rain as being negative cannot be described with the words "true" or "false".

35.3.2 Abundance of Latent Concepts

Compared to many disciplines in the natural science, the social sciences seem to be concerned with an abundance of latent concepts. That is, social scientists study many phenomena that are not directly observable and, therefore, require more or less complex measurement techniques for quantification. A typical example of a latent concept from the natural sciences is temperature, which is defined as the energy of the random motion of the microscopic particles in a system. Temperature cannot be observed directly, because particles and their movement are too minuscule. However, one can take advantage of the close relationship between the temperature and the volume of a system, which inspired Fahrenheit to develop of the mercury thermometer in 1714. As temperature increases, the mercury expands and rises in the thermometer's tube, which makes it possible to relate the mercury level in the tube to temperature. While the thermometer actually quantifies the volume rather than the movement of particles, it allows to quantify temperature on the so-called "ratio level." This scale level allows to rank order the temperature of systems in that, for instance, a temperature of 5 °C is considered warmer than 3 °C. It is also possible to interpret temperature differences. For example, the difference between 5 °C and 10 °C is the same as the difference between 95 °C and 100 °C. Furthermore, when measured on the Kelvin temperature scale, there is even a meaningful zero point, which allows statements about the relative temperature. 100 K, for instance, is twice as warm as 50 K.

Unfortunately, there are very few social-scientific concepts that can be quantified directly or that have been measured with a similar accuracy as temperature. On the level of social collectives, for instance, social scientists struggle with quantifying key concepts such as a nation's wealth, social inequality, or a collective's level of cohesion. On the level of the individuals, there are debates about measuring concepts such as an individual's satisfaction, happiness, social identity, social and human capital, or social status. Even sex and gender are often hard to measure unequivocally.

Measurement problems are a fundamental roadblock for the validation of social-influence models. A typical example of an opinion that these models are concerned with is individuals' political orientation, which is prominently measured on the left-–right political spectrum. The General Social Survey, one of the most important social-scientific surveys, for instance, has been asking respondents already since 1972 the following question: "In politics people sometimes talk of left and right. Where would you place yourself on a scale from 0 to 10 where 0 means the left and 10 means the right?"[3] This measurement instrument seems to have a similar scale as temperature, but there are many differences. One the one hand, it is possible to rank-order measurements. For instance, if a respondent's answer to the question switches from 2 to 4, one can conclude that her political orientation shifted toward the right. On the other hand, can one conclude that a respondent choosing answer category 2 is more leftist than a participant who selected option 4? The problem is that respondents might interpret the question and the answer categories differently,

[3] https://gssdataexplorer.norc.org/variables/4971/vshow.

which makes comparison across individuals difficult. Likewise, the same individual may change her interpretation of the scale over time, because the meaning of the terms “right” and “left” changed over time. Political stances that were considered very leftist a few decades ago (e.g., the support of abortion) are today conceived as being moderate. Thus, a person whose actual orientation remained unchanged, might nevertheless choose today a different answer to the survey question than 20 years ago, and vice versa. Another problem is that the scale may not be linear. Does an opinion shift from 1 to 2 indicate the same absolute opinion change as a shift from 4 to 5? We do not know.

Both the macro- and the micro-approach to validating social-influence models are affected by these measurement problems. Empirically testing model predictions about the emergence of consensus, fragmentation, or polarization requires a quantification of these macroconcepts. For instance, finding that the distribution of political orientation in a population has two modes, one would conclude that the population is fragmented into two camps. However, it may be that individuals actually agree much more than the distribution suggests, but the population consists of members of different generations who interpret the labels “right” and “left” in different ways.

Another problem for macrovalidation efforts is that the measurement problems make it difficult to aggregate individual-level data. The degree of polarization in a population, for instance, is typically measured with the variance of individuals’ opinions or Esteban and Ray’s polarization measure (Boxell et al. 2017; DiMaggio et al. 1996; Esteban and Ray 1994; Bramson et al. 2016), outcomes that require opinion measures on a continuous scale. To calculate variance, for instance, one sums up the squared gap between individuals’ opinions and the opinion average. This gap cannot be calculated, however, if individuals might have based their answers on different interpretations of the answer scale.

Likewise, micro-validation is seriously affected by these measurement problems. One approach to empirically testing which of the three weight functions shown in Fig. 35.1 is accurate, is to measure as the dependent variable in a statistical regression model individuals’ opinion shifts resulting from social influence. To this end, one calculates the difference between the opinions before and after exposure to the source of influence. Next, one estimates statistical models with the initial opinion distance between the individual and the source of influence as the main independent variable, as opinion distance is the determinant of the influence weights in the formal models (see Fig. 35.1). Finding that individuals who were exposed to a dissimilar source adjusted their opinions away from the source would be considered support for the negative-influence assumption. The problem is that both the dependent and the independent variable of this statistical model require opinion measurements on a continuous scale, as both quantify a difference between two opinions. Most opinion scales, however, do not provide continuous measures.

Such measurement problems are well known in the social sciences. Some of them can be overcome with elaborated statistical methods, such as structural equation modeling or regression techniques for ordinal dependent variables (Loehlin and Beaujean 2016). Another approach has been to shoehorn opinion measurements into ratio scales. For instance, researchers measured individuals’ opinions about the smoking ban in restaurants and bars by asking which share of tables in a restaurant should be reserved for nonsmokers, which can be expressed in percentages (Takács et al. 2016). In another study, participants had to find the optimal spot for a new leisure center (Mäs and Flache 2013). They could choose between two cites that were 50 km apart or any spot between the two cities. Thus, geographic distance provided a continuous scale. While this approach was applied successfully in laboratory experiments where the actual opinion issue was not important for the purpose of the study, it seems difficult to apply it for instance to the measurement of individuals' political orientation. A third approach to deal with measurement problems of latent concepts is to operationalize opinions with multiple indicators and aggregate (e.g., mean, or median) individuals' answers. For instance, in an attempt to measure political conservatism, an important dimension of political orientation, a scale has been proposed that builds on 12 items (Everett 2013). Respondents are confronted with topics such as abortion, gun ownership, and religion" and are asked to indicate for each item how positive or negative they feel about each item on the scale of 0–100. However, this solves problems that researchers interested in social influence encounter only if the mentioned measurement inaccuracies are captured by in the aggregation process, which is a strong and hard to test assumption.

Despite these efforts to overcome measurement problems, model validation remains a challenge. The core problem is that models are hard to falsify with problematic measurement tools. For instance, studies that did not find support for the assumption that influence can turn negative when the individual and the source of influence disagree can always be criticized for basing their finding on problematic opinion measurements (Takács et al. 2016; Clemm von Hohenberg et al. 2017). As a consequence, it is hard for the discipline to decrease the number of candidate social-influence models and identify a standard model. A second problem is that empirical research does not generate insights with a high precision. On the one hand, the theoretical models make very precise assumptions about social influence, specifying the exact change in opinions depending on the attributes of the source. What is more, often even seemingly minimal changes in these assumptions have critical impact on model predictions (Mäs et al. 2010; Flache and Macy 2011a). On the other hand, the described measurement problems make it difficult if not impossible to feed models with empirically validated assumptions with a comparable precision.

35.3.3 Representation of Time

Any model developed to capture the dynamics of a social system needs to represent time, independent of whether the time is modeled continuously or as a sequence of discrete events. The sketched models of social influence, for instance, assumed that the dynamics can be broken down to a number of discrete events t. At every event, each agent's opinion is updated exactly once. Figure 35.2 shows that it took the model that assumes only positive social influence about 40 events to generate opinion consensus. It remains unclear, however, to what amount of time this number of simulation events corresponds. Is it a month, a year, or 30 min?

This lack of information can create problems for validation efforts testing the macropredictions of models, as it makes them immune to empirical falsification. For instance, if one does not find support for one of the macropredictions shown in Fig. 35.2, one could always argue that the empirical study is based on an insufficiently long time frame and would have confirmed the predictions had it involved longer time frames.

To be sure, this is not always a problem. A model of East Germans' participation in the political protests of 1989, for instance, could build on the fact that demonstrations took place every Monday. Thus, in this case, one can safely assume that individuals updated their beliefs about the number of participants and also their own behavior exactly once a week. In many other settings, however, such information is not available. How often, for instance, do individuals update their evaluations of political candidates? Does this frequency change during election periods?

An additional complication is that many models and also the social systems they represent display so-called "broken ergodicity". Dynamics tend to reach rest points, system states where dynamics have settled in that the central characteristics of entities on the micro and/or macrolevel remain constant. The three dynamics displayed in Fig. 35.2, for instance, have reached such rest points. However, in a probabilistic world, these system states will not be totally stable, even though the dynamics can rest for a very long time. At some moment, a sequence of random events will always cause the system to leave the stable state and may, in relatively short amount of time, lead the system into a new state. The problem with models that generate such dynamics is that their predictions are difficult to compare to empirically observed dynamics. While stochastic models allow one to predict that a system state will be left at some moment and also how probable a shift from one state to the other is at any given moment, it is usually impossible to predict when this shift will occur. This makes timing empirical studies challenging.

35.3.4 Interplay of Multiple Processes

A common limitation of many field studies in the social sciences is that observed dynamics can result from many different processes that act in conjunction. It is, therefore, often very challenging to draw conclusions about which processes have actually been responsible for the empirical observations in a given context.

For instance, one of the most robust findings in the social sciences is called "homophily," the notion that individuals tend to have social relationships with similar others (Lazarsfeld and Merton 1954; Wimmer and Lewis 2010; McPherson et al. 2001). Obviously, social influence can generate homophily, because it makes individuals grow more similar during the interaction. Another explanation, however, points to selection processes. Research along the similarity-attraction paradigm has demonstrated that individuals tend to develop more and closer social relationships with others who hold similar views (Byrne 1971). It might also be that individuals do not seek to develop relationships with similar persons but actually want to avoid interaction with dissimilar individuals (Rosenbaum 1986). This also generates homophily. It may even be that neither selection nor influence are active but individuals who have social relationships tend to expose themselves to similarly biased media or political opinion leaders who then affect their opinions in similar ways (Slater 2007; Iyengar and Hahn 2009).

The problem that arises from the fact that multiple social processes might be active and lead to the same consequences in a given setting is that in order to quantify social influence and validate influence models, one needs to either experimentally or statistically control for all alternative explanations. Consider, for instance, a researcher trying to validate social-influence models with online polling tools. These tools are abundant these days on online news websites, as many users consider it fun to share their opinions online and receive information about others' views. If users were positively influenced by an online news outlet, one would expect that for instance, opinions measured on rightist websites will shift toward more rightist views over time. However, this observation can also be explained by a simple selection process. It may just be that users with leftist opinions switch to alternative news outlet when they realize that their original news source has a rightist audience. As these leftist users do no longer indicate their opinions, the opinion average measured on the website will shift to the right, even without social influence.

A powerful approach to tackle the problem of multiple parallel processes is to develop highly controlled laboratory experiments (Friedkin and Johnsen 2011; Takács et al. 2016; Mäs and Flache 2013). In the laboratory, researchers can reduce the number of possible explanations, by either designing artificial settings where the impact of a given exogenous factors (e.g., media) can be excluded and/or by exposing all participants to the exact same stimuli. As a consequence, it can be excluded that observed differences in the behavior of participants result from these factors. However, since laboratory experiments control away the influence of processes that are present in the field, controlled experiments are often criticized for studying humans in artificial settings where their behavior might differ from the behavior in their normal habitat.

This problem of limited external validity in laboratory studies is also known in other disciplines. In the natural sciences, for instance, researchers study bacteria in Petri dish, to create settings where bacteria and their behavior are affected by a very limited number of external factors. An increasingly recognized problem, however, is that bacteria quickly adopt to the environment of the Petri dish. Within a few bacterial generations, researchers might thus study bacteria that differ in potentially important ways from the bacteria from natural settings that they sought to study in the first place.

A second powerful approach to tackle the problem of multiple parallel processes has been to statistically control for potentially intervening mechanisms. In the field of social-influence research, for instance, so-called "Stochastic actor-oriented models" (Steglich et al. 2010) have been elaborated to disentangle selection from influence effects. This method requires longitudinal network data about all relevant social relationships in a given population and information about individuals' characteristics such as their opinions. With a combination of statistical inference and computer simulation, this method estimates which combination of alternative processes best explains the coevolution of network ties and individuals' characteristics observed in the data. One problem is that stochastic actor-oriented models require very detailed data, which is usually gathered only in small population such as school classes or work teams. As a consequence, it is usually hard to quantify the effects of factors acting outside of the bounds of the studied populations.

35.3.5 Context Characteristics Matters

One of the key achievements of many fields in the natural sciences was the development of accepted, standard models that made correct predictions in a wide range of applications. With the exception of economics, social sciences lack such a general and accepted theoretical paradigm. The problem is not that there are no general theories in the social sciences (for an introduction, see Turner 1974, 1995) but unlike in many other disciplines, these efforts have hardly contributed to the development of a canon, a set of widely accepted assumptions, predictions, and methods. Existing grand theories have been criticized for being too remote from reality and, as a consequence, hard if not impossible to test empirically (Merton 1957).

Having made negative experiences with existing general models, many contemporary social scientists follow a so-called "middle-range approach" (Hedström and Udehn 2009; Hedström and Ylikoski 2010; Boudon 1981; Merton 1957). Rather than being applicable to a wide or even unlimited range of contexts, middle-range theories have been developed to accurately represent dynamics in a restricted set of contexts, such as only college dormitories (Garrison and Babcock 2009), or only work teams (Mäs et al. 2013). Thus, rather than developing a general model that can be applied to a wide range of social contexts, models are being tailored to very specific social setting.

Specific characteristics of a given social context can be relevant in two different ways. First, context characteristics can affect whether certain microprocesses give rise to macrooutcomes or not. There are, for instance, social-influence models that incorporate the notion that demographic differences between individuals can trigger the microprocess of negative influence, an assumption that has been derived from theories of intergroup relations and social differentiation (Tajfel 1978; Tajfel and Turner 1986; Elias 1969; Bourdieu 1984). As negative influence can explain opinion polarization, one would thus expect that demographic diversity in a given population should intensify opinion polarization. Modeling work, however, shows that polarization arises only under certain context conditions, even when there is high demographic diversity and when demographic differences between agents is assumed to motivate negative influence (Flache and Mäs 2008b; Mäs et al. 2013; Flache and Mäs 2008a; Grow and Flache 2011). For instance, in contexts where the underlying social network is highly segregated along demographic boundaries, in that there are demographically homogenous network clusters, individuals hardly interact with members of a demographic outgroup. As a consequence, negative influence is rare and social-influence dynamics do not foster opinion polarization. Another important context condition is that individuals either consider relevant only a single demographic dimension or, when multiple dimensions are salient, there is a high correlation between demographic attributes (Lau and Murnighan 2005; Flache and Mäs 2008b; Mäs et al. 2013). When demographic attributes are not aligned, pairs of individuals might differ on one dimension but will likely share other demographic characteristics. These similarities prevent negative influence, despite high levels of diversity in the population.

The fact that context characteristics can affect whether or not microprocesses have certain macroeffects is a challenge for validation projects, because the researcher needs to make sure that all important characteristics of the context of an empirical study are captured. For instance, if one is unaware of the critical effects of network segregation, finding no opinion polarization in a demographically diverse setting would lead one to the false conclusion that there is no negative influence. As one can never be 100% certain that one has captured all potentially important context characteristics, it makes it hard to falsify model assumptions.

Second, context characteristics can affect whether core model assumptions about microprocesses are true or not, which may make certain models inapplicable to some contexts. Most models of social influence, for instance, represent social-influence from the perspective of the target of influence. That is, when updating the opinion of an agent i, a network neighbor j or a set of network neighbors is selected who exerts some form of influence on i. This form of communication has been called "one-to-one communication" or "many-to-one communication" (Flache and Macy 2011b) and represents many forms of communication in offline settings. The models sketched in Sect. 35.2, for example, assume many-to-one communication, as *one* agent's updated opinion is similar to the weighted average of *many* other agents' opinion. On the Internet, however, communication is different. Bloggers and users of online social networks, for instance, emit online content that is then shared with all of the contacts at once. Thus, in these online contexts communication is reversed. Modelers referred to this form of communication as "one-to-many communication" and demonstrated that it can generate very different social-influence dynamics than other forms of communication (Flache and Macy 2011b; Keijzer et al. 2018). It turned out, for instance, that in Axelrod's model of cultural dissemination, there is an increased chance that nodes become culturally isolated when there is one-to-many communication rather than one-to-one communication. When a node i emits a message to all of her network contacts and one neighbor j is not socially influenced, then j will grow more dissimilar to those contacts that he shares with i if these actors were influenced by i. This process, it turns out, is less likely under other communication regimes, which shows that models of social influence in online settings require a different representation of communication than social influence in offline settings. The implications for model validation are important. The problem is that the results of model validation efforts in a given settings may not be applicable to other contexts. Modelers interested in social-influence processes on the Internet, for instance, debate whether their theoretical models can be informed by empirical research on offline settings (Mäs and Bischofberger 2015).

35.4 Discussion

In the remainder, I formulate four recommendations for future theoretical and empirical research that will help in tackling these challenges.

35.4.1 Compare Models and Identify Critical Assumptions!

The literature on social-influence modeling is very rich in terms of the number of models that have been developed to understand and predict the opinion dynamics that social influence generates in networks. Models are based on markedly different assumptions about how individuals influence each other and how they select influential peers. Furthermore, similar theoretical assumptions are often formally implemented in various forms. Classic models, for instance, implement opinions as positions on a continuous scale, while Axelrod assumed that agents influence each other on a nominal scale (Axelrod 1997; Flache et al. 2017; Huckfeldt et al. 2004). This diversity of models is a strength of the literature, but this strength can only be exploited when models are compared and differences are identified. So far, however, modelers tend to propose a novel model without specifying which assumptions have been adopted from an existing model and which assumption has been added.

Herbert Gintis once criticized that many social scientists handle theories like a toothbrush, in that they would never use another researcher's theory. This behavior limits scientific progress. To improve, modelers should build on existing models and add assumptions in a step-wise process. Importantly, authors should formally demonstrate that the predictions of their new model differ from the predictions of an existing model, as this shows that they have added a critical ingredient. Great examples of this approach are given by Flache and Macy (2011b) and Huckfeldt et al. (2004) in their work on Axelrod's model of cultural dissemination. Likewise, there are very few contributions to the literature where competing models are integrated into a joint framework and systematically compared to demonstrate critical differences (Mäs et al. 2014; Mäs and Bischofberger 2015).

Model validation profits from more model comparison, because resulting insights will guide empirical research. First, empirical research is most needed for those assumptions that model comparison has identified as being responsible for alternative predictions. Second, the model comparison will highlight the conditions under which alternative models make different predictions, and thus point to social contexts that allow to test the competing predictions of alternative models.

In particular, computer simulation is a powerful tool to compare competing models, because, unlike many other modeling methods, it imposes very few restrictions on the choice of theoretical assumptions. As a consequence, it is relatively easy to represent multiple competing theoretical assumptions in a single computer program and then explore how model predictions change when one switches from one assumption to the other.

35.4.2 Defend Your Assumptions!

Focusing on demonstrating which predictions follow from their model assumptions, many contributions to the modeling literature do a poor job in defending the assumptions. In particular, assumption that change model predictions need to be backed up with theoretical arguments explaining why an assumption is plausible. What is more, authors should refer to empirical research that supported their assumptions.

Figure 35.2, for instance, illustrates that the negative-influence assumption changes the predictions of influence models. General theories from social psychology and sociology explain why individuals can be negatively influenced by interaction partners, but these theories also imply that negative influence results only under certain conditions. Social-identity theory, for instance, implies that the negative influence is only activated when the individual and the source of influence belong to different social categories that the individual considers salient (Hogg et al. 1990; Tajfel and Turner 1986). Sociological theories, in contrast, imply that negative influence is triggered by status differences between individuals (Bourdieu 1984; Elias 1969). In particular, individuals belonging to high classes are assumed to distance themselves from low-class individuals. These low-class individuals, in contrast, are assumed to be positively influenced by higher classes, according to these sociological theories.

Efforts to validate model assumptions can profit from a better theoretical and empirical foundation of theoretical assumptions, because it informs about the conditions under which assumptions should be tested. The negative-influence assumption, for instance, should be tested in settings where theories predict that such forms of influence actually exist. An interesting candidate, for instance, would be a school class characterized by high ethnic diversity and status differences. In contrast, negative influence may be very unlikely when individuals do not communicate face-to-face as students do in a class room, but in a computer-mediated setting like a comment board on the Internet. Social psychological research suggests that computer-mediated communication is less affected by individual's physical appearance as individuals are not informed about demographic differences (Postmes et al. 2001).

The flexibility that computer simulation creates for modelers implies that the theoretical consequences of various assumptions can be explored with relatively little effort. This is a great advantage of computer simulation, but it should not lead researchers to implement assumptions without a solid theoretical foundation.

35.4.3 Explore Model Scope and Its Boundaries!

Every theory has a scope, a domain of contexts where it can be applied and where its predictions can be put to the test. Above, I argued that modelers should clearly define the scope of their model in order to inform empirical research about the social contexts in which their model can be tested, because theories can only be falsified with data gathered in contexts covered by the theory's scope (see also Chap. 6 by Beven and Lane in this volume).

In addition, however, empirical research should also explore the boundaries of a model's scope, testing whether the scope of a theory may be bigger than expected. This is often possible without high costs. Above I have argued for instance that diverse school classes are a promising setting to test the negative-influence model. Furthermore, I argued in Sect. 35.3.1 that negative influence is more plausible when individuals influence each other's opinions rather than their beliefs. This, suggests that the dynamics of music tastes in a diverse school class are within the scope of the negative-influence model (Lewis et al. 2012). Nevertheless, it is worthwhile to test whether the negative-influence model also predicts the dynamics of students' beliefs and behavior rather than only opinions. Finding that the negative influence makes accurate predictions also in contexts that are outside of the model's scope will guide the development of more general models.

35.4.4 More Validation!

The described challenges to the validation of formal models in the social sciences are fundamental. Since these challenges also often act in tandem, there is certainly no simple solution. Nevertheless, the approach with the highest potential to eventually lead to the development of standard and accepted models in the social sciences is to intensify validation efforts.

Diversification of empirical approaches appears to be promising. First, empirical research should test model assumptions and predictions in various settings, in order to explore which model is supported under certain conditions. The higher the number of empirical studies that do not support a given model and the higher the number of different contexts that these studies explored, the more confidently one can consider a model false. Second, researchers should apply alternative empirical methods to measure social-scientific concepts (e.g., opinions), in order to exploit their complementary advantages. Third, researchers should also make more use of the huge arsenal of empirical methods that the social sciences provide. For instance, large-scale representative surveys allow to reliably describe opinion distributions and their dynamics. Controlled experiments in the laboratory, in contrast, make it possible to directly test model assumptions about the opinion changes resulting from social influence. Furthermore, the Internet provides new opportunities to gather detailed observational data about human behavior and interaction. The emerging field of computational social science applies methods developed in physics, mathematics, and computer science to gather, manage, and analyze this information (Lazer et al. 2009; Conte et al. 2012; Salganik 2017; Golder and Macy 2014). The social sciences will profit greatly from these empirical efforts when they are systematically targeted at those contexts, model assumptions, and predictions where models disagree.

References

Abelson, R. P. (1964). Mathematical models of the distribution of attitudes under controversy. In N. Frederiksen & H. Gulliksen (Eds.), *Contributions to mathematical psychology* (pp. 142–160). New York: Rinehart Winston.

Ahrweiler, P., & Gilbert, N. (2015). The quality of social simulation: An example from research policy modelling. In M. Janssen, M. Wimmer, & A. Deljoo (Eds.), *Policy practice and digital science* (pp. 35–55). Heidelberg/New York: Springer.

Axelrod, R. (1997). The dissemination of culture-A model with local convergence and global polarization. *Journal of Conflict Resolution, 41*(2), 203–226.

Baldassarri, D., & Bearman, P. (2007). Dynamics of political polarization. *American Sociological Review, 72*(5), 784–811.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural-change as informational cascades. *Journal of Political Economy, 100*(5), 992–1026.

Bonacich, P., & Philip, L. (2012). *Introduction to mathematical sociology*. Princeton and Oxford: Princeton University Press.

Boudon, R. (1981). *The logic of social action. An introduction to sociological analysis*. London: Routledge and Kegan Paul.

Bourdieu, P. (1984). *Distinction: A social critique of the judgment of taste*. Cambridge, MA: Harvard University Press.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences of the United States of America, 114*(40), 10612–10617.

Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., & Flocken, C. (2016). Disambiguation of social polarization concepts and measures. *The Journal of Mathematical Sociology*.

Brewer, M. B. (1991). The social self-On being the same and different at the same time. *Personality and Social Psychology Bulletin, 17*(5), 475–482.

Brousmiche, K.-L., Kant, J.-D., Sabouret, N., & Prenot-Guinard, F. (2016). From beliefs to attitudes: Polias, a model of attitude dynamics based on cognitive modeling and field data. *Journal of Artificial Societies and Social Simulation*, *19*(4).

Byrne, D. (1971). *The attraction paradigm*. New York, London: Academic Press.

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics, 81*(2), 591–646.

Chattoe-Brown, E. (2014). Using agent based modelling to integrate data on attitude change. *Sociological Research Online*, *19*(1).

Clemm von Hohenberg, B., Mäs, M., & Pradelski, B. S. R. (2017). Micro influence and macro dynamics of opinion formation. SSRN. https://ssrn.com/abstract=2974413.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., et al. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics, 214,* 325–346.

David, N. (2009). Validation and verification in social simulation: Patterns and clarification of terminology. In F. Squazzoni (Ed.), *EPOS 2006: Epistemological Aspects of Computer Simulation in the Social Sciences* (pp. 117–129). Berlin: Springer.

Deffuant, G., Huet, S., & Amblard, F. (2005). An individual-based model of innovation diffusion mixing social value and individual benefit. *American Journal of Sociology, 110*(4), 1041–1069.

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have Americans' social attitudes become more polarized? *American Journal of Sociology, 102*(3), 690–755.

Elias, N. (1969). The civilizing process. In *The history of manners* (Vol. I). Oxford: Blackwell.

Esteban, J. M., & Ray, D. (1994). On the measurement of polarization. *Econometrica, 62*(4), 819–851.

Everett, J. A. C. (2013). The 12 item social and economic conservatism Scale (SECS). *PLoS One*, *8*(12), e82131. (Edited by Pete Roma. Public Library of Science).

Flache, A., & Mäs, M. (2008a). Why do faultlines matter? A computational model of how strong demographic faultlines undermine team cohesion. *Simulation Modelling Practice and Theory, 16*(2), 175–191.

Flache, A., & Mäs, M. (2008b). How to get the timing right. A computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams. *Computational and Mathematical Organization Theory, 14*(1), 23–51.

Flache, A., & Macy, M. W. (2011a). Small worlds and cultural polarization. *The Journal of Mathematical Sociology, 35*(1–3), 146–176.

Flache, A., & Macy, M. W. (2011b). Local convergence and global diversity: From interpersonal to social influence. *Journal of Conflict Resolution, 55*(6), 970–995.

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, *20*(4).

Friedkin, N. E., & Johnsen, E. C. (2011). *Social influence network theory*. New York: Cambridge University Press.

Garrison, L. A., & Babcock, D. S. (2009). Alcohol consumption among college students: An agent-based computational simulation. *Complexity*, *14*(6), 35–44 (Wiley Subscription Services, Inc., A Wiley Company).

Gilbert, N., & Troitzsch, K. G. (1999). *Simulation for the social scientist*. Buckingham Philadelphia: Open University Press.

Gilbert, N., & Ahrweiler, P. (2005). Caffè Nero: The evaluation of social simulation. *Journal of Artificial Societies and Social Simulation*, *8*(4).

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology, 40*(1), 129–152.

Grow, A., & Flache, A. (2011). How attitude certainty tempers the effects of faultlines in demographically diverse teams. *Computational and Mathematical Organization Theory, 17*(2), 196–224.

Hedström, P., & Udehn, L. (2009). Analytical sociology and theories of the middle range. In P. Hedström, & P. Bearman (Eds.), *The oxford handbook* (pp. 25–47). Oxford University Press.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology, 36*(1), 49–67.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, *5*(3).

Hegselmann, R., Schelling, T. C., & Sakoda, J. M. (2017). The intellectual, technical, and social history of a model. *Journal of Artificial Societies and Social Simulation (JASSS)*, *20*(3), 15.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15,* 135–175.

Hogg, M. A., Turner, J. C., & Davidson, B. (1990). Polarized norms and social frames of reference: A test of the self-categorization theory of group polarization. *Basic and Applied Social Psychology, 11*(1), 77–100.

Huckfeldt, R., Johnson, P. E., & Sprague, J. (2004). *Political disagreement. The survival of diverse opinions within communication networks*. Cambridge University Press.

Iyengar, S., & Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication, 59*(1), 19–39.

Keijzer, M. A., Mäs, M., & Flache, A. (2018). Online social networks foster cultural isolation. Groningen.

Lau, D. C., & Keith Murnighan, J. (2005). Interactions within groups and subgroups: The effects of demographic faultlines. *Academy of Management Journal*, *48*(4), 645–659.

Lazarsfeld, P. F., & Merton, R. K. (1954). Friendship and social process: A substantive and methodological analysis. In M. Berger, T. Abel, & C. H. Page (Eds.), *Freedom and Control in Modern Society* (pp. 18–66). New York, Toronto, London: Van Nostrand.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., et al. (2009). Computational social science. *Science, 323*(5915), 721–723.

Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences of the United States of America, 109*(1), 68–72.

Liu, C. C., & Srivastava, S. B. (2015). Pulling closer and moving apart: Interaction, identity, and influence in the U.S. Senate, 1973 to 2009. *American Sociological Review, 80*(1), 192–217.

Loehlin, J. C., & Alexander Beaujean, A. (2016). *Latent variable models: An introduction to factor, path, and structural equation analysis* (5th ed.). New York and London: Taylor & Francis.

Lorenz, J. (2005). A stabilization theorem for dynamics of continuous opinions. *Physica a-Statistical Mechanics and Its Applications, 355*(1), 217–223.

Lorenz, J. (2007). Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C, 18*(12), 1819–1838.

Lorenz, J. (2017). Modeling the evolution of ideological landscapes through opinion dynamics. In W. Jager, R. Verbrugge, A. Flache, G. de Roo, L. Hoogduin, & C. Hemelrijk (Eds.), *Advances in social simulation 2015* (pp. 255–266). Springer International Publishing.

Macy, M. W., Kitts, J., Flache, A., & Benard, S. (2003). Polarization and dynamic networks. A hopfield model of emergent structure. In R. Breiger, K. Carley, & P. Pattison (Eds.), *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers* (pp. 162–173). Washington, DC: The National Academies Press.

Mark, N. P. (2003). Culture and competition: Homophily and distancing explanations for cultural niches. *American Sociological Review, 68*(3), 319–345.

Marsden, P. V., & Friedkin, N. E. (1993). Network studies of social-influence. *Sociological Methods & Research, 22*(1), 127–151.

Mäs, M., & Flache, A. (2013). Differentiation without distancing. Explaining opinion bi-polarization without assuming negative influence. *Plos One* 8(11).

Mäs, M., & Bischofberger, L. (2015). Will the personalization of online social networks foster opinion polarization. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2553436.

Mäs, M., Flache, A., & Helbing, D. (2010). Individualization as driving force of clustering phenomena in humans. *PLoS Computational Biology, 6*(10), e1000959.

Mäs, M., Flache, A., Takács, K., & Jehn, K. A. (2013). In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization. *Organization Science*, *24*(3), 716–736.

Mäs, M., Flache, A., & Kitts, J. A. (2014). cultural integration and differentiation in groups and organizations. In V. Dignum & F. Dignum (Eds.), *Perspectives on culture and agent-based simulations*. Cham: Springer.

Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review, 11*(3), 279–300.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27,* 415–444.

Merton, R. K. (1957). *Social theory and social structure*. Ill: Glencoe.

Nagel, E. (1979). *The structure of science: Problems in the logic of scientific explanation*. Indianapolis, Cambridge: Hackett.

Opp, K. D. (2009). *Theories of political protest and social movements: A multidisciplinary introduction, critique, and synthesis*. New York: Routledge.

Opp, K. D., & Gern, C. (1993). Dissident groups, personal networks, and spontaneous cooperation-The East-German revolution of 1989. *American Sociological Review, 58*(5), 659–680.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Postmes, T., Spears, R., Sakhel, K., & De Groot, D. (2001). Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin, 27*(10), 1243–1254.

Rosenbaum, M. E. (1986). The repulsion hypothesis: On the nondevelopment of relationships. *Journal of Personality and Social Psychology, 51*(6), 1156–1166.

Sakoda, J. M. (1971). The checkerboard model of social interaction. *The Journal of Mathematical Sociology*, *1*(1), 119–132 (Taylor & Francis Group).

Salganik, M. J. (2017). Bit by bit: Social research in the digital age.

Salzarulo, L. (2006). A continuous opinion dynamics model based on the principle of meta-contrast. *Journal of Artificial Societies and Social Simulation*, *9*(1).

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology, 1,* 143–186.

Slater, M. D. (2007). Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication Theory, 17*(3), 281–303.

Steglich, C., Snijders, T. A. B., & Pearson, M. (2010). Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology, 40*(1), 329–393.

Tajfel, H. (1978). Social categorization, social identity and social comparison. In *Differentiation between social groups: Studies in the social psychology of intergroup relations*.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel, & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Chicago: Nelson-Hall Publishers.

Takács, K., Flache, A., & Mäs, M. (2016). Discrepancy and disliking do not induce negative opinion shifts. *PLoS One, 11*(6), e0157948.

Turner, J. H. (1974). *The structure of sociological theory*. Chicago, Illinois: The Dorsey Press.

Turner, J. H. (1995). *Macrodynamics. Toward a theory on the organization of human populations*. New Brunswick, N.J: Rutgers University Press.

Wimmer, A., & Lewis, K. (2010). Beyond and below racial homophily: ERG models of a friendship network documented on Facebook. *American Journal of Sociology, 116*(2), 583–642.

Chapter 36
Validation and the Uniqueness of Historical Events

Josef Köstlbauer

Abstract Historians have been slow to include computer simulations into their discipline's methodological apparatus. This chapter details the challenges faced when trying to employ simulations for historical research. Central to this is the idiographic character of historical research, which leads to problems regarding computer simulations and validation. Historians are concerned with the unique, with distinct historical processes, whose ultimate result is known. They do not formulate general laws or rely on deductive-nomological approaches. But this should not keep historians from exploring the potentials of computer simulations to the full extent: Big-data projects may help to dissolve the nomothetic-idiographic divide, microhistorical research may profit from simulations for contextualization or to compensate for fragmentary sources. In all cases, validation has the potential to make historians reflect more on evaluative assumptions, and on the ways, they pose questions and explain processes.

Keywords Simulation · Computer Simulation · History · Humanities · Big Data in Historical Research · Simulation in History · Validation · Video Games and History

History and computer simulations make strange bedfellows; they even may seem to be antithetical. Simulations are about creating models of a system and tinkering with variables to see what happens. History, on the other hand, seems to forbid such tinkering: Everything has happened already, the past remains a forbidden land. There is nothing to meddle with. Or is there?

Before proceeding further, I need to caution readers that the author himself is a historian. While I have done research on digital games and participated in research projects which employed digital tools and created digital content, my knowledge of quantitative methods and their mathematical prerequisites is much inferior to that of any trained sociologist—as is my knowledge of designing and employing computer simulations, including the problems of validation contained therein. So the following is not a sociologist's perspective on matters of historiography, instead

J. Köstlbauer (✉)
History Department, University of Bremen, Bremen, Germany
e-mail: josef.koestlbauer@uni-bremen.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_36

it is a perspective from the inside of the discipline. To the knowledgeable reader, this will be apparent in the language and terminology used. By necessity the historian's perspective on computer simulations and validation is interdisciplinary: input and insights from games studies or cultural studies are as important as from the social sciences. By presenting perspectives and methodologies of history, I hope to facilitate comparison and contribute to interdisciplinary cooperation.

In the following pages, I will present some observations on the problems and potentials associated with the use of computer simulations and validation in historiography. Doing this, I follow a very basic definition of simulation as the use of a formal model of reality to imitate processes taking place in a system. According to game theorist Frasca (2003, p. 223), to simulate "is to model a (source) system through a different system which maintains (for somebody) some of the behaviors of the original system." Philosopher Hartmann (2011, p. 83) phrases it more abstractly, "a simulation results when the equations of the underlying dynamic model are solved." Historiography usually is based on the presumption that the world is something tangibly real (leaving aside radically constructivist notions), which can be ascertained through historical evidence. This reality can be understood as a system and represented in a model. A model may be classified as "a set of assumptions about a system" (Redhead 1980, p. 146), with a dynamic model including assumptions of time-evolution (Hartmann 2011, p. 82). The latter seems central for the simulation of processes in history, but there might also be problems where static models may be viable.

Validation is understood as an iterative process integral to model design and simulation. It includes validation of assumptions guiding model design as well as the problem of ascertaining the fidelity of a model or establishing deviations. Validation poses particular challenges to the discipline of history, which has a strong idiographic orientation.

Being a historian, I feel my discipline can add to a volume on computer simulation validation by adding a bona fide historical perspective. Therefore, the first chapter contains a brief historical reflection. Since a summary of the history of simulation would far exceed the bounds of this essay, I kept myself to a glimpse on the historical semantics of simulation. The intention is to demonstrate that simulation is a transient term which has moved through different, if related, meanings and has been used in a range of contexts.

The second chapter provides a brief introduction to history's methodological dispositions and traditions. Subsequent observations in chapter two deal with the question of why there is so little interest in models and simulation in history. The third chapter discusses the methodological challenges involved with using models and simulation in history and the concomitant problems of validation. The fourth chapter is about uses and potentials of simulations in history. This chapter also includes simulations devised as a heuristic tool as well as simulation games developed outside of academia. Together these three chapters hopefully will give an idea of major aspects affecting the employment of simulations in historical research and the challenges involved in validation.

36.1 A Brief History of Simulation's Semantics

Upon close inspection, the definition of the term simulation, like many supposedly straightforward terms, turns out to be surprisingly slippery. Although in this particular case this is absolutely fitting; deriving from Latin verb "simulo", to simulate is to emulate, to pretend, to fake, to make believe in English as well as in French. This is the original meaning of the term but looking up the noun "Simulation" in current encyclopedias, like English and German Wikipedia or the Encyclopaedia Britannica, turns up definitions that broadly define simulation as a scientific technique to analyze systems. Simulation in this sense is a heuristic method. Usually, references to the use of simulations and simulators for training, education, and entertainment are included, too.

So an inquiry into the history of the term and its meanings reveals striking changes over time. In early modern Europe, there was a fascinating discourse on simulation/dissimulation as techniques employed by statesmen and courtiers (Castiglione 2007; Gracián 1993). Partly philosophical, partly didactic, it sought to explain as well as to publicize the culture of secrecy and disguise, which was perceived as permeating European courts (Snyder 2009). Due to the ascendancy of new culture espousing enlightened civic virtues (Hazard 2013) the term "simulation" increasingly acquired a strong negative bias, becoming the opposite of truthfulness and close to hypocrisy (Zedler 1746; Crabb 1824).

During the nineteenth century, the term simulation entered into the vocabularies of medicine and psychology. Acquiring a scientific veneer, the term nevertheless retained a negative connotation, most tangibly expressed in the German language. There the "Simulant" is a malingerer, who uses simulation as a strategy trying to avoid military service or work. That usage of the term acquired unprecedented currency during World War I, when European states frantically tried to mobilize their populations for warfare (War Office Committee 2004, pp. 141–144; Hirschfeld et al. 2009, p. 216; Michl 2007, pp. 185–194, 218).

A cursory inspection of printed English and German publications between 1700 and 2000 using the Google Ngram Viewer shows a substantial increase in the usage of the term simulation from the middle of the twentieth century onwards. What is the reason for this? It can be interpreted as indicative of both the diffusion of applied mathematics and the onset of the digital age. First, during the twentieth century, the language of mathematics and science completely usurped the term simulation and superseded its former meanings. Subsequently, with the dissemination of personal computers, the profusion of digital technologies, and the rise of digital games to mass media, the term became ubiquitous. As part of an emergent language of computers, digitality, and virtuality, gained a new semantic field and entered popular culture. The former meanings may not have ceased, but undoubtedly they have been relegated to a peripheral position. Today it is unimaginable to think of simulations without thinking of computers, too. Telling examples are the encyclopedia entries mentioned above. This also signifies that a very profound change in the term's meaning has taken place. No longer is simulation tainted by negative connotations. It has become a technical or methodological term.

This seemingly inevitable connection between computers and simulation is misleading insofar as computers are not necessary for conducting simulations. For example, one of the most famous simulation models, economist Thomas Schelling's agent-based model of racial segregation from 1971 was first presented using nickels and dimes on a checkerboard (Schelling 1978). Also, simulation games are still being published on paper (Corbeil 2011, p. 421). This book is on computer simulations; nevertheless, some observations will apply to models and simulations in the broadest sense.

There is an ironic twist to simulation's evolution has taken from art of courtiers to morally dubious pretense to supposedly neutral semantics of technology and science when considered in the context of our increasingly digitized world. After all, our mundane interactions with digital media are made possible only by increasingly powerful techniques of simulation, disguise, and pretense: The graphical user interface, which has been one of the prime enabling factor in turning computers from specialist scientific and military equipment into ubiquitous companions, whose presence pervades almost all aspects of our professional and private lives. Today's glossy interfaces imbue visualized abstractions like desktop icons with a treacherous aspect of materiality. The shiny arrays of virtual buttons, sliders, and spaces niftily conceal the structures and ontologies of the software, the workings of algorithms, and the banal flip-flop/yes–no of energy levels. Therefore, one might argue that nowadays simulation and dissimulation form the very essence of the computer. The interface has become the place of the cyborg, fusing man and machine, software and brain. These instances of simulation are rarely reflected upon since they have become such integral parts of our lives (Köstlbauer 2013).

What needs to be kept in mind is that terms we use so were understood quite differently a mere hundred years ago. The question of why names and meanings change points us to problems of historical continuity and discontinuity, to shifting paradigms and perspectives.

36.2 History

The eminent American historian Bernard Bailyn once described history as "*sometimes* being an art, *never* a science, *always* a craft" (italics in the original, Ekirch 1994, p. 636). Historians establish historical facts through careful research of sources. Historical sources are all documents or objects that have significance for a historical research issue. As such their nature is not predetermined and for example can encompass archival documents as well as movies, paintings, architecture, digital games, or Instagram accounts. However, sources are not mirrors of the past. Instead, they are products of interests, ambitions, ideas, and perceptions of the individuals and societies that produced them. Sources are also media, and thus products of technology and practices evolving around them, like writing, narrating, painting, filing, archiving, etc. So the information on events, persons, and places contained within sources gets further distorted and warped. Therefore, the discipline's craft, its quintessential and most rigorous methodologies, is about analyzing, contextualizing, and interpreting sources. Based on the findings further hypotheses and questions can be formulated (Schmale 2006, p. 17–18; Sellin 1995, p. 17).

Furthermore, sources are incomplete: The set of sources we use to reconstruct the past is the result of the vagaries of chance. Decay, war, fire, and floods, negligence, and ignorance are constantly and successfully conspiring against historians and making historical inquiry ever more difficult and haphazard. Therefore, the identification of lacunae in sources is of central importance for historical contextualization.

The historian's work can also be described as a medial operation: History as an academic discipline is very much a child of the Gutenberg Galaxy (McLuhan 1962). It is profoundly shaped by the forms and logic of typographic media, which accompanied the formation of historiography as an academic discipline since the nineteenth century. It tells something about the persistence of such medial formations that even massive changes in technology–like the one experienced today–seem to cause little disruption. Historians are writers, and a sizable part of them will remain so for the foreseeable future—no matter how much number crunching their actual work contains and notwithstanding the digital media environs they operate in today. Symptomatic is a statement by historian William Turkel: "To some extent we're all digital historians already, as it is quickly becoming impossible to imagine doing historical research without making use of e-mail, discussion lists, word processors, search engines, bibliographical databases and electronic publishing." (Turkel 2008) One might say that historians have adapted to technological change without changing the stance and setup of historical research. So far even the new field of digital humanities has done little to alter that.

So history still is historiography–the writing of history–and language both spoken and written is its premier medium. The way to a Ph.D., to tenure, influence and, maybe, fame leads across the pages of books, articles, and essays. Emergent disruptive transformations in the academic publishing system like open access publishing are not necessarily going to alter the importance of writing. A well established academic discipline like history has a lot of inertia. Therein lays one of the reasons why simulations are not part of historian's standard toolbox. When publishing books and articles is rewarded, it is risky to concentrate on devising or learning new methodologies. It is also harder to acquire funding, and it will be pretty lonely as long as there is no scholarly community within the discipline to discuss problems and results. But it is entirely possible to imagine other ways to study history, some of these will be addressed on the following pages.

Another reason why simulations are rarely employed by historians has to do with the way historians regard their discipline's subject. Like other academic disciplines history attempts to explain the world. Its interpretations of the past are based on the basic premise of a chain of causality tying the "Then" to the "Now". This does not mean that Leopold von Ranke's famous dictum still holds true, which says that history is about finding out "what actually happened" ("wie es eigentlich gewesen ist").[1] Nowadays, it is well understood that the past remains elusive and is only partially accessible to reconstruction and interpretation. What we see is distorted by the bias and incompleteness of sources used. History remains by necessity a fragmentary and preliminary narrative. Also, the importance of causality does not translate into bland determinism. History always is open: historical events are not predetermined but essentially contingent (Koselleck 1989).

Sources and the methodologies of interpreting them are central to the self-conception of the discipline. It causes a profoundly idiographic approach: Historians identify sources, which they treat as unique manifestations of a distinct past, and which lead to unique, distinct historical events. New historical knowledge is perceived to be gained by uncovering new sources, very much like archeologists excavating new objects, or by researching topics that have not yet been the subject of historical inquiry. This stance does not translate into ignorance regarding theory and methodology; indeed there is a heterogeneous multitude of methods and approaches. (Lengwiler 2011; Herbst 2004). But it is not conducive to research aiming to formulate general theories and to devise models for analyzing hypotheses.

36.3 Challenges

In their "History Manifesto", Historians Jo Guldi and David Armitage asserted that "the world around us is clearly one of change, irreducible to models." (Guldi and Armitage 2014, p. 3) This strong statement is rather astonishing as both authors are quite open to methods and tools of the digital humanities. Of course, the world can be represented in models. But the idiographic stance characterizing history leads to problems regarding computer simulations and validation. Some areas in the social sciences are comfortable with deductive-nomological approaches, in which general laws (explanans) provide the explanation of a phenomenon or event (explanandum), the latter being logically deducible from the first (see Chap. 35 by Mäs in this volume). In history, on the other hand, the result of historical processes is already known. Now, one could produce a deterministic model of the world, which always provides the same dynamics as long as the same conditions remain in place. Running simulations, one could observe which conditions replicate the known historical results. In such a case validation is the process of observing results and tweaking the model. It would be a more inductive-statistical approach, which analyzes the probabilities of the assumptions expressed in the model. Such simulations are of limited use when causal connections within a historical event are known in detail, but validation would be relatively straightforward, essentially being achieved by comparing the simulation to the established chain of events. They also may be helpful if there are lacunae which

[1] Franz Leopold von Ranke (1795–886) is considered one of the founding figures of modern historiography. He coined the famous phrase regarding the task of historiography, "zu zeigen wie es eigentlich gewesen ist" ("To show what actually was").

cannot be filled through more orthodox historical research. This would be a way of complementing poor data. One problem of validation is that results only provide information on whether the assumptions of the model can give explanations for a historical phenomenon or not. They do not provide proof. Still, simulations may point researchers to new avenues of research or aid them in reframing research questions.

An example case is the “Simulation on War, space, and the evolution of Old World complex societies”, an agent-based, spatially and temporally oriented simulation. The researchers modeled “cultural evolution mechanisms” to examine variables governing the rise of “large-scale complex societies” throughout an extraordinarily long period, from 1500 BCE to 1500 CE (Turchin et al. 2013). The fate of the relevant societies is well established, at least as far as rise and decline are concerned. While the researchers involved do seek a better understanding of historical processes, they operate within the methodological framework of mathematics, anthropology, and evolutionary biology. The label applied to this specific area of transdisciplinary research is Cliodynamics. But the interest is more anthropological and sociological than historical; the authors emphatically assert that they are trying to describe general principles at play in the *longue durée* processes. Thus, they are not primarily interested in the societies whose development is being modeled; rather they are interested in analyzing large-scale societies per se. Klaus Troitzsch (1994, pp. 41–44) pointed out that many other simulations devised to simulate prolonged periods of social evolution containing a historical aspect are more sociological in outlook than historical. Of course, that does not mean that historiography is not going to profit from such efforts. But without rejecting other disciplines interest general theories or their viability, historians usually endeavor to explain historical processes through distinct historical conditions.

Interesting to the historian is the researchers’ reference to established historical knowledge both by framing their hypotheses as well as by validating simulation results in the abovementioned simulation project. Validation here means that the results conform to a significant percentage of historical evidence (Turchin et al. 2013). But is that sort of validation satisfying? There seem to be certain risks, for example privileging preestablished conceptions. Validation, therefore, needs to encompass the basic assumptions of the model, too. The study’s authors based their model on the premise that the institutions needed to enable the existence of large groups evolved as a consequence of “intense competition, primarily warfare” (Turchin et al. 2013, p. 16384). This assumption seems rather simplistic. Further assumptions regarding the importance of certain military technologies like horse chariots or cavalry equipment seem equally questionable when considering the findings of experimental archeology (see for example Sidnell 2006).

Calling into question the basic historical assumption used in the creation of the model also calls into question the usability of results. Does it still demonstrate the significance of warfare of large complex societies? Does it show something else? Such speculation may lead to outright dismissal, but at the same time, it can be quite helpful as it inspires new interpretations. The simulation is not going to tell us what happened or why it happened. It does facilitate the analysis of hypotheses regarding the long-term significance of large-scale conflict and the impact of technologies, thereby potentially aiding researchers of various disciplines in identifying new avenues of research. Thus, it conforms to two functions of simulation detailed by philosopher Stephan Hartmann: simulation is a technique of research that makes it possible to investigate the dynamics of a process or system in detail, and it is a heuristic tool that allows one to develop hypotheses, theories and models (Hartmann 2011, p. 85; Troitzsch 1994).

Validation of such simulations also entails thinking about complexity. Refining simple but easy to work with models will create layers of complexity of models and make it increasingly difficult to comprehend. Transparency of models is an important prerequisite for validation (Hartmann 2011, pp. 108–112; Turchin et al. 2014).

Deterministic models might come into their own when the historical outcome of a specific situation is not known. For example, this would be true for simulations used for predicting future developments (Murauer 2014, p. 15). But it is unclear how this simulation might be validated. And while there always have been historians, who unabashedly asserted the prognostic powers of history (Geiss 1998, p. 18–24), to me, the discipline seems neither conceptually nor methodologically equipped to deal with prognostics.

More difficult to hypothesize is the use of stochastic models, in which the basic assumptions change with each iteration of a simulation and produce a different outcome each time. How might the results be validated? The outcome of history is evident (facts); simulation results which contradict the actual outcomes of historical developments can hardly be regarded as valid.

Data poses another problem. Few historians regard computable data a central result of their research. Mostly this is a specialty of economic and social history, where sources like church registers, toll registers, shipping lists, or census list are being used. These are also subdisciplines which have a long tradition of employing the methods of sociology or economics. Network analysis, too, produces computable data, for example, on familial, institutional, or correspondence networks. Historical network analysis also provides examples of the challenges posed by historical data sets: often, they are uneven, and the unevenness may be unsystematic, making it difficult to deal with (Meeks 2015).

Typically data assembled by historians during research in archives and the evaluation of literature is used toward analyzing and explaining causal relationships. It is not necessarily regarded as a product of research which merits dissemination and being shared within the community. As of yet, there is no widespread culture of publishing and sharing datasets which in turn might foster the formulation of hypotheses analyzed through models and simulation. This may be expected to change as the interdisciplinary field of digital humanities is creating new possibilities and tools for research and increasing the number of researchers.

It is telling that a vast historical database project is being developed outside of the discipline of history: The Seshat project, named after an Egyptian goddess, tries to harness big data to quantify historical processes. Founded in 2011, it is financed both by public and private institutions and claims no less than “to bring together the most current and comprehensive body of knowledge about human history in one place” (Seshat 2017). Leaving aside the slightly megalomaniac tone of this statement, the generation of huge databases has its merits, not only for historians but also for scholars from other disciplines and it undoubtedly can engender interdisciplinary cooperation. Seshat is created with simulations in mind (Turchin et al. 2015).

The enthusiasm concerning "big data" in business as well as in academic research has kindled renewed interest in quantitative work, database design, and in models.[2] A remarkable project directed by art historian Maximilian Schich extracted spatiotemporal information on "notable individuals" from various existing datasets, spanning a period of two millennia. This data was used to construct a worldwide historical migratory network and to study statistical patterns. The project intended to create a macroscopic perspective of cultural history based on quantitative methods and network theory (Schich et al. 2014).

Such projects also indicate a requirement for the sort of specialist knowledge possessed by trained historians. Historical sources provide peculiar pitfalls: Throughout history, archives were purged, documents forged, biographical dates lied about, etc. Rarely can biographical data be taken at face value. Datasets that are easy to deal with, like, for example, birth and death records, tend to have limited informative value (for examples regarding network analysis see Meeks 2015).

Simulations also are intricately linked to the problem of counterfactual history (Bunzl 2004; Demandt 2011). Any simulation of counterfactuality, in the end, is running counter to historical factuality aspired to by historians. Nevertheless counterfactual musings not only have a kind of persistent allure, which has found expression in popular culture (Rodiek 1997; Brendel 2010), they also are hard to avoid for historians (Bunzl 2004, p. 846). A century ago Max Weber pointed out that every assumption on the causal significance of an event or actor presupposes counterfactual musings (Herbst 2004, p, 75; Nonn and Winnerling 2017, p. 8).

But the counterfactual perspective inherent in historical interpretation does not necessarily make "what if…?" a valid *modus operandi* of historical research. To historians more important than to explore alternative futures is to understand "past futures", meaning the range of possibilities perceived and imagined by contemporaries (Demandt 2011; Koselleck 1989). Simulations may aid us in comprehending such possibility-spaces and the ways they closed or opened. But they need clear delineations and definitions, and they need to be supported by concomitant reflection on the scientific method, model building, and interpreting results. Exploring the spectrum of possibilities faced or imagined by historical actors is an important way of analyzing historical perceptions, hopes, and fears (Bunzl 2004, p. 848). Remarkably there is little public recognizance of this fact in German historiography. Christoph Nonn and Tobias Winnerling recently pointed out that there was no major discussion of counterfactuality since the 1980s. The situation is different in Anglo-Saxon academia, where there is a longer tradition of pursuing such questions (Nonn and Winnerling 2017, pp. 9–12). But even there, counterfactuality is far from being a central part of historian's methodological apparatus.

[2] For a detailed analysis and several examples see Guldi and Armitage (2014, pp. 88–116, 151), Porsdam (2011, pp. 2–7).

36.4 Uses and Potentials of Simulations in History

Some scholars, like Schich (2016), argue that the employment of what he dubs "natural science methods" within disciplines like art history might overcome the separation between nomothetic and idiographic disciplines described by Wilhelm Windelband in 1894. Windelband himself pointed out the dependency of idiographic disciplines on general laws formulated by the nomothetic sciences (Windelband 1915). To bring history and simulations together there seem to be two ways: either identify the nomothetic strands within the discipline (for example, in economic history or quantitative social history) or set up simulations in a way that makes them useful to idiographic explanation.

36.4.1 Big-Data and Longue Durée History

Big-data projects, like those introduced in Chap. 3 offer new possibilities of analyzing evolutionary processes over very long time periods and for testing hypotheses by running simulations. Taking up a phrase coined by famous French historian Fernand Braudel, some proponents speak of a new history of the *longue durée* (François et al. 2016, para. 4–6). Others fervently hope for big-data making possible the escape from the curse of "short-termism" (Guldi and Armitage 2014, pp. 1–10). That is studying persistent or very slowly changing structures, which evolve largely untouched by the outcomes of wars or the actions of individuals (being the *pouvoir* of the *histoire événementielle*) (Braudel and Colin 1987; Braudel 1990).

The most intriguing promise of big-data analysis lies in what Manovich (2017, pp. 60–61) has termed *Cultural Analytics*. Its vision is "to describe, model, and simulate the global cultural universe". When it becomes possible to analyze "everything by everyone" (Manovich 2017, p. 61), bias rooted in canon, tradition, or other more or less arbitrary or nonscientific qualitative categorization can be transcended. The result is a new ability to distinguish the exceptional from the general. Thus, the nomothetic concern with the general increases the ability to identify and subsequently research the particular (Manovich 2017, p. 62–63; see also Manovich et al. 2014).

Simulations of the big-data/*longue durée* type are operating on a macro-level, and they are often initiated in other disciplines than history. Micro-level research using simulations is more unusual to come by. But simulations potentially could bridge the macro–micro gap which has become recognized as a central methodological and theoretic problem of the historical discipline (Herbst 2004, p. 17). Microhistory partly was a reaction to social history in the 1970s and 1980s. Partly it was a continuation of a tradition of the Annales school of history (Herbst 2004, pp. 192–194). Well-known representatives are Carlo Ginzburg, Natalie Zemon Davies, and Robert Darnton. Microhistory is concerned with reconstructing the lives of individuals and communities. "Micro" indicates the microscopic perspective, the attempt to get at the details of everyday life. Microhistorical studies delve deeply into archival material and have proven effective in demonstrating how religious beliefs, trade, literacy, etc., influenced the lives of specific individuals and communities. While ostensibly denouncing grand theories, microhistorians developed quite a sophisticated methodology, allowing them to deconstruct discourses and use sources in innovative ways. In doing this it contributes to our understanding of the complexity of historical societies, and it adds variables which have to be taken account of when doing macrohistory.

36.4.2 Microhistorical Research and Simulation

At the same time, this means that microhistory lays bare the often hidden general assumptions at work in a seemingly idiographic discipline like history. Such general assumptions about, for example, early modern estate based society in Western Europe are by no means arbitrary and rest on the knowledge of many cases. Also, they are generally acknowledged to be broad approximations or representations of general characteristics. Nevertheless, this means that there are certain models (however, ill-defined) influencing historical research and the discipline could do worse than to carefully reframe such assumptions (and the attendant nomenclature) (for the twentieth century debate on the problem see Herbst 2004, pp. 68–70).

At first glance, microhistorical research and its methodologies do not lend itself to simulation easily. But given microhistory's sense for lacunae in the source material and their interest in small-scale patterns, simulations could provide interesting settings for simulations. Microhistorical research might provide material well suited for counterfactual simulation, too. As has been pointed out already, simulations could be used to explore historical alternatives as perceived by contemporaries (Bunzl 2004, p. 848). The microscopic nature of studies in this field might make it easier to devise simple and easy to control models and to validate them.

An example of a relatively low-key but effective simulation intended for microhistorical research is Jeremy Throne's agent-based digital simulation of Robert Darnton's "communications circuit". Darnton's model was created to explain a unique system: the French book trade in Enlightenment Europe. It was about how books were produced, sold and read, how they spread ideas and influenced discourse. Throne's simulation does not intend to assess some general economic theory. But it allows testing of basic assumptions articulated in the underlying model for consistency. As Gavin points out, the validation process attains central importance. When agents do not behave in the way the designer expects them to, modifying the model becomes a process of intellectual inquiry in itself which helps in the process of reformulation and refinement (Throne 2014; Gavin 2014; Darnton 1982).

36.4.3 Digital Games and Simulation Games

Simulations also hold promise as pedagogical tools or as tools of demonstration (see also Hartmann 2011, p. 85). Most importantly, they can teach students the indeterminate nature of history (Winnerling 2017; Köstlbauer 2015; Vowinckel 2009) and provide very helpful aids to cognitive and affective understanding (Bigelow 1978, p. 209–10). The pedagogical use of simulations also brings us to the topic of digital games: Right from the beginning digital games with historical themes have been envisaged and indeed used as pedagogical tools by teachers at schools and universities (Bigelow 1978). The first historically themed digital game seems to have been an economic simulation game developed between 1962 and 1967 depicting ancient Sumer (Winnerling 2018). Especially games like those of the Civilization Series (Microprose, Activion, and Firaxis 1991–2016) or the grand strategy titles developed by Swedish studio Paradox Interactive are cited as examples of simulation games suitable for teaching.

History provides popular settings for games, which are a testament to the mimetic desire shaping a popular approach to history. More than that, many games are designed and marketed with the outspoken claim to provide a simulation of history, and this is not simply a marketing trick or a pedagogical streak in game design communities. There is a considerable demand; it is something players want and ask for loudly and persistently. They intensely scrutinize games, on forums, there are huge discussions about the accuracy of the way the past is represented in games, right down to the level of nitpicking. A striking example is the discussion about the depiction of Samurai in Total War: Shogun II (Creative Assembly 2011). These were shown with Horo cloaks on their backs, which seemed strange and inauthentic to many players despite valid historical existence for their use. Eventually, “no horo” modifications were produced (Pfister 2017).

It is, of course, debatable whether games (“video games”) should be included in a discussion of simulation and history. But it is hard to deny that there is a simulation aspect within games. Gonzalo Frasca, an influential proponent of Game Studies, even proclaimed simulation the central characteristic of digital games in opposition to narrative (Frasca 2003). On the other hand, interaction with simulations can be described as containing basic elements of play (see recently Saam and Schmidl 2018 or Köstlbauer 2015). There is also a historical connection between simulations, games, and the computer. John von Neumann built one of the first working computers, and together with Oskar Morgenstern, he published the first works on game theory (Lévy 1998, pp. 937–944; Hilgers 2008, pp. 175–179).

Digital games also confront us with the vagueness of the term “simulation” in popular usage. So-called simulation games encompass a heterogeneous range of games, from vehicle and battlefield simulations such as the Digital Combat Simulator (Eagle Dynamics 2008) or the Armed Assault Series (Bohemia Interactive 2009) to more fanciful games like SimCity or The Sims 4. Vehicle simulations and some military simulations are straddling the divide between serious games/training aid and leisure games. In the case of The Sims, one wonders whether the latter are simulations in anything but the loosest sense of the word or whether they are better understood as a warped representation of the dreams and fears harbored by a contemporary American middle class. They are more akin to "simulacra", simulations of ideas or visions (Baudrillard 1994). As such they become a fascinating source of historical inquiry regarding US culture and society.

What distinguishes a historical simulation game from a game with a bit of historical varnish on top, remains open to debate. So far any attempt at drawing generally accepted boundaries has proven more or less futile, the distinction remaining dependent on context (Köstlbauer 2013; Sauvé et al. 2007). But like bland academic simulations historically themed games need to relate to concepts of reality.

No matter whether off-the-shelf entertainment products or specially developed pedagogical instruments, simulation games can offer alternatives to deterministic and teleological representations of history still prevalent in Western culture. Designed around simple models such simulation games can allow the demonstration of specific mechanisms at work in historiography. They can also facilitate the reflections of the possible futures governing the actions of historic actors. And they also should be used to reflect on the ways the design of the simulation influences results. Therefore, validation becomes a central part of the learning process. Such goals in mind two historians at the University of Düsseldorf in Germany (Heinrich Heine Universität) developed a history video game named Lienzo. Built around a well-known chain of events, the sixteenth century conquest of Mexico, it intends to demonstrate both historical causalities as well as the fundamentally undetermined character of history (Winnerling 2017).

Given their status as mass media and the significant number of sales some games achieve (Bogost 2007), any historical game makes for a worthwhile subject of research. Furthermore, some games offer tantalizing prospects to be appropriated as simulations for historical research. In 2015, several historians founded the German *Arbeitskreis Geschichtswissenschaft und Digitales Spiel* (research group for History and digital games). During the German Historikertag 2017 (biannual conference of German historians), it published a manifesto that also includes a statement on simulations (gespielt.hypotheses.org 2017). Many of the points regarding simulations raised in this book chapter are also emphasized in the manifesto.

36.5 Conclusion and Outlook

History will retain its idiographic posture. But that need not keep historians from exploring more fully the potentials of computer simulations. Ultimately, the use of models and simulations and the concomitant validation processes have the potential to make researchers reflect more sharply on the hypotheses and evaluative assumptions influencing model design. It leads to reflections on the ways historians comprehend their discipline and its subject, how they pose questions and explain processes. Therefore, even more important than the results produced by runs of a simulation may be the questions springing from it. Is a solely or primarily idiographic approach still feasible? Or, how can an idiographic approach be enhanced or buttressed by computer simulations and the related scientific methodology?

Introducing computer simulations into historical research requires renewed analysis of what is particular and what is general. Such analysis may provide opportunities to transcend the idiographic-nomothetic divide, since formulating general laws and focusing on the unique become near-simultaneous and reciprocal.

The more abstract a simulation and the more limited the hypothesis, the simulation is going to test, the easier it is to understand the model and the dynamics at work and the easier and cheaper it is to create. Generally, there are persuasive arguments for simplicity of models (Turchin et al. 2014, in reply to Thomas 2014; also Troitzsch 1994, p. 62), but it seems doubtful whether they can be particularly useful to historians (Weber 2007, p. 109–110) because they are likely to produce very general statements. Creating complex agent-based simulations leads to greater opacity of the simulation itself, both regarding understanding the code and understanding what happens during simulation and why. This makes it so much harder to judge the validity of results. But it also may create a blind spot regarding technical design decisions and epistemological assumptions or theoretical bias influencing model design. These decisions may range from the programming language used to the ways space determines agent movements (Weber 2007, p. 113) or the way the factor time is incorporated.

The new interest in historical simulations is a recent phenomenon, hardly a decade old in 2018. It is part of a "visionary discourse and transformative sentiment" (Svensson 2012) associated with digital humanities. The assertion and often enthusiastic endorsement of the game-changing nature of digital technology for the humanities (as well as for other disciplines) has lead to many research initiatives, methodological innovation, and a new sense of interdisciplinary endeavor (see for Porsdam 2011; Burdick et al. 2012; Weller 2011). Whatever one might think of these, there is no doubt that there are very serious efforts underway. This is also proven by the increasing number of chairs, research positions, digital humanities labs being created at universities or research centers in the last ten years.

On the other hand, there may be concern that the digital humanities are just the latest fad sweeping the humanities, another turn of the revolving door of academic fashion, moving in the same merry roundabout as formerly the linguistic, the visual, the cultural or the spatial turn (for a discussion of this see: Herbst 2004, p. 14–17; also Porsdam 2011; Weller 2011).

As far as the discipline of history is concerned, it remains yet to be seen whether there will be a profound methodological and conceptual change anytime soon. There also remain questions about the ways academic institutions will deal with the expenses in terms of time and funds that will be required to utilize simulation. How will historians argue the necessity of such endeavors and how do they assess risk and cost when preparing applications for research funds? The progressing institutionalization of digital humanities may alleviate these problems, but the type of interdisciplinary liaison required is still found far too rarely.

References

Baudrillard, J. (1994). *Simulacra and simulation*. Ann Arbor: University of Michigan Press.

Bigelow, B. E. (1978). Simulation review: Simulations in history. *Simulation & Gaming, 9,* 209–220.

Bogost, I. (2007). *Persuasive games: The expressive power of videogames*. Cambridge, MS: MIT Press.

Braudel, F. (1990). *Das Mittelmeer und die mediterrane Welt in der Epoche Philipps II*. Frankfurt am Main: Suhrkamp.

Braudel, F., & Colin, A. (1987). Histoire et sciences sociales: La longue durée. *Réseaux, 27,* 7–37.

Brendel, H. (2010). Historischer Determinismus und historische Tiefe - oder Spielspaß? Die Globalechtzeitstrategiespiele von Paradox Interactive. In A. Schwarz (Ed.), *Wollten Sie auch immer schon einmal pestverseuchte Kühe auf Ihre Gegner werfen?* (pp. 95–122). Münster: Lit-Verlag.

Bunzl, M. (2004). Counterfactual history: A user's guide. *American Historical Review*, *109*, 845–858.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). *Digital_Humanities*. Cambridge, MS: MIT Press.

Castiglione, B. (2007). Il Libro del Cortigiano. Milano: Garzanti Libri 2007.

Corbeil, P. (2011). History and simulation/gaming: Living with two solitudes. *Simulation and Gaming, 42,* 418–422.

Crabb, G. (1824). *English synonyms explained*. London, UK: Baldwin & Cradock.

Creative Assembly. (2011). *TotalWar: Shogun II*. Creative Assembly. Windows, Mac.

Darnton, R. (1982). What is the history of books? *Daedalus, 111*(3), 65–83.

Demandt, A. (2011). Ungeschehene Geschichte. Ein Traktat über die Frage: Was wäre geschehen, wenn …? Göttingen: Vandenhoeck & Ruprecht.

Ekirch, A. R. (1994). Sometimes an art, never a science, always a craft. A conversation with Bernard Bailyn. *The William and Mary Quarterly*, *51*(4), 625–658.

François, P., et al. (2016). A Macroscope for Global History: Seshat Global History Databank, a methodological overview. *Digital Humanities Quarterly, 10*(4).

Frasca, G. (2003). Simulation versus narrative. Introduction to ludology. In M. J. P. Wolf & B. Perron (Eds.), *The video game theory reader* (pp. 221–235). New York: Routledge.

Gavin, M. (2014). Agent-based modeling and historical simulation. *Digital Humanities Quarterly, 8*(4).

Geiss, I. (1998). *Zukunft als Geschichte. Historisch-politische Analysen und Prognosen zum Untergang des Sowjetkommunismus, 1980–1991*. Stuttgart: Franz Steiner.

gespielt.hypotheses.org (2017). Manifest für geschichtswissenschaftliches Arbeiten mit Digitalen Spielen! Version 1.1. Retrieved from http://gespielt.hypotheses.org/manifest_v1-1.

Gracián, B. (1993). *Hand-Orakel und Kunst der Weltklugheit*. Zurich: Diogenes.

Guldi, J., & Armitage, D. (2014). The history manifesto. Cambridge, UK: Cambridge University Press. Retrieved from http://www.cambridge.org/core/what-we-publish/open-access/the-history-manifesto.

Hartmann, S. (2011). The world as a process: Simulations in the natural and social sciences. In R. Hegselmann, U. Mueller, & K. G. Troitzsch (Eds.), *Modelling and simulation in the social sciences from the philosophy of science point of view* (pp. 77–100). London: Springer.

Hazard, P. (2013). *The crisis of the European mind, 1680–1715*. New York: New York Review Books.

Herbst, L. (2004). *Komplexität und Chaos. Grundzüge einer Theorie der Geschichte*. Munich: C. H. Beck.

Hirschfeld, G., Krumeich, G., & Renz, I. (2009). *Enzyklopädie Erster Weltkrieg. Aktualisierte und erweiterte Studienausgabe*. Paderborn: Schöningh.

Koselleck, R. (1989). *Vergangene Zukunft*. Zur Semantik geschichtlicher Zeiten. Frankfurt am Main: Suhrkamp 1988.

Köstlbauer, J. (2013). The Strange attraction of simulation: Realism, authenticity, virtuality. In M. Kapell & A. B. R. Elliott (Eds.), *Playing with the past. Digital games and the simulation of history* (pp. 169–184). New York: Bloomsbury Academic.

Köstlbauer, J. (2015). Spiel und Geschichte im Zeichen der Digitalität. In Schmale, W. (Ed.): *Digital Humanities. Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität* (pp. 95–124). Stuttgart: Franz Steiner.

Lengwiler, M. (2011). *Praxisbuch Geschichte*. Einführung in die historischen Methoden. Zurich: Orell Füssli.

Lévy, P. (1998). Die Erfindung des Computers. In M. Serres (Ed.), *Elemente einer Geschichte der Wissenschaften* (pp. 937–943). Frankfurt am Main.

Manovich, L. (2017). Cultural analytics, social computing, and digital humanities. In M. T. Schäfer (Ed.), *The datafied society. Studying culture through data* (pp. 55–74). Amsterdam: Amsterdam University Press.

Manovich, L., Tifentale, A., Yazdani, M., & Chow, J. (2014). The exceptional and the everyday: 144 hours in Kiev. Retrieved July 28, 2018, from http://manovich.net/content/04-projects/084-kiev-article/sk219_1225.pdf.

McLuhan, M. (1962). *The gutenberg galaxy*. Toronto: University of Toronto Press.

Meeks, E. (2015). Networks in the study of culture and society. In H. Kaper & C. Rousseau (Eds.), *Mathematics of planet earth: Mathematicians reflect on how to discover, organize, and protect our planet* (pp. 157–159). Philadelphia, PA: SIAM.

Michl, S. (2007). *Im Dienste des "Volkskörpers": Deutsche und französische Ärzte im Ersten Weltkrieg*. Göttingen: Vandenhoeck & Ruprecht.

Murauer, J. (2014). Modellbildung und Simulation als Methode zur Bearbeitung soziologischer Fragestellungen aus dem Bereich der Entwicklungszusammenarbeit. Erörterungen anhand einer Fallstudie zu einem Erosionsschutzprojekt in Burkina Faso (Westafrika). Berlin: Lit-Verlag.

Nonn, C., & Winnerling, T. (2017). *Eine andere deutsche Geschichte 1517–2017*. Paderborn: Schöningh.

Pfister, E. (2017). „Wie es wirklich war." – Wider die Authentizitätsdebatte im digitalen Spiel. Arbeitskreis Geschichtswissenschaft und Digitale Spiele, 18 May 2018. Retrieved from https://gespielt.hypotheses.org/1334.

Porsdam, H. (2011). Too much 'digital', too little 'humanities'? An attempt to explain why many humanities scholars are reluctant converts to Digital Humanities. Retrieved July 28, 2018, from https://www.repository.cam.ac.uk/handle/1810/244642.

Redhead, M. (1980). Models in physics. *British Journal for the Philosophy of Science, 31,* 145–163.

Rodiek, C. (1997). Erfundene Vergangenheit. Kontrafaktische Geschichtsdarstellung (Uchronie) in der Literatur. Frankfurt am Main: Klostermann.

Saam, N., & Schmidl, A. (2018). A distinct element of play. Scientific computer simulation as playful investigating. In A. Friedrich, P. Gehring, Ch. Hubig, A. Kaminski, & A. Nordmann (Eds.), *Arbeit und Spiel. Jahrbuch Technikphilosophie 2018* (pp. 99–118). Baden-Baden: Nomos.

Sauvé, L., Renaud, L., Kaufman, D., & Marquis, J.-S. (2007). Distinguishing between games and simulations: A systematic review. *Educational Technology & Society*, *10*(3), 247–256. Retrieved from http://www.ifets.info/journals/10_3/17.pdf.

Schelling, T. C. (1978). *Micromotives and microbehavior*. New York: Norton.

Schich, M. (2016). Personal Website. Retrieved from http://www.schich.info/research.htm.

Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., et al. (2014). A network framework of cultural history. *Science, 345*(6196), 558–562.

Schmale, W. (2006). *Schreib-Guide Geschichte*. Wien: Böhlau.

von Hilgers, P. (2008). *Kriegsspiele. Eine Geschichte der Ausnahmezustände und Unberechenbarkeiten*. Paderborn: Fink, Schöningh.

Sehsat: Global History Databank. (2017). Retrieved from http://seshatdatabank.info/.

Sellin, V. (1995). *Einführung in die Geschichtswissenschaft*. Göttingen: Vandenhoeck & Ruprecht.

Sidnell, P. (2006). *Warhorse: Cavalry in ancient warfare*. London: Continuum.

Snyder, J. R. (2009). *Dissimulation and the culture of secrecy in early modern Europe*. Berkeley: University of California Press.

Svensson, P. (2012). Envisioning the digital humanities. *Digital Humanities Quarterly, 6*(1).

Thomas, R. C. (2014). Does diffusion of horse-related military technologies explain spatiotemporal patterns of social complexity 1500 BCE-AD 1500? *Proceedings of the National Academy of Sciences of the United States of America, 111*, E414. Retrieved from http://www.pnas.org/content/111/4/E414.

Throne, J. (2014). Modeling the communications circuit. In P. A. Youngman & M. Hadžikadić (Eds.), *Complexity and the human experience. Modeling complexity in the humanities and social sciences* (pp. 105–119). Singapore: Pan Stanford Publishing.

Troitzsch, K. G. (1994). The evolution of technologies. In N. Gilbert & J. Doran (Eds.), *Simulating societies. The computer simulation of social phenomena* (pp. 41–62). London, UK: Routledge.

Turchin, P., Currie, T. E., Turner, E. A. L., Gavrilets, S. (2013). War, space, and the evolution of Old World complex societies. *Proceedings of the National Academy of Sciences of the United States if America*, *110*, 16384–16389. Retrieved from http://www.pnas.org/content/110/41/16384.

Turchin, P., Currie, T., Turner, E. A. L., Gavrilets, S. (2014). Reply to Thomas. Diffusion of military technologies is a plausible explanation for the evolution of social complexity, 1500 BCE-AD 1500. *Proceedings of the National Academy of Sciences*, *111*, E415–E415. Retrieved from http://www.pnas.org/content/111/4/E415.

Turchin, P., Brennan, R., Currie, T., Feeney, K., François, P., Hoyer, D., et al. (2015). Sehsat: The Global History Databank. *Cliodynamics*, *6*, 77–107. Retrieved from http://www.researchgate.net/publication/279849138.

Turkel, W. J. (2008). Towards a computational history. *Digital History Hacks* (August 20, 2008). Retrieved from http://digitalhistoryhacks.blogspot.com/2008/07/towards-computational-history.html.

War Office Committee. (2004). Report of the War Office Committee of Enquiry Into "Shell Shock" (pp. 141–144). First published 1922, London.

Weber, K. (2007). Erklärung historischer Abläufe mit Computersimulationen. *Historical Social Research, 32*(4), 94–121.

Weller, M. (2011). *The digital scholar: How technology is transforming scholarly practice*. London, UK: Bloomsbury Academic.

Windelband, W. (1915). Geschichte und Naturwissenschaft (Starßburger Rektoratsrede). In W. Windelband, *Präludien. Aufsätze und Reden zur Philosophie und ihrer Geschichte* (Vol. 2, pp. 136–160). Tübingen: Mohr.

Winnerling, T. (2017). Selbstversuch: Wenn zwei Historiker ein Spiel machen… *Gespielt. Arbeitskreis für Geschichstwissenschaft und Digitale Spiele*. Retrieved from http://gespielt.hypotheses.org/1231.

Winnerling, T. (2018). Projekt Sumerian Game: Digitale Rekonstruktion eines Spiels als Simulation eines Modells. *Gespielt: Arbeitskreis Geschichtswissenschaft und Digitale Spiele.* Retrieved from https://gespielt.hypotheses.org/1796.

Zedler, J. H. (1746). *Grosses vollständiges Universal-Lexicon Aller Wissenschafften und Künste* (Vol. 47). Leipzig und Halle.

Part IX
Reflecting on Simulation Validation: Philosophical Perspectives and Discussion Points

Chapter 37
What is a Computer Simulation and What does this Mean for Simulation Validation?

Claus Beisbart

Abstract Many questions about the fundamentals of some area take the form "What is ...?" It does not come as a surprise then that, at the dawn of Western philosophy, Socrates asked the questions of what piety, courage, and justice are. Nor is it a wonder that the philosophical preoccupation with computer simulations centered, among other things, about the question of what computer simulations are. Very often, this question has been answered by stating that computer simulation is a species of a well-known method, e.g., experimentation. Other answers claim at least a close relationship between computer simulation and another method. In any case, correct answers to the question of what a computer simulation is should help us to better understand what validation of simulations is. The aim of this chapter is to discuss the most important proposals to understand computer simulation in terms of another method and to trace consequences for validation. Although it has sometimes been claimed that computer simulations are experiments, there are strong reasons to reject this view. A more appropriate proposal is to say that computer simulations often model experiments. This implies that the simulation scientists should to some extent imitate the validation of an experiment. But the validation of computer simulations turns out to be more comprehensive. Computer simulations have also been conceptualized as thought experiments or close cousins of the latter. This seems true, but not very telling since thought experiments are not a standard method and since it is controversial how they contribute to our acquisition of knowledge. I thus consider a specific view on thought experiments to make some progress on understanding simulations and their validation. There is finally a close connection between computer simulation and modeling, and it can be shown that the validation of a computer simulation is the validation of a specific model, which may either be thought to be mathematical or fictional.

Keywords Definition · Experiments · Thought experiments · Argumentation · Models · Internal vs. external validity

C. Beisbart (✉)
Institute of Philosophy, University of Bern, Bern, Switzerland
e-mail: Claus.Beisbart@philo.unibe.ch

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_37

37.1 Introduction

What is validation of computer simulations and how does it work? One strategy to make progress on these questions is to put another, apparently more fundamental, question first: What is a computer simulation, to begin with? The idea is that a closer understanding of what computer simulation is usefully constrains any sensible view on its validation.

The aim of this chapter is to pursue this strategy. I will thus address the question of what a computer simulation is and then consider consequences for understanding validation. The question of what computer simulation is has indeed been at the center of a lively philosophical debate (see Imbert 2017 and Saam 2017 for overviews). I will draw on this debate and consider important proposals about what a computer simulation is.

The question of what computer simulation is can most naturally be answered in terms of a definition (see Gupta 2015 for a primer on definition). In the recent philosophical literature about computer simulations, we do find attempts at such a definition (Hartmann 1996, Sect. 2 and Humphreys 2004, pp. 110–114). But the question of what computer simulation is has often been answered in a loser sense by subsuming it under, or relating it to, some known method such as experiment. In this chapter, I concentrate exclusively on proposals that spell out what computer simulations are by claiming a close association between computer simulation and some other method. The reason is that such accounts seem particularly promising for a better understanding of the validation of simulations because they open pathways into known territory. I do not require that the proposals under consideration aim at a full-fledged definition of simulation. Some proposals that have been much discussed in the literature do not attempt to give such a definition, and it seems inappropriate to exclude them. There is of course a downside, when I include accounts that do not provide a full definition and that do not even typify simulations in terms of a genus: The accounts are weaker and less informative. But more precision as to what a simulation really is would not likely be of much help for the understanding of validation. It has in general proven difficult to specify what sort of things other artifacts or creations of the human mind, e.g., novels or symphonies, are precisely. Fortunately, the type of thing to which artifacts belong seems rather immaterial for other questions we may ask about them.[1]

In this chapter, I will thus go through a number of *methods*, ask whether computer simulation may be understood by relating it to the method and then trace consequences for validation. Now a specific proposal to the effect that computer simulation is closely associated with such and such method will not deepen our understanding of the validation of simulations, if the proposal itself is implausible.

[1] If somebody claims that computer simulations are, say, experiments, then what is claimed may either be regarded as essential of computer simulation (such that it should be included in its definition), or it may be supposed to be a contingent claim about computer simulations. I take it that the views under consideration are meant to capture essential properties of simulations, but this is not necessary for my argument.

So I'll briefly evaluate each proposal under consideration. My conclusions in this respect may not be shared by every author in the field because the nature of computer simulations has remained controversial. I nevertheless hope to be fair and give most positions due consideration. For reasons of space, the discussion has to be brief, so I cannot fully cover the existent literature. Note further that the consequences that the accounts under consideration have for validation haven't yet been worked out; so in this respect, this chapter will move beyond the existent literature.

After some preliminaries, I'll start with the method of experiment (Sect. 37.3), move to thought experiment (Sect. 37.4) and modeling (Sect. 37.5), before I conclude in Sect. 37.6. Views that take computer simulations to be genuinely novel (see, e.g., Winsberg 2001; see Frigg and Reiss 2009 for discussion) are not covered in this way, but they do not promise an easy route to better understanding validation, so we can bracket them in what follows (in this volume, Chap. 43 by Imbert addresses the questions of whether, and how, computer simulation is novel).

37.2 Preliminaries

Before I look at various proposals that relate computer simulation to other methods, it is useful to comment on two important concepts, viz., those of computer simulation and of validation. As already implicit in the discussion so far, I assume that computer simulation qualifies as a scientific method. A scientific method, in turn, is a type of activity that scientists engage in to promote the ultimate aims of science, e.g., to gain knowledge and understanding. For the purposes of this chapter, I take it that a computer simulation crucially involves the run of a simulation program that provides possibly partial and approximate solutions to equations that trace the dynamical evolution of a target system (this is close to Humphreys' definition 2004, pp. 110–114). In this way, the dynamical evolution of the target is imitated or modeled. There are different ways to specify in more detail what sort of activities form part of a computer simulation (see, e.g., Parker 2009, p. 488), but how exactly this is done will not matter for our purposes. Simulations that do not involve a digital computer, e.g., analogue simulations, are neglected for the purposes of this chapter.

Regarding the notion of validation, we have to be very careful. In the sciences, not only computer simulations are said to be validated; rather, scientists talk about the validation of models and experiments too. This suggests that there is a general idea of validation that covers more specific notions like the validation of experiment, etc. In what follows, validation in this general sense will be called validation$^{\text{gen}}$. It comprises, very roughly, the activities that make a case that the results that have been, or can be, obtained by applying the method in a particular case do in fact hold true of one or more real-world systems. It is important to note that the activities of validation$^{\text{gen}}$ refer to reality.

In my discussion, I will sometimes focus on the validation of few actual results that have been obtained by applying the method in a specific case. But my considerations are meant to carry over to a more comprehensive validation. The latter covers not only the results that have in fact been obtained but also results that could be obtained by applying the method in a specific case, e.g., by running the same experiment in slightly different ways. I allow that validation$^{\text{gen}}$ can also be concerned with the question of whether the assumptions built into a specific application of the method hold.

As far as the validation of computer simulations, validation$^{\text{cs}}$, is concerned, I assume that it is validation$^{\text{gen}}$, as applied to computer simulations. In this case, the results are constructed from the output, which consists of values of characteristics such as position, energy, etc. The results can be cast in claims about the target system, which are either quantitative or qualitative. In the simplest case, such a claim has it that certain characteristics of the target system, say, the luminosity of a star, has such and such value, as output from the simulation. Since the outputs from simulations are affected by all kinds of errors, the output numbers reflect the target system at best only up to some accuracy. The question of whether the results of a simulation hold thus is meant to be the question of whether the results are sufficiently accurate for the intended applications. The point of validation then is to show that the results are sufficiently accurate. This understanding of simulation validation accords well with the famous definition of validation$^{\text{cs}}$ by Schlesinger et al. (1979, p. 104).

My focus in this chapter is on validation$^{\text{cs}}$ as applied to a simulation program or the model implicit in it, call it the computational model. At this point, it does not make a difference whether we talk about the program or the computational model because the program delivers exact solutions to the computational model (this is just how the latter is defined). The computational model needs to be distinguished, however, from the conceptual model which is typically the scientific model that scientists are interested in before they run the simulations. In many simulations, this model consists of differential equations, while the computational model approximates the latter in some way. So-called verification is supposed to show that the computational model is a faithful representation of the conceptual model for the purposes of the inquiry. If a simulation is properly verified, then the distinction between the conceptual and the computational model does not much matter. If it is not clear whether the simulation is verified, then we can in principle distinguish the validation of the conceptual model, validation$^{\text{con}}$, from that of the computational model (or the computer program), which has been called validation$^{\text{cs}}$.

Now if computer simulation turns out to be closely related to another method, then it is likely that we obtain consequences for validation$^{\text{cs}}$. In particular, if computer simulation is intimately connected to a method for which validation is an issue, it may turn out that validation$^{\text{cs}}$ boils down to the validation of the method to which computer simulation is assimilated. Since we will examine several distinct attempts to relate computer simulations to other methods, it is likely that we will consider different, possibly inconsistent claims that arise for validation$^{\text{cs}}$. It may be thought that this leads to different concepts of validation$^{\text{cs}}$, for instance, the concept of validation$^{\text{cs,e}}$ (validation of computer simulation under the assumption that the latter is an experiment), etc. In what follows, I will refrain from distinguishing such concepts because we are ultimately interested in one concept, viz., that of validation$^{\text{cs}}$, and since it is possible to discuss various views on validation^{cs} without assuming several concepts. Further, as I will briefly argue in Sect. 37.6 below, the proposals that prove to be sensible in our discussion are compatible with each other.

37.3 Computer Simulations and Experiments

Computer simulations are often called computer experiments (e.g., Beeler 1983), where the term "experiment" is sometimes put in scare quotes (e.g., in the title of Verlet 1967). Likewise, the term "experiment in silico" is used for simulations (e.g., Naumova et al. 2008). All this is no accident. There are in fact close parallels between experiments and computer simulations (see Beisbart 2018, Sect. 3). As is well known, experiments crucially involve two types of causal interaction between the working scientist and the system she is working on: intervention—an experimental system is set up or at least manipulated in some way—and observation—the reaction of the system is observed (see, e.g., Heidelberger 2005; Radder 2009 and Franklin and Perovic 2016 for reviews about experiment). Computer simulations seem to function in a parallel way: Simulation scientists interfere with the hardware of a computer to set up a system, which is then investigated. After the program has been run, they obtain outputs that are interpreted in the same way as are data from observation of an experiment. It is no surprise then that some philosophers have tried to understand computer simulation in terms of experiment.

37.3.1 Computer Simulations as Experiments

Some authors have gone as far as to claim that computer simulations are, or crucially involve, experiments. In more detail, there are two ways in which this claim may be spelled out (Beisbart 2018, Sect. 4): Either *the computer* itself is supposed to be the system experimented on. The results obtained for this system are then transferred to the target system of the simulation, e.g., to a cell or a galaxy that is simulated. Parker (2009) defends this view. She takes it to be obvious that a simulation involves the twofold causal interaction with a computer that is constitutive of experiment (ibid., p. 488). She further argues that the experimental status of simulations is important for a comprehensive epistemology of computer simulation because certain concerns that may arise about a simulation hinge on the fact that a material system (here, the computer) is under investigation (pp. 489–491). As an alternative view, it is suggested that at least some simulations are really experiments *on the target system* of the simulation (e.g., a galaxy). Morrison (2009) takes some steps in this direction, although she never claims simulations to be experiments. But she definitely takes some simulations and experiments to be on par epistemically, e.g., because models function in the same way in both methods. Massimi and Bhimji (2015) make a stronger case for the view that simulations are experiments by arguing that some computer simulations from particle physics involve causal interactions that are not relevantly different from the causal interactions between experimenters and the systems they experiment on (cf. also Morrison 2015, Part III).

Suppose now, for the sake of argument, that at least some computer simulations are, or crucially include, experiments. What would this mean for validation^{cs}?

It is first interesting to note that the term "validation" or "validity" is well-established for experiments too. In some sciences, e.g., the social sciences, it is common to distinguish between internal and external validity^{e}.[2] For our purposes, we may understand the distinction as follows: Internal validity^{e} is about results that concern the system with which the experimenter interacts causally (i.e., the system experimented on) during the time when the experiment is run. External validity^{e}, by contrast, is about the generalization to other times, condititions, systems, etc. For a simple example, an experimentalist may suspect that, in the system on which she has experimented, a particular medical treatment of a person has caused her recovery. The experiment is internally validated^{e} in this respect, if the experimenter shows that the treatment did cause the recovery in the specific case under consideration. The experiment is externally validated^{e} if the effect can be shown to generalize to other patients. Both ways, validity^{e} is a matter of inference. Note that external validity^{e} is only a concern if scientists wish to generalize their results. This condition is not met in all experiments. It is possible to run an experiment on, say, a population of animals just to learn about this very population at one time.

The distinction between internal and external validity^{e} is quite rough. In some areas, other, more fine-grained distinctions are drawn. For instance, in educational and psychological research, people discriminate between construct, content and criterion validity^{e} (see, e.g., Newton and Shaw 2014). Here, construct validity^{e} is roughly supposed to ensure that a measurement does indeed reflect a theoretical construct. In what follows, we cannot discuss such domain-specific notions of validity^{e}, but we will below comment on construct validity^{e} and relate it to validity^{cs}.

If simulations were, or included, experiments *on the computer hardware*, they would be internally valid^{e} qua experiment if results about the computer were shown to be genuine (this is in fact suggested by Parker 2008, p. 168). But when running a computer simulation, scientists do not typically establish any results about the computer hardware. The outputs of the computer simulation are immediately interpreted in terms of the target system, and not in terms of the computer hardware. For instance, if a simulation program prints a series of numbers, the latter are interpreted as temperatures of the target system for a series of times. In fact, most simulation scientists cannot even use the output to infer anything interesting about the hardware because they know virtually nothing about the hardware. So assuming that computer simulations are experiments, internal validity^{e} is not a matter of concern for computer simulations. But then, nor can external validity^{e} be. The point of external validity^{e} is the generalization of results that have been established about the system

[2] The distinction goes back to Campbell 1957 (see Winsberg 2009, p. 579 following Parker). See also Campbell and Stanley (1963, p. 5) and Cook and Campbell (1979, p. 37). It was originally restricted to experiments in social science that aim at causal claims.

experimented on. If no such results have been obtained, external validity^e cannot get started.

So the notions of internal and external validity^e do not make much sense for computer simulations, if the latter are considered to be experiments on the hardware. This casts doubts on the very idea that simulations are experiments on the hardware. It is in fact problematic to say that computers are observed qua experimental system if working simulation scientists only understand the output in terms of the target system. Likewise, it seems problematic to suggest that computers are manipulated in the way experimental systems are, if the working scientists don't really know what they are doing with the computer qua material system when they, e.g., type commands in the keyboard. For these reasons, computer simulations are not, and do not include, experiments on the hardware of the computer (see Beisbart 2018, Sect. 5 for details).

Turn now to the proposal that computer simulations are experiments *on the target*. This view fits better with the distinction between internal and external validation^e of experiments. Suppose for instance that a merger between two known galaxies is simulated. Qua experiment, the simulation would be internally valid^e, if it produced genuine results about the specific galaxies involved in the merger. It would be externally valid^e, if the results were shown to extend to other mergers of galaxies.

The problem though is that the proposed view, viz., that computer simulations are experiments on the target, itself isn't plausible. First and quite obviously (and pace Massimi and Bhimji 2015), computer simulations do not involve the characteristic twofold causal interaction between the experimenter and the experimental system. If a galaxy is simulated, this system is neither manipulated nor observed (see, e.g., Beisbart 2018, Sect. 6). Second, there is a crucial epistemological difference between computer simulations and experiments in that the assumptions built into a simulation in some sense imply the results, whereas this is not so for experiments (Arnold 2013, pp. 59–60, Beisbart 2018, Sect. 6). A consequence is that, in the words of Morgan (2005, p. 324), whereas simulations may surprise, only experiments can confound and thus lead scientists to question their assumptions.

37.3.2 Computer Simulations as Modeled Experiments

Now if computer simulations are neither experiments on the hardware nor on the target of a simulation, how can we explain the striking similarities between both methods? One proposal is that computer simulations can *model* possible experiments and do in fact often do so (cf. the title of Winsberg 2003; see Beisbart 2018 for an elaboration). The idea is that simulations first allow the scientist to model interventions on the target by setting the initial conditions and the parameter values that serve as input to a simulation program. The reaction of the target then is traced by running the program. Finally, observation and analysis of the data are modeled by those activities with which simulation scientists process the output of the simulations. In this way, computer simulations allow the representation of possible experiments.

I'm here talking of *possible* experiments because the experiments that are represented may in fact never be carried out on the real target system. Note further that some actual computer simulations (qua runs of a simulation program) do not model an experiment because they just try to represent the dynamics of a real-world target system as it happens to be like without any intervention. It is finally important to note that the various steps of an experiment are modeled in different ways: While the reaction of the target system is just modeled using the model implicit in the computer simulation, the intervention on the target system and the observation are not (the program does not contain variables that trace the working scientist looking at the target system, for instance). Rather, intervention and observation are modeled by activities on the part of the computer scientist, when she sets the initial conditions and observes the result on the screen.

If this proposal is on the right track (and it may be debated whether it is), then simulation scientists should appropriately model activities that establish internal and, if applicable, external validation^{e}. This consequence is not implausible. To show this, we consider a simple schematic example. We concentrate on internal validity^{e}, because it is generic. So suppose that nano-scientists want to know how the flow of a fluid through a nano-channel is influenced by the roughness of the wall of the channel. They run a molecular dynamics simulation of the system (see Liu et al. 2010 for an example of such a simulation). Suppose that their simulation outputs indicate that "roughness reduces the electro-osmotic flow rate dramatically even though the roughness is very small compared to the channel width" in a specific case (ibid., p. 7834). In an analogous way, measurements from an experiment may in principle indicate such an effect. In what follows we assume that this (qualitative) claim is the result that the scientists are interested in. Now to internally validate^{e} this result, experimentalists need to show that the effect is genuine. For instance, they have to make sure that there is in fact a dramatic reduction of the flow rate. Also, since the reported result has some generality because there are many ways in which the surface of a wall may be rough, the experiment has to be run for different realizations of roughness. But if this is needed in the experiment, then simulation scientists should model this internal validation^{e}, when they run the simulation. That is, they need to make sure that the simulation output does in fact imply a dramatic reduction of the flow rate and that the effect holds for many realizations of roughness. They will thus run the program with different parameters for the roughness. These activities form certainly some part of validation^{cs} because they are needed to show that there is this effect in the target. So we can say that, to some extent, validation^{cs} models internal validation^{e}.

Note though that certain concerns that matter for the validation^{e} of experiments are often not a real issue in the validation^{cs} computer simulations: For instance, in many experiments, some characteristics are measured using extremely complicated measurement devices. Internal validation^{e} has to make sure that the measurement devices function as intended. And it's a matter of construct validity^{e} that the measurements do in fact reflect the construct scientists are interested in. In many simulations, all this is not an issue because the characteristics are traced by the computer simulation program such that their values can be output and directly inspected by the scientists (there are some simulations that cover the measurement devices too, for instance in particle physics; see Massimi and Bhimji 2015 for a philosophical account). For another issue that need not concern simulation scientists, experimentalists cannot perfectly shield their experiments from external influences. If they observe a specific effect, they have to exclude that it was produced by external factors not controlled for in the experiment. This is not a concern for most simulations because they typically isolate the system under consideration in a perfect manner simply by not modeling external factors.[3]

Conversely, there is also a task in validation^{cs} that does not have a counterpart in validation^{e}: Simulation scientists need to show that their simulation faithfully traces the behavior of the real-world target system. If it doesn't, then what is claimed as result doesn't hold. The focus here is on the reaction of the target system. The computer program may after all misrepresent the way in which the target system behaves under the conditions that have been set. The proposal that computer simulations can model experiments can account for such practices. The reason is that, under the proposal, the experiment is only modeled and the model needs of course validation^{m} too (see below for validation^{m} of models).

Thus, assuming the proposal, we can distinguish between two layers of validation^{cs}: First, it must be shown that the modeled experiment really has such and such as result. This is to model internal validation^{e} from the experiment. As indicated, this is typically much easier than for real experiments. Second, it must be shown that the model of the experiment delivers a faithful representation of the way in which the target system reacts to the setup produced initially. This covers most part of the validation^{cs}, as it is known for simulations.

Now when simulation scientists validate^{e} their simulation, this closely resembles the validation^{e} of experiments. As Parker (2008) argues, at least five validation^{e} strategies known from the validation^{e} of experiments have close parallels in the validation^{cs} of computer simulation. For instance, both experimentalists and simulationalists can argue that their apparatus/simulation is built upon well-confirmed theory. Most importantly, both experimentalists and simulationalists can choose the so-called Sherlock Holmes strategy, which is to exclude all sorts of errors. Parker is certainly right in observing such parallels. Maybe, they can to some extent be accounted for by saying that the simulations model experiments. But the parallels should not lead us to assimilate validation$^{e/cs}$ of experiments and of computer simulations too much. As noted above, only the latter has to make a case that the reaction of the target system is faithfully modeled. This point is also clear from Parker's discussion when she notes that a simulation may be validated^{cs} by validating^{m} the underlying model and then showing that the computer program does in fact yield approximate solutions to the model (ibid., pp. 166–7). Showing that a program delivers such suitable approximations doesn't have a parallel in experiments. When even this part of validation^{cs} works in a similar way as does the validation^{e} of experiments, the reason is not a deeper parallel between computer simulation and experiment, but

[3] In a computer simulation, the computer hardware may of course be subject to influences not controlled for. But this is typically excluded by activities of verification, see below.

rather that prescriptions such as “check that there are no errors” apply quite generically to situations in which many errors are possible (ibid., pp. 178–179).

If validation does in fact differ between experiment and computer simulation, we may use validation to discriminate between both methods. This strategy has been adopted by Winsberg (2009). He assumes that experiment and simulation involve an inference from the system that is directly studied to a target system that is typically different from the system studied first. His proposal is (ibid., p. 586):

> what distinguishes simulations from experiments is the *character of the argument given* for the legitimacy of the inference from object to target and the *character of the background knowledge* that grounds that argument.

Very roughly, the crucial idea is that the arguments used in the validation$^{\mathrm{cs}}$ of computer simulation draw on trust that the working scientist has the right sort of principles for modeling the target system under consideration (ibid., p. 587). This is compatible with the view that computer simulations can model experiments.

To sum up then this section: Experimentation and computer simulation resemble each other in many respects. This cannot be explained by saying that the latter are, or include, experiments since simulations do not obey the conditions constitutive of experiments. Rather, many simulations can be said to model possible experiments. The consequence for validation$^{\mathrm{cs}}$ is that, to some extent, the validation$^{\mathrm{cs}}$ of computer simulations may be understood as modeling the validation$^{\mathrm{e}}$ of a possible experiment. But this does not exhaust the validation$^{\mathrm{cs}}$ needed for computer simulations. Rather, the validation$^{\mathrm{cs}}$ of computer simulation needs also to show that the possible experiment is after all well traced. As a matter of coincidence, even this part of validation$^{\mathrm{cs}}$ follows general strategies that are used in experimentation such as the exclusion of errors.

37.4 Computer Simulations, Thought Experiments and Argumentation

If computer simulations are not really experiments, they may still qualify as *thought experiments*. Very roughly, when a scientist runs a thought experiment, she considers a certain scenario and tries to anticipate in thought what will happen in this scenario. For instance, Einstein used a thought experiment involving a train running through a station to show that different observers do not agree on whether two events are simultaneous or not (the thought experiment is described in Einstein 1920, pp. 11–27; see Brown and Fehige 2017 for more examples and a review of the philosophy of thought experiments).[4]

Thought experiments do not involve any causal interaction of the scientist with the system investigated. They are thus not a subclass of experiments. Accordingly,

[4] Our focus in this section is exclusively on *scientific* thought experiments. Philosophers too engage in thought experimentation, but it is at least arguable that thought experimentation in philosophy and the sciences function quite differently.

crucial objections against the view that computer simulations are experiments do not apply anymore. It is indeed plausible to say that, in a computer simulation, a certain scenario is thought through as it is in a thought experiment. The role of the computer here is to expand the human capacities to think (see Humphreys 2004, Chap. 1 for this idea).

It does not come as a surprise then that the philosophical literature has closely associated computer simulations and thought experiments. Humphreys (2004, p. 115) has noted that computer simulations have taken the role that thought experiments had in less technologically advanced times. El Skaf and Imbert (2013) argue that thought experiments and simulations alike fall under the same general description of unfolding a scenario.[5] Beisbart (2012) argues that computer simulation and thought experiments fall under the broader category of scientific argumentation, although they differ in a couple of respects. Lenhard (2011), by contrast, draws a starker contrast between the methods and argues that thought experiments have a transparency that computer simulations lack.

For the purposes of this chapter, we need not take a stance on whether computer simulations qualify as thought arguments or whether they are species of the same genus. For even if a close connection between both methods can be established, this connection does not much help understanding computer simulation and its validation^{cs}, unless more is said about thought experiments, and this is in fact difficult. If thought experimentation is a method of its own at all, it is quite peculiar. It is not a method that is applied as widely or as standardly as is experimentation. The examples of scientific thought experiments identified in the philosophical literature are few. And there is no established methodology of running and validating thought experiments.

From a philosophical perspective, it is controversial how thought experiments achieve their tasks. The last two decades have seen a lively philosophical debate on this topic with a wide spectrum of positions. To mention the most extreme ones, whereas Norton (1996, 2004a, b) claims that thought experiments are arguments, Brown (1991, 2004) thinks that some type of thought experiment provides a priori epistemic access to laws of nature. Other positions hold that at least some thought experiments rely on quasi-observational intuitions that can provide justification for belief (Gendler 2004) or that thought experimenting is based upon mental modeling (Nersessian 1992, 2007; see Sect. 5 below for more on models).

So thought experimentation is neither a particularly well-established method nor well understood, as far as its philosophical account is concerned. It will thus not much enhance our understanding of computer simulation if we establish a close connection between the latter and thought experimentation. We will thus turn to a particular philosophical account of thought experiments that promises at least some insight into the validation^{cs} of computer simulations, viz., Norton's so-called argument view. As already indicated, the view has it that thought experiments form a species of arguments and thus instantiate scientific inference. Norton makes a case for

[5] They also include experiments under this description, but in the last section, we have already noted crucial differences between experiment and simulation.

this view by, e.g., reconstructing known thought experiments in terms of arguments (e.g., Norton 1996, Sect. III).[6]

For our purposes, it is interesting to note that, in defending his view, Norton addresses potential problems with thought experiments. For one thing, he draws the attention to pairs of thought experiments that yield incompatible results. So at least one of the thought experiments must be deficient. If thought experiments are arguments, the incompatibility of their results can be explained by saying that at least one of the underlying arguments is not sound (Norton 2004b, Sect. 3). For another thing, Norton claims that the past record of thought experimentation is not impressive because many thought experiments have arrived at a wrong conclusion. He thus demands a mark of reliability for thought experiments. His own proposed mark is that the form of a thought experiment is taken to be legitimate by some logic (Norton 2004b, Sect. 4).[7] Finally, he suggests that a reconstruction of a thought experiment in terms of an explicit argument may clarify its merits, e.g., by uncovering hidden premises (Norton 1996, Sect. 3.1).

So the argument view has some resources to address the reliability of thought experiments and, maybe, to develop a related methodology. Argumentation is in fact some part of the scientific method and has been extensively studied in logic and philosophy of science. The findings and techniques from logic in particular are of great help in the assessment of arguments. If the argument view of thought experiments has it right, logic may turn out helpful in the assessment of thought experiments too.

Let us thus try to extend the argument view to computer simulations (see Beisbart 2012 for details). The rough idea is that each run of a computer simulation program goes through some argument. The premises of the argument are the assumptions that underlie the simulation, e.g., about the dynamics of the target system or about the initial conditions. The conclusions are the results, which can be obtained from the output of the simulation. They have it that certain characteristics (temperature, positions of particles, etc.) take these and these values. Clearly, if all goes well, the conclusions follow from the model assumptions. So each simulation can at least be reconstructed as an argument. If we adopt the extended mind hypothesis (Clark and Chalmers 1998), we can even show that a coupled system consisting of the working scientist and the computer runs through the argument as a matter of fact.

Set up in this way, the argument view about computer simulations has some plausibility. When we were discussing above whether computer simulations are experiments, we have argued that they are not because the model assumptions implicit in the simulation imply in some way the result. This is to say that there is an argument running from the model assumptions to the result of a simulation. The argument view is focused on this very argument.

[6] There is no need here to draw on Nersessian's view that thought experimenting involves mental modeling since we'll examine simulations and models in due course in Sect. 37.4.

[7] This mark is not sufficient for a good thought experiment because even a valid argument can arrive at a wrong conclusion if some premise is false. But this complication does not matter for our argument.

Suppose now that this view is on the right track. What would the implications for the validation^{cs} of computer simulations be?[8] Well, if a computer simulation is an argument with the result as a conclusion, the result is likely true if i. the premises are likely true and ii. if they strongly support the conclusion. This suggests a certain two-step strategy to validation^{cs}.

How exactly this strategy to the validation^{cs} of computer simulation looks like turns on what exactly we take the argument and its premises to be. From the viewpoint of working scientists, it is natural to say that the argument takes the assumptions from the conceptual model as premises. After all, it is the conceptual model that is at the center of what scientists think about the target system. Thus, if we assume that the argument starts from the assumptions of the conceptual model as premises, then, in the first step in the strategy to validate^{cs} the simulation, scientists need to check that the assumptions of the conceptual model are likely true. For instance, the premises may be considered likely true either because they draw on well-confirmed theory or because they report measurements (e.g., about the initial conditions or of some parameter values). Now the assumptions of a conceptual model are often known to be false, because they are based upon approximations and idealizations. But even then, the assumptions may still be sufficiently accurate for the purposes of the inquiry. So if we weaken the premises and let them claim that the model assumptions are to some extent accurate, then scientists may be able to make a case for their likely truth.

In a second step, scientists have to check whether the premises strongly support the conclusion, i.e., what they obtain as result. Now what the computer does in order to produce the result is to go through a number of calculations. Exact calculations can be cast as deductive arguments (in fact, the point of a calculus is to obtain deductive arguments). This would mean that the arguments are as strong as they can be because the truth of the premises guarantees the truth of the conclusions. But in a computer simulation, the computer does not carry out exact calculations about the conceptual model. As is well known, the calculations done by the computer involve all kinds of errors with respect to the conceptual model, e.g., errors that arise from the discretization of differential equations or roundoff errors (see Chap. 5 by Roy in this volume). There may even be hardware failures or programming errors that prevent the computer program from working as intended. Now at least the known errors are usually taken into account when the results are formulated. As indicated in the preliminaries, scientists do not assume that the value of a characteristics in their target system is precisely, say, 4325 in some units, if this is the number output by the computer program. They rather assume that the output number reflects the true value up to some accuracy. Taken in this way, there is a chance that the premises (i.e., the model assumptions) do in fact imply the conclusion (i.e., the result), as is expected for a deductive argument.

But clearly, work is needed to show that the conclusion does in fact follow from the premises in a specific case. For instance, unknown errors due to hardware failure have to be excluded. Further, the accuracy of the results needs to be determined in

[8] Baumberger et al. (2017) have recently proposed to frame validation^{cs} using notions from argumentation theory. But this conceptualization of validation^{cs} is independent of the argument view.

such a way that they follow from the fact that the model assumptions hold to some accuracy. And if there are uncertainties in some model assumptions, scientists have to check what uncertainties they produce for the results. All these activities are well known as verification of a computer simulation program. To achieve this, scientists cannot write down an argument and check it using the standards of some logic (as might somehow be suggested by the argument view). This is too complicated because too many calculations are involved in a simulation. The reason is that computer simulations are opaque: We cannot see how the results follow from the premises in the way in which we can do this in simple thought experiments (see Humphreys 2004, Sect. 5.3 and Humphreys 2009, pp. 618–9 for opaqueness).

So far then, the argument view suggests a two-step procedure for validation^{cs}. If a computer simulation is an argument with premises from the conceptual model, then, in a first step, the accuracy of the premises is to be secured. A second step is supposed to make a case that the argument is sufficiently strong such that the conclusions follow from the premises. In principle, this is certainly a sensible approach to validation^{cs}. If validation^{cs} of a computer simulation is supposed to make a case that the results hold (up to some accuary), then a viable route is to show that the conceptual model that underlies the simulation is sufficiently accurate and that the computer program delivers results that are sufficiently accurate with respect to the conceptual model (see also Chap. 42 by Beisbart in this volume).

But there is a problem with this two-step strategy. The challenge is to make a sufficiently strong case for the premises, i.e., for the assumptions of the conceptual model. Even if the dynamcis of the target system is well understood in terms of a well-confirmed theory, this does not guarantee sufficient accuracy of the conceptual model because the theory needs to be combined with additional assumptions, e.g., about parameter values, initial and boundary conditions to produce a concrete model, and many of these assumptions will at best be uncertain. The only route to make a case for the conceptual model as a whole then is to run the simulation and to see whether the results match with measurements from the target system to a sufficient degree of accuracy. This strategy is of course familiar from validation^{cs}, as it is known and described in the literature. A comparison between data from the target (or, maybe, a system sufficiently similar to the target) and simulation output is what many people take to be crucial about validation^{cs}. This data-oriented approach to validation is not incompatible with the argument view. In terms of this view, what is crucial in the comparison between simulation output and measured data is, very roughly, this: Some premises of the argument behind the simulation are uncertain (e.g., because there is uncertainty about the values of certain parameters), so some of their consequences are derived using a simulation. If these consequences turn out to be true, then the premises are to some extent confirmed, and this confirmation extends to other consequences of the premises in new applications of the program. This way of reasoning is often called hypothetico-deductive approach. It makes use of deductive arguments, but in a way that is not as straightforward as to reason from the premises to the conclusion. A closer analysis of the inferences involved is beyond the scope of this chapter (but see Chap. 42 by Beisbart in this volume).

All in all, the argument view about computer simulations accommodates the activities of validation$^{\mathrm{cs}}$ as follows: It conceptualizes validation$^{\mathrm{cs}}$ in terms of an argument that is used as a reconstruction of the simulation. The basic point of validation$^{\mathrm{cs}}$ then is to argue that this very argument is sound. Because validation$^{\mathrm{cs}}$ is thus an argument about an argument, it can be called a meta-inference. What is a natural suggestion from the viewpoint of the argument view is a separation between the examination of the premises and of the way they support the conclusion (i.e., the results). If the premises are supposed to be assumptions from the conceptual model, then the argument view invites a 2-step procedure: A case is made that the conceptual model is sufficiently accurate (we can call this validation$^{\mathrm{con}}$ of the conceptual model) and a case is made that the results do follow from the premises (this is verification of the simulation program). However, this two-step procedure is often not viable because the conceptual model cannot be validated$^{\mathrm{con}}$ independently from the simulations, and to this extent, the suggestion on behalf of the argument view is not useful. Note though that validation$^{\mathrm{cs}}$ activities that compare simulation results with measured data can be accommodated within the argument view too. Nevertheless, the argument view remains a bit artificial in that the argument that has been proposed as reconstruction does not lend itself to an investigation because it is unclear how the conclusion follows from the premise. The argument is also quite far from the calculations done in the computer, which use all kinds of approximation schemes.[9]

37.5 Models and Simulations

The proposal that there is a close connection between computer simulation and modeling, indeed that computer simulations are some sort of models is now more than just in the air. In Sect. 37.3 above, we have proposed to say that computer simulations can model possible experiments. In Sect. 37.4 we have observed a continuity between simulation and thought experiment, where some authors take the latter to crucially involve mental modeling.

Talk about simulations too indicates a strong link between simulations and models, e.g., when people speak of simulation models. Hartmann's 1996 definition of computer simulation, according to which a simulation forms a process that imitates another process, also establishes a connection to models, insofar as imitation is a sort of modeling or representation. What is further promising for our purposes is that validation is an issue for models too (see, e.g., Koblick 1959, p. 642 for an example).

[9] We might also have provided a slightly different argument to represent a run of a computer simulation program: The idea would be that the premises state the computational model. Now the latter is defined such that results of the simulations are exact solutions to the computational model. So it is not an issue anymore to check that the argument is deductive. But the work of validation is only shifted to the examination of the premises. For instance, if we want to make a case that the computational model is sufficiently accurate by drawing on prior commitments to theory, we must show that the theory is likely sufficiently accurate and that it is appropriately reflected in the computational model.

Let us thus try to conceptualize simulations in terms of models and probe possible conclusions for the understanding of validation^{cs}.

A challenge that we face when thinking about models is an embarrassment of riches. There are not just many models, they also belong to different categories (see Frigg and Hartmann 2017 for a philosophical overview). Some models are material systems, e.g., scale models of cities or cars. Other models are merely imagined, e.g., a number of point particles connected by massless springs. Sometimes, a set of equations is said to form a model too. So we have at least material, fictional and mathematical models. There are likely more types of models, but the three categories will suffice for our purposes. Note too that models of different types are often closely related to each other; for example, mathematical equations can describe a fictional system.

In view of the plurality of (types of) models, it will hardly be illuminating to call computer simulations models, unless more is said about models. What then does modeling amount to and is there anything common to all types of models? A first observation that is relevant in this respect is that models are typically based upon simplifications. Various features of the target are abstracted away, idealizations are assumed and approximations made. For instance, a scale model of a city leaves out small-scale decorations of houses, it gives all buildings the same color and approximates the marketplace as a square. This gives rise to the following proposal: A model is a system that is distinct from the target system but used as a surrogate for the latter. Since the model is simpler than the target, scientists can more easily learn about the model; nevertheless, some of the findings obtained for the model can be transferred to the target system. I take this to be the core of insightful philosophical accounts of modeling, e.g., by Hughes (1997), Suárez (2004) and Weisberg (2007). This core suggests that modeling is an indirect research method that takes a detour via a surrogate (Weisberg 2007, p. 207). It further implies that modeling may be split into three stages, viz., construction of the model, analysis of the model and coordination between model and target (ibid., pp. 222–226). This view nicely accounts for material models and fictional models. It is less clear, however, how it applies to mathematical models. We may either stretch words a bit and say that sets of mathematical equations too constitute systems that are studied as surrogates for their targets (cf. Suárez 2004). Alternatively, we may say that mathematical equations are not really models, but model descriptions (see Weisberg 2007, p. 217). We can then say that they specify the dynamics of other, e.g., fictional models. This makes a lot of sense as long as the equations involve significant simplifications. If, by contrast, some equations are not built upon simplifications and iterally hold true of the target system, then we can take them to be descriptions of the target, and we need not call them a model.

Validation^{m} is an issue for models because modelers first obtain results for their models which they then need to translate to their targets. This is not to deny that the issue may arise whether results obtained for a model are genuine. Echoing the distinction known for experiments, we can say that this is a matter of internal model validity^{m}. But issues of internal validity^{m} are not specific to models. For instance, if an experiment on a material model is run, internal validity^{m} of the results about the model is internal validity^e of the experimental results. The crucial and characteristic question of validation^m is rather external: What is the justification to assume that some results obtained for the model apply to the target too? Note that external validity^e of experiments is about a similar sort of inference.

Let us now go back to our main topic and to simulations. What precisely is their relationship to models? This is an intricate issue because computer simulations involve various models of several types (see Beisbart 2014 for a more extensive discussion).

First, each computer simulation crucially involves a mathematical model. This holds true not only of computer simulations that attempt to solve ordinary or partial differential equations. It is also true of, e.g., agent-based models. Such models need not involve variables that take numbers as values, but they nevertheless involve equations, which trace the time evolution of purely qualitative variables (e.g., the preferred political party). More generally, every simulation contains rules that are supposed to trace the dynamics of the target system in some respect, and these rules can be cast as mathematical equations.

When a set of mathematical equations from a computer simulation involve a lot of simplifications with respect to the target system, it is natural to say that they directly refer not to the target system, but rather to a system distinct from the latter that is then used to understand the target. Since this system typically only exists in thought, we are talking about a fictional model. In fact, when computer scientists describe their simulations, they often refer to point particles that collide fully elastically and so on. Such point particles clearly form a fictional model. Accordingly, at least some computer simulations involve a fictional model. The latter is of course intimately connected with the mathematical model in a simulation, because the latter describes the former, in particular, its dynamics.

Does a computer simulation also involve a material model? Hartmann's definition of simulation in terms of a process that imitates another one refers to processes in a computer hardware. One can in fact show that, in successful deterministic simulations, the dynamics of the programmed computer represents the dynamics of the target system: The computer runs through a sequence of states that each correspond to states in the target system that follow the same order (Beisbart 2014). But it stretches things a bit to say that the computer itself serves as a surrogate for the target system. The reason should be clear from our discussion of experiments: We cannot really say that the computer hardware itself is observed or investigated by simulation scientists because no information about the computer is obtained. So in what follows, our focus will be on the mathematical and the fictional model involved in a simulation.

Above, we have distinguished between the conceptual and the computational model. This distinction is orthogonal to the distinction between mathematical and fictional models behind simulations. It is clear that there are both a conceptual and a computational mathematical model depending on whether we talk about equations that the scientists are really interested in on the basis of their knowledge or whether these are approximations that have to be made to implement the former equations in the computer. Likewise, we can apply the distinction between conceptual and computational models at the level of fictional models, if the assumptions about the imagined system that are inherent in the simulation program are not exactly the assumptions that scientists have started with. In particular, very often, the dynamics of the fictional system traced by the simulation program is not exactly the dynamics of the fictional system that scientists were originally interested in. This is of course due to approximations needed for the implementation of the conceptual model in a digital computer.

Given that there are various models associated with simulations, the question arises of what the point of running the computer program is. The answer is that a run of the computer program contributes to the analysis of the model (i.e., the second stage of modeling). As far as the mathematical equations are considered to be a model, the run of the simulation program yields information about its solutions. As far as a fictional model is concerned, information about its dynamical behavior is derived by working out an (approximate) solution to the equations that describe the fictional system. In either case, if all goes well, the simulation scientist first and foremost gains knowledge about the model.

The analysis of a conceptual model (be it mathematical or fictional) can be very difficult and lead to errors. The reason is that the conceptual model is some distance away from what the computer does in fact do. The verification of a simulation thus is supposed to make a case that the analysis produces genuine results about the conceptual model. Thus, what is called verification of a simulation is the internal validation$^{\mathrm{m}}$ of results about the conceptual model.

If such results have been established, they have to be translated to the target to make some progress in knowledge about, or understanding of, the latter. In terms of our modeling terminology, we may say that we need external validation$^{\mathrm{con/com}}$ of the conceptual or computational model to make sure that what the model suggests for the target holds true of the latter with sufficient accuracy. But the values that the computational model suggests for certain characteristics of the target are just the numbers output from the simulation runs. Thus, the results from the computational model (be it fictional or mathematical) are the results from the simulation. So, to validate$^{\mathrm{cs}}$ the simulation is to validate$^{\mathrm{com}}$ the computational model, and vice versa. This is in fact what we have proposed in the introduction, where validation$^{\mathrm{cs}}$ was defined to be validation$^{\mathrm{com}}$ of the computational model. We now see the justification for this. If the results from the simulation have been verified regarding the conceptual model, then the validation$^{\mathrm{cs}}$ of the simulation will also establish the external validity$^{\mathrm{con}}$ of the conceptual model; then, with some right, this validity$^{\mathrm{con}}$ may be called the validity$^{\mathrm{cs}}$ of the simulation too.

It thus turns out that both verification and validation$^{\mathrm{cs}}$ of a simulation can be understood in the terms familiar from modeling. In particular, the validation$^{\mathrm{cs}}$ of the computer simulation turns out to be the validation$^{\mathrm{com}}$ of the computational model, when we adopt the modeling terminology. As a consequence, principles from the methodology of modeling can be used to validate computer simulations. The problem is only that a neat and tidy methodology for validating models is as much missing as one for computer simulations. When we talk about fictional models, validation$^{\mathrm{m}}$ would have to show that some results on the fictional model carry over to the target because both are similar in relevant respects. Concerning mathematical equations, the point of validation$^{\mathrm{m}}$ is to show that the equations provide results that are accurate enough for the purposes of a simulation. Either way, there doesn't seem any general principled approach to achieve this. So we cannot simply draw on rich insights into modeling to make progress on understanding the validation$^{\mathrm{cs}}$ of simulations.

37.6 Conclusions

One strategy to understand validation$^{\mathrm{cs}}$ of computer simulation is to begin with the question of which sort of method computer simulations are, or, maybe, how they are associated with other methods, and then to derive consequences for validation$^{\mathrm{cs}}$. In this chapter, we have pursued this strategy. What did we earn by doing so?

Computer simulations can be associated with several methods that have some independent life and that have in fact been practised before the advent of computer simulation. As it happens, the term "validation" has currency regarding some of these methods too.

Although computer simulations are not running *experiments*, properly speaking, some simulations model possible experiments. An immediate consequence is that, to some extent, the validation$^{\mathrm{cs}}$ of experimental results need to be modeled too. But there is more to the validation$^{\mathrm{cs}}$ of computer simulation: It must be shown that they properly trace the target system, in particular, its reaction to an intervention. It turns out that techniques for doing so have close parallels in the methodology of experiments.

Carrying out a computer simulation is much closer to going through a *thought experiment* than running a real experiment. Both computer simulation and thought experiments do not involve causal interaction with the target system. But thought experimentation is a peculiar method; there is no worked out methodology, and the philosophical explanation of how thought experiments work (if they do) is controversial. A useful approach to at least many thought experiments is the so-called argument view. It can be extended to computer simulations and then basically cashes out the idea that computer simulations infer what a model implies for the dynamics of its target. From this perspective, the main question of validation$^{\mathrm{cs}}$ is whether the inference constituted by a computer simulation is sound. It is natural to split this question into two questions, viz., whether the premises are sufficiently accurate (at least as far as their impact on the results is concerned) and whether the argument is such that the premises support the conclusion sufficiently. Whether the premises are sufficiently accurate is a matter of validation$^{\mathrm{m}}$ of the underlying model. Whether or not the conclusion is sufficiently supported by the premises is in principle a matter of logic but cannot be investigated using techniques from logic or argumentation theory in the case of simulations. The reason is that it is not explicit how the results follow from the premises because simulations are in some sense opaque. Further, in typical examples of computer simulations, no independent case for the validity$^{\mathrm{m}}$ of the conceptual model can be made. Thus, the comparison between simulation outputs and measured data is the preferred method of validation. The argument view can accommodate this but doesn't have particularly interesting implications for it.

Computer simulation is finally closely related to *modeling*. Each simulation implements a mathematical model. If the mathematical model is very simplistic, when compared to the target, it is most natural to say that its equations directly refer to a fictional system that is then used to learn about the target system. So in this case, the computer simulation is closely associated with a fictional model too. Regarding both the mathematical and the fictional model, we may distinguish between the conceptual and the computational model. What the computer simulation qua run of the computer program does is to analyze a computational model. If verification is successful, the results on the computational model can be interpreted in terms of the conceptual model. Since the most interesting part of validation$^{\mathrm{m}}$ of a model establishes that the results obtained for the model can be translated to the target, validation$^{\mathrm{cs}}$ of the simulation is validation$^{\mathrm{com}}$ of the computational model. But this doesn't allow for very interesting insights about the validation$^{\mathrm{cs}}$ of simulations.

Our results are not without irony. Although computer simulations are not experiments (or so has been argued), the methodology of experiment seems to provide the most fruitful perspective on the validation$^{\mathrm{cs}}$ of simulations because many strategies of validating$^{\mathrm{cs}}$ simulations have close parallels in the methodology of experimentation (Parker 2008). After some reflection, this shouldn't come as a surprise, however. Experimentalists can draw on a long track record of successful experimentation. There is nothing like this for thought experiments; also, the arguments behind computer simulations cannot be surveyed and assessed with the techniques from argumentation theory. Modeling, finally, is too close to simulation as to allow for an interesting perspective on the validation$^{\mathrm{cs}}$ of simulations. These days, modeling and computer simulation are so much intertwined that we cannot expect that there is an independent storehouse of recipes for modeling that may then be used for the validation$^{\mathrm{cs}}$ of computer simulations. It is true that there has been, and still is, a lot of modeling without computer simulation. But modeling of this sort has often remained content with qualitative agreement with the target and is not much concerned with predictions of high accuracy, which is a vital issue for many simulations.

A possible objection against the claims of this chapter may be that it has been friendly to various accounts of computer simulation. But do they really fit together? I don't see any problems in this respect. That computer simulations can, and often do, model possible experiments, nicely fits with the view that computer simulations implement models. Now when we talk about the models implicit in simulations, we often do not say that they model a possible experiment. To some extent, this is so because some simulations (viz., those that trace a target system that is not manipulated during an experiment) do *not* in fact model experiments. For other simulations, we can say that they model an experiment but this is not absolutely necessary for their understanding. The view that computer simulations can, and do in fact sometimes do, model possible experiments is also compatible with the idea that they are something like thought experiments or arguments. The reason is that the result of the modeled experiment is inferred using the help of a computer and thus in a way anticipated in thought.

As we can fit together the various accounts of computer simulations accepted in this chapter, we can piece together the implications for the validation^{cs} of computer simulations. What the accounts suggest, for instance, are certain distinctions within the activities of validating^{cs} simulations. Clearly, several ways of drawing such a distinction can be appropriate and useful. The distinctions may further be used to propose certain strategies to validation^{cs} of simulations, and it is no contradiction to say that there are various strategies to validate^{cs} simulations.

Acknowledgements I'm grateful to Julie Jebeile and Nicole J. Saam for useful comments and criticism.

References

Baumberger, C., Knutti, R., & Hirsch Hadorn, G. (2017). Building confidence in climate model projections: An analysis of inferences from fit. *WIREs Climate Change*, *8*(3), e454.

Beisbart, C. (2012). How can computer simulations produce new knowledge? *European Journal for Philosophy of Science, 2*(3), 395–434.

Beisbart, C. (2014). Are we sims? How computer simulations represent and what this means for the simulation argument. *The Monist*, *97*(3), 399–417, (special issue edited by P. Humphreys).

Beisbart, C. (2018). Are computer simulations experiments? And if not, how are they related to each other? *European Journal for Philosophy of Science, 8*(2), 171–204. https://doi.org/10.1007/s13194-017-0181-5.

Beeler, J. R. (1983). *Radiation effects computer experiments*. Amsterdam etc: North-Holland.

Brown, J. R. (1991). *The laboratory of the mind: Thought experiments in the natural sciences*. London: Routledge.

Brown, J. R. (2004). Peeking into Plato's haeven. *Philosophy of Science,* 71, 1126 –1138.

Brown, J. R., & Fehige, Y. (2017). Thought experiments. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2017 Edition). https://plato.stanford.edu/archives/sum2017/entries/thought-experiment/.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297–312. https://doi.org/10.1037/h0040950.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gale (Ed.), *Handbook of research on teaching* (pp. 88ff). Chicago, IL: Rand McNally.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin Company.

Einstein, A. (1961). Relativity, the special and the general theory. A Popular Exposition. London: Methuen (1920, here quoted after edition published by Crown, New York).

El Skaf, R., & Imbert, C. (2013). Unfolding in the empirical sciences: experiments, thought experiments and computer simulations. *Synthese, 190*(16), 3451–3474.

Franklin, A., & Perovic, S. (2016). Experiment in physics. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2016 Edition). https://plato.stanford.edu/archives/win2016/entries/physics-experiment/.

Frigg, R., & Hartmann, S. (2017). Models in science. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2017 Edition). https://plato.stanford.edu/archives/spr2017/entries/models-science/.

Frigg, R. P., & Reiss, J. (2009). The philosophy of simulation: Hot new issues or same old stew? *Synthese, 169,* 593–613.

Gendler, T. S. (2004). Thought experiments rethought and reperceived. *Philosophy of Science, 71,* 1152–1163.

Gupta, A. (2015). Definitions. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2015 Edition). https://plato.stanford.edu/archives/sum2015/entries/definitions/.

Hartmann, S. (1996). The World as a process: Simulations in the natural and social sciences. In R. Hegselmann et al. (Eds.), *Modelling and simulation in the social sciences from the philosophy of science point of view, Theory and decision library* (pp. 77-100). Dordrecht: Kluwer.

Heidelberger, M. (2005). Experimentation and instrumentation. In D. Borchert (Ed.), *Encyclopedia of philosophy*. Appendix (pp. 12–20). New York: Macmillan.

Hughes, R. I. G. (1997). Models and representation. *Philosophy of Science (Proceedings), 64,* S325–S336.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. New York: Oxford University Press.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169,* 615–626.

Imbert, C. (2017). Computer simulations and computational models in science. In L. Magnani & T. Bertolotti (Eds.), *Springer handbook of model-based science* (Vol. 34, pp. 733–779). Cham: Springer.

Koblick, D. C. (1959). An enzymatic ion exchange model for active sodium transport. *The Journal of General Physiology, 42*(3), 635–645.

Lenhard, J. (2011). Epistemologie der Iteration. *Gedankenexperimente und Simulationsexperimente. Deutsche Zeitschrift für Philosophie, 59*(1), 131–145.

Liu, J., Wang, M., Chen, S., & Robbins, M. O. (2010). Molecular simulations of electroosmotic flows in rough nanochannels. *Journal of Computational Physics, 229*(20), 7834–7847.

Massimi, M., & Bhimji, W. (2015). Computer simulations and experiments: The case of the Higgs boson. *Studies in History and Philosophy of Modern Physics, 512,* 71–81.

Morrison, M. (2009). Models, measurement and computer simulation: The changing face of experimentation. *Philosophical Studies, 143,* 33–57.

Morrison, M. (2015). Reconstructing reality: Models, mathematics, and simulations. New York: Oxford University Press.

Naumova, E. N., Gorski, J., & Naumov, Y. N. (2008). Simulation studies for a multistage dynamic process of immune memory response to influenza: Experiment in silico. *Annales Zoologici Fennici, 45,* 369–384.

Nersessian, N. J. (1992). In the Theoretician's laboratory: Thought experimenting as mental modeling. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (Vol. 1992, pp. 291–301).

Nersessian, Nancy J. (2007). Thought experimenting as mental modeling. *Croatian Journal of Philosophy, 7*(2), 125–161.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: SAGE Publications.

Norton, J. D. (1996). Are thought experiments just what you thought? *Canadian Journal of Philosophy, 26,* 333–366.

Norton, J. D. (2004a). On Thought experiments: Is there more to the argument?. In *Proceedings of the 2002 Biennial Meeting of the Philosophy of Science Association. Philosophy of Science* (Vol. 71, pp. 1139–1151).

Norton, J. D. (2004b). Why thought experiments do not transcend empiricism. In C. Hitchcock (Ed.), *Contemporary debates in the philosophy of science*. Blackwell: Oxford, pp. 44–66.

Parker, W. S. (2008). Franklin, Holmes, and the epistemology of computer simulation. *International Studies in the Philosophy of Science, 22*(2), 165–183.

Parker, W. (2009). Does matter really matter? *Computer Simulations, Experiments, and Materiality, Synthese, 169,* 483–496.

Radder, H. (2009). The philosophy of scientific experimentation: A review. *Automatic Experimentation 1*. Open access. http://www.aejournal.net/content/1/1/2.

Saam, N. J. S. (2017). What is a computer simulation? *A Review of a Passionate Debate, Journal for General Philosophy of Science, 48*(2), 293–309.

Schlesinger, S. et al. (1979). Terminology for Model Credibility, *Simulation, 32*, 103–104.

Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science, 71,* 767–779.

Verlet, L. (1967). Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical Review*, *159*(1), 98.

Weisberg, M. (2007). Who is a modeler? *British Journal for Philosophy of Science, 58,* 207–233.

Winsberg, E. (2001). Simulations, models, and theories: Complex physical systems and their representations. In *Proceedings of the Philosophy of Science* (Vol. 68, pp. 442–454).

Winsberg, E. (2003). Simulated experiments: Methodology for a virtual world. *Philosophy of Science, 70,* 105–125.

Winsberg, E. (2009). A tale of two methods. *Synthese, 169,* 483–496.

Chapter 38
How Do the Validations of Simulations and Experiments Compare?

Anouk Barberousse and Julie Jebeile

Abstract Whereas experiments and computer simulations seem very different at first view because the former, but not the latter, involve interactions with material properties, we argue that this difference is not so important with respect to validation, as far as epistemology is concerned. Major differences remain nevertheless from the methodological point of view. We present and defend this distinction between epistemology (the domain of scientific operations that are justified by rational principles aiming at improving current knowledge) and methodology (the domain of scientific operations that are governed by rules, not all of which are grounded on rational, explicit principles). We illustrate this distinction and related claims by comparing how experiments and simulations are validated in evolutionary studies, a domain in which both experiments in the lab and computer simulations are relatively new but mutually reinforcing.

Keywords Theory-ladeness · Holism of confirmation · Opacity · Duhem–Quine problem · Verification and Validation · Calibration · Benchmarking · Parameter tuning · Sensitivity analysis · Measurement errors · Numerical errors · Evolutionary studies · Richard Lenski's Long-Term Experimental Evolution

38.1 Introduction

At first view, computer simulations and experiments in the lab seem to be very different methods to obtain information on target phenomena. There seems to be hardly anything in common between a material setup including measuring instruments and a computer or supercomputer running through lines of code: in the first case, scientists

A. Barberousse (✉)
Sorbonne Université, Paris, France
e-mail: Anouk.Barberousse@paris-sorbonne.fr

J. Jebeile
Université Catholique de Louvain, Louvain-la-Neuve, Belgium
e-mail: julie.jebeile@uclouvain.be

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_38

are engaged in physical interactions with a material system, whereas, in the second case, the physical interactions with the computer are only a means to execute and to correct the computer program. However, considered with respect to their scientific aims, computer simulations and experiments do share common features as they are both designed to learn about physical, biological, or social phenomena that have been selected as worthy objects of investigation. They do so by using mediators that are the system that is experimented upon in the case of experiments and the computer program in the case of simulations.[1] A major common feature of experiments and simulations is that they rely on hypotheses about the relations between the target phenomenon and the mediator. The experimenter hypothesizes that the system she is interacting with faithfully represents her target system, at least with respect to the properties she is interested in, whereas the computer program implements equations that are hypothesized to correctly represent the behavior of the target system. What does the concept of correct representation involve in each case? This will be a major topic of this chapter.

The common features shared by experiments and computer simulations may be revealed, as above, by a description of their aims and theoretical aspects, but do not seem to be connected with the actions that experimenters and simulationists have to perform in order to make sure that their setup (material or code) actually realizes what it is meant to do. In other words, the description of experiments and computer simulations that points out their common features does not seem to pertain to their *validation*. The validation of experiments and simulations is indeed a rather concrete endeavor including specific inquiries that their designers and users have to carry out in order to make the outcomes of their setup acceptable, and thus usable for further investigations. Therefore, the difference between experiments and computer simulations strikes back when scientific practice is taken into account. Experiments involving physical interactions with material systems seem to make a huge difference with respect to validation even though this difference does not play a decisive role from the conceptual perspective. Materiality appears as a major component of the validation of experiments.

Is this so? The aim of this chapter is to show that, despite the obvious differences between experiments and computer simulations, these differences are not that important from the epistemological point of view. Section 38.2 argues in favor of this claim and introduces a distinction between methodological and epistemological components of validation. Section 38.3 discusses an example in a domain that is potentially revealing a lot about validation of experiments and simulations, namely evolutionary biology, because both methods are relatively new therein, compared with physics.

[1] Here and throughout the chapter, we assume that an experiment has a target that may be different from the system experimented on. The experiment thus crucially involves an inference from the system with which the experimenter interacts to the target system. This view of experiments is not uncontroversial, see e.g., Chap. 37 by Beisbart in this volume.

38.2 Epistemology and Methodology of Validation

Both experiments and computer simulations may be usefully seen as artifacts or setups, whose aim is to perform certain processes that result in the production of signals that are in turn interpreted as answers to questions scientists ask about target phenomena. This analysis of experiments and simulations, which focuses on the inferences that both allow, helps us to emphasize that these setups are built up in the first place because they are thought to be able to provide these answers, that is, to be epistemically useful. The reasons why experiments on the one hand, and simulations on the other, are considered as epistemically useful surely differ. For instance, in an experiment, the target phenomenon that is being studied may be reproduced using relevant material properties, whereas in a computer simulation, epistemic usefulness is commonly inferred from the quality of the underlying model. Materiality seems to play an important role in the epistemic usefulness of experiments, whereas the epistemic usefulness of simulations seems to rely on assumptions in the models on which they are based. Materiality makes reproducibility more difficult in experiments than in simulations. Experiments, to be validated, should be reproducible or replicable, and yet, running again the same experiment does not systematically yield identical results, whereas simulations seem less threatened by lack of reproducibility of results. The concept of epistemic usefulness will play a major role in this chapter, as it will provide us with a means to describe how experiments and simulations function *with respect to their epistemic aims.* This framework of analysis will be the ground on which we shall compare validation of experiments and simulations.

In this section, we will first propose a working analysis of validation as potentially applicable to both experiments and simulations, despite differences that we will also discuss. Second, we will introduce a distinction concerning two aspects of validation, namely epistemology and methodology, that will enable us to complete our analysis of the differences between the validation of experiments and simulations.

38.2.1 The Concept of Validation

What does it mean to validate an experiment or a simulation? In order to answer this question, it may be useful to examine what it means to count on an experiment's or simulation's results when they have been validated. From this point of view, it appears that a validated experiment or simulation is one about which it has been established that it provides its users with the types of results they expected from their setup, namely, results that they consider relevant and trustworthy given the questions they ask themselves—which does not preclude the results from being surprising. Accordingly, a validated experiment or simulation is a reliable generating knowledge *process*, while validated results are reliable *products* of such a process. "Validation" thus refers to the intellectual activity by which researchers establish that the results of experiments and simulations are reliable. Another important distinction is between internal and external validation. The latter is about assessing the representational quality of the mediating setup, namely about checking that one can draw correct inferences about the target system from the obtained results. In an experiment, the mediating setup is the experimental system, while in simulations, it is the computer program. Internal validation, on the other hand, is a precondition for external validation, as it is about ensuring that the setup properly functions independently of being representationally well established. As it will become clear, our concept of validation is user-centered in that the assessment of results depends on the users' aims and expectations about these results. Starting from this user-centered concept of validation, we will discuss how it relates to experiments and simulations faithfully representing target phenomena and how both methods may do so in their respective realms.

Now under which conditions does an experiment or simulation deliver the types of results that their users expect (again, not necessarily the results they predicted, but results that, however surprising, they will count as relevant)? It seems fair to say that this happens when the results faithfully represent the target phenomenon (although there may be other cases). Does the concept of faithful representation help us to analyze the concept of validation? Unfortunately, faithfulness in this context is relative to the experimenter's or simulationist's aims and (well informed) expectations about what counts as a relevant result. For instance, depending on the question the experimenter or simulationist asks, the idealizations that are included in the experiment's design or the simulation program can be more or less appropriate. Let us illustrate this point with an example. Drugs are often tested on rats, used as model organisms, in the hope of curing humans from their diseases. In these cases, the idealization involved is that the human bodies' reaction to the treatment depends upon a physiological pathway that has an analog in rats. When this condition is obtained, tests on rats are significant for human beings. When the physiological pathway has no analog in rats, another model organism is required, which shares the relevant material properties with human bodies (see Parker 2008b, 2009 about this point). Let us take another example of idealizations being more or less appropriate depending on the context. The ideal gas model is satisfactory for normal temperatures and pressures; it allows deriving the ideal gas law ($PV = n\mathrm{R}T$). In the model, molecules are assumed to be perfectly elastic spheres, exerting no force, and their volume is negligible in comparison with the volume occupied by the gas. Such a model nevertheless fails to predict the properties of biphasic systems or monophasic systems which evolve toward a biphasic state (phase transitions). Here, the van der Waals equation $P + (a/V2)\ (V - b) = RT$ (where a and b are associated with the intermolecular forces) is used instead of the ideal gas law. By adding attractive and repulsive intermolecular forces, the van der Waals equation yields more accurate results at high temperatures and low pressures than the ideal gas law. Furthermore, the Dieterici equation, which is given by $P(V - b) = RT\ \exp(-a/VRT)$, provides us with even more accurate results than the van der Waals equation in the case of a heavy complex gas. Since the concepts of appropriate idealization and faithful representation are relative to the users' aims, they cannot provide us with a definite criterion for validation of experiment or simulation that would help researchers in any circumstance. To be sure, the designers of experiments and simulations do rely on these concepts in their attempts at validating the results they have obtained, because they can only establish the faithful representation relation when the setup, experiment or simulation, is validated itself. So, there is a strong link between faithful representation and validation. But it does not provide us with any criterion that would have a practical impact. Let us add that the concepts of appropriate idealization and faithful representation are only dependent on the users' minds in so far as the users' aims determine what is relevant. However, whether some results are appropriate or faithful given the aims of the users depends only on the physical properties instantiated in the experiment or on the properties represented in the simulation's outcomes, once analyzed. The latter ones themselves depend on whether the involved equations do capture at least a part of the studied phenomenon.

From the user's point of view, focusing on the many problems that can let the setup go wrong, an experiment or a simulation can be said validated when most doubts about its capacity to answer her questions have been dispelled. This is hardly an all-or-nothing affair. Rather, the setup may be said to be "validated enough" with respect to the questions at hand, as well as the availability of other results on the topic of investigation. The fact that validation comes in degrees (that may not be measurable) has an important consequence, namely, that it may be progressive (Morrison 2015): an experiment or a simulation may reach a higher degree of validation if some of its elements are modified. The gradual nature of validation forces one to take its dynamics into account: validation has to be considered as a process rather than as a state of an experimental or computational setup, about which nothing could be done after it has been reached.

The gradual nature of validation also helps us understand in what sense validation is the object of rational evaluation. As emphasized above, expected answers are usually obtained when the setup correctly represents the target phenomenon. The heart of the (external) validation operation is to check that the representational link between the target phenomenon and the mediator is sound and well established. This link is of semantic nature: it ensures that the processes happening in the artificial setup provide its users with reliable and usable information on the target phenomenon. This is why validation belongs to the realm of reasons as having to do with the assessment of representational links that are realized, e.g., in the material setup of the experiment and in the states of the measuring instruments, but are not of a material nature, for they are established by the activity of the mind. However, the semantic nature of the link between mediator and target phenomenon does not help realize the required monitoring, consisting in the careful examination of each instrument and place where relevant interactions are supposed to occur in order to get the expected outcomes. Such monitoring aims at establishing the material conditions enabling the production of the outcomes; it focuses on material versus rational aspects of the experiment. In a simulation, the monitoring focuses on aspects like the stability of the code, the interactions among modules within the program, etc. All these aspects may be improved step by step and independently of each other because they are conditions for the quality of the outcomes but are not part of the semantic properties of the outcomes properly. This is why, from the practical point of view, the experimenter or simulationist does not draw anything useful and applicable from this semantic link between the target phenomena and the mediator. Help for these practical activities have to be sought from elsewhere.

With respect to the practical concerns of the experimenters and simulationists, it may seem at first sight that experiments are easier to validate as mediators that are able to gain reliable information about the target phenomenon because they share more with the latter. They indeed realize part of the phenomenon itself by instantiating in the lab the very same physical interactions and processes that are being investigated, like electromagnetic interactions, chemical reactions, etc. In order to assess whether this sharing of material characters actually facilitates validation of experiments compared with validation of simulations, let us now examine which actions are taken in order to establish the required link between the target phenomenon and the mediator in both cases of experiments and computer simulations. These actions are meant to ensure the epistemic control that is tantamount to validation. Their main aim is to anticipate noise or unlooked-for outcomes that would make the overall outcomes unreliable. For sure, they differ significantly between the experiment and the simulation because, in the first case, they are directed toward instruments and other, material elements within the setup, whereas, in the other case, they exclusively consist in code checking and rewriting. But do they also differ conceptually? This question has already been addressed by Parker (2008a); however, in the following, we offer a slightly different answer than hers.

In practice, validation includes control of errors and uncertainties. It is a terribly difficult set of tasks because their aim is to try to identify errors that had first been undetected. "Error" here refers to any event during the process (experiment or simulation), or structural feature in the setup, that went wrong (cf. Chap. 5 by Roy in this volume). Validation amounts to remove as much doubt on possible errors or mismatch as possible. *In principle*, if the design of the setup is good and has been implemented correctly, it is not necessary to worry about validation. *In practice*, however, the possibility of error cannot be eliminated beforehand. It is thus necessary to have safeguards in place. These may be tricks of the trade but some of them may fall into the category of formalized methods, like the Verification & Validation method (see below) whose aim is to limit risks of error, that is, to increase control on the ongoing process. Some of these methods are well grounded. This grounding belongs to epistemology. The distinction between methodology and epistemology that we develop in the next section relies on the observation that some practices are justified by rational principles whereas others are best seen as governed by explicit rules that cannot be justified otherwise than by referring to their efficiency. The former obey epistemological principles; the latter may do so, but not always. It is a matter of philosophical discussion on whether they do.

Validation also aims to convince users of the experiment's or simulation's outcomes that major epistemological obstacles (like theory-ladeness, holism of confirmation, and opacity) have been overcome. Let us present some of these obstacles. (There are other epistemological obstacles whose scope is restricted to either experiments or simulations.) First, both experiments and simulations are more often than not theory-laden and therefore face the risk of circularity. Computer simulations are obviously theory-based, whereas experimental measurements may be considered theory-laden as well, since, in order to perform them, scientists rely on the theoretical background of the instrument, which may include some assumptions that pertain to the hypotheses they are examining. Second, both experiments and simulations have to face the Duhem–Quine problem. Experiments do not allow for any hypothesis to be tested in isolation, because an empirical test also requires auxiliary hypotheses. According to the model-oriented version of the Duhem–Quine thesis (Lenhard and Winsberg 2010; Winsberg 2010; Jebeile and Barberousse 2016), the theoretical assumptions in a computer model cannot be tested separately either. When a model's outputs are found to be irrelevant or untrustworthy, with respect to what is known about the investigated system and the models at hand, the modeler has usually no way to cut the model into pieces that could be confirmed or refuted in isolation. As a result, she cannot identify which part is responsible for the failure. Conversely, when a model's outputs are considered relevant, it is not easy to tell whether it is only due to adjustments or to the model's core hypotheses. Third, both simulations and experiments sometimes function as black boxes. An experiment functions like a black box when the experimenter does not (need to) know some (or all) of the physical processes at work in what she observes. At first glance, it seems that a simulation can never be a black box since the program contains the theoretical equations and data that allow for the simulation, or at least supposedly so. In other words, nothing in the program is *prima facie* hidden from the scientists. But simulations involve long and complex calculations that cannot be mentally surveyed by the human mind and are thus epistemically opaque (Humphreys 2004).

Before we go further into the distinction between methodology and epistemology, let us briefly present the Verification and Validation (V&V) method, which has been proposed for computer simulations and which has recently received some philosophical attention (Winsberg 1999, 2010; Oreskes et al. 1994; Morrison 2015). V&V aims to check, first, that the code is well implemented, and will not lead to inaccurate results for mere computer software reasons, and, second, that the simulation model is a good representation of the target system; the first phase corresponds to verification, the second to validation (Oberkampf and Trucano 2002; Oberkampf et al. 2002). Verification is divided into code verification and solution verification (cf. Chap. 11 by Rider in this volume). Code verification emphasizes the good functioning of the code; it is about checking that the code contains no algorithmic error, and functions properly on the chosen hardware and system software. Solution verification aims to assess whether the computed solutions derive satisfactorily from the model assumptions. Consistency and stability of the numerical scheme are controlled for reducing, respectively, truncation errors and computer round-off errors (discretization errors are often controlled a posteriori, i.e., after calculation, through back-and-forth tests which consist in changing the meshing size and the discretization steps, cf. Chap. 11 by Rider and Chap. 12 by Roache in this volume). Validation consists in checking that the computed solutions match the available empirical data, including those obtained from experimental measurements and from already validated simulation models.

38.2.2 Epistemology and Methodology

We can now come back to the distinction between methodology and epistemology with respect to validation. Methodology with respect to validation is the domain of the rules and strategies that are established by the scientific community for avoiding errors and legitimating the results of an experiment or of a simulation in practice. Epistemology with respect to validation is the domain of the principles making either the experimental setup or the computer program a legitimate epistemic mediator by justifying (at least) some of the methodological rules. This distinction has been the topic of a debate between Winsberg (2010) and Morrison (2015). According to the former, the main difference between experiments and simulations is more methodological than epistemological. By contrast, according to the latter, the epistemological grounding of simulations differs significantly from that of experiments. How are we to understand the distinction between methods and their rational foundations? This section is devoted to these questions.

Understood broadly, methodology provides experimenters and simulationists with rules allowing them to control possible errors and disfunctionings. These rules may have a variety of epistemological statuses: some of these are context-dependent (and may depend on the most local details of the setup), whereas others are grounded in well-established regularities about the investigated phenomenon. In simulations, an example of the latter is the acceptance of certain idealizations, like the incompressible flow condition, commonly introduced in models of fluid dynamics, which allows neglecting the influences of pressure and temperature on mass density. Examples of the former are the set of rules aiming at controlling discretization errors in simulations (Roy 2010). This is achieved by making the grid size and the discretization steps as small as necessary. But this, in turn, can create convergence issues as the values of the tested parameters are limited by the computing power of the machine. So, generally, discretization errors are controlled through back-and-forth tests which consist in successively changing the grid size and the discretization steps. The user systematically tests a set of parameters about the grid size and the discretization steps, and then checks whether the calculation has converged. She also has to check whether the results seem plausible, as too coarse a meshing could lead to very approximate or incorrect results.

Be they context-dependent or well grounded, the aim of methodological rules is to establish the epistemic authority and credibility of the outcomes of the experiment or simulation, thus warranting their users to trust these outcomes. Making these rules explicit is a nice way to make progress on the road of providing scientists with an operational concept of validation, allowing for replication. This is why the methodology of validation is so important: as validation is not achieved automatically at the end of some previously determined procedure but is rather the object of evaluation and judgement, relying on explicit rules is a nice way to improve the robustness of validation assessments. However, not all validation rules are explicit, especially in the case of experiments, as some of them are just local routines that may have been adopted a long time ago in the lab without anyone remembering why. Moreover, others may have been based on good reasons without their users being fully aware of these reasons. To say the least, trust in the outcomes of the mediator, experiment or simulation, may be increased by relying on methodological rules, but cannot be definitively established by means of these rules. In order to reach better prospects for validation, it is necessary to examine the foundations of methodological rules, that is, to turn to epistemology. Let us emphasize that the distinction between methodology and epistemology is not strict: whether methodological rules obey epistemological principles is a matter of a philosophical discussion.

Epistemologically well-grounded practices are those whose reasons can be explicated. These reasons may provide methodological rules with foundations or explain why the simulation's or experiment's users are warranted to expect the results they look for by experimenting or simulating. If the setup has been well designed and well implemented, that is, if it relies on sound epistemological foundations, resorting to the methodology of error-seeking and-erasing is dispensable; but it is never in practice since human realizations are affected by contingency. This forces the designers of experiments and simulations to put various methods in place that take care of unforeseen errors.

Let us now survey various operations that are commonly associated with validation, like calibration, parameter tuning, and uncertainty management, in order to assess whether they fall under the scope of epistemological principles or are just suggested by methodological rules, not to mention rules of thumb".[2] This will help us clarify which specific procedures, taken together, compose validation.

- Calibration in experiments and its analog in simulations, when used for validation purposes, is an epistemologically well-grounded practice because it is based on principles that lie at the bottom of the very experimental or simulationist project. In an experiment, calibration is "the use of a surrogate signal to standardize an instrument" (Franklin 1997, p. 31). If the experimental apparatus reproduces known phenomena, calibration gives us reasons to trust the experimental results (Morrison 2009) and may, therefore, be used as part of the validation process. The precise counterpart of this practice in simulations would consist in a comparison between computed solutions and benchmarks as conducted in the second phase of V&V (called "validation"; see above). Benchmarks include available empirical data as well as computed solutions from already well-confirmed models (cf. Chap. 18 by Saam in this volume). But the term "calibration," in the context of simulations, is sometimes used to mean something else, i.e., "parameter tuning" (Trucano et al. 2006). We have to say at this point that terminology is fluctuating a lot among scientists. Overall, calibration in experiments and its analog in simulations are best seen as a practice whose aim may be reached through different means, such as benchmarking, parameter tuning, and sensitivity analysis.
- Benchmarking as a validation practice applied to simulations consists in designing test cases for comparing model outputs with data from other origins. It can be done using various points of reference, e.g., empirical data, computed solutions

[2] As mentioned above, some operations are not clearly grounded on epistemological principles, so that the distinction is not clear-cut and is open to discussion.

from other models, etc. When the data come from experimental sources, benchmarking is about selecting the physical conditions under which the data should be adequately measured. Let us consider a mathematical function representing the evolution of a magnitude against another variable; for example gas pressure against temperature. If variation in pressure is tested for temperature range [0–100 °C], the measured points at 91, 93, and 95 °C will not help validate the model on the entire temperature domain because they are too close to each other. Thus, a selection criterion for data is their regular distribution on the physical domain to test: the data have to cover the entire domain to be useful in the validation process. In order to properly choose the data for comparison, there exist optimization methods, like optimal designs (e.g., Hadamard matrix and Doehlert matrix), which, coupled with least squares regression, aim at choosing the optimal conditions for measuring empirical data in order to validate a mathematical model within a physical domain.

- By contrast, parameter tuning is seldom epistemologically well-grounded, because it is contingent on the details of the realization of the setup. In simulations, parameter tuning is a corrective process which consists in tuning some parameters, i.e., numerical constants in the model, in order to make model results better fit known data. Generally, the results to fit are associated with observables while the parameters that are tuned are not well known (Hourdin et al. 2017). This is why rules of thumb are often used for parameter tuning. Parameter tuning may be viewed as belonging to validation proper or as a precondition for validation (Lenhard 2018). For Lenhard (2018), indeed, adjustable parameters in simulation models cannot be kept separate from the model proper. They are part of the representational content of the model, and as such, have first to be assigned with numbers before the model can be validated. As Lenhard writes, these parameters "also belong to the model form, because without assignment of parameters neither the question about representational adequacy nor the question about behavioral fit can be addressed. [...] Before the process of adjustment, the mere form of the scheme can hardly be called adequate or inadequate." (For Lenhard, parameter tuning is even a precondition for verification, so it reinforces the entanglement between verification and validation.)
- Sensitivity analysis is also governed by methodological rules rather than well-grounded epistemological principles. It aims to quantify uncertainties in model outputs and to trace back to their sources in model inputs. It is about rerunning simulations with slight modifications in relevant parameters supposed to generate uncertainties, and assess how significantly these modifications modify simulation outputs.

Now that we have made clear that establishing a sound representational link via a variety of practices is an important part of validation, let us turn back to the supposed privilege of experiments with respect to validation. The fact that experiments are realized in the material world is supposed to give them some sort of superiority; however, this does not hold when the representational link is poorly established.

With the priority of the representational link in mind, many aspects of validation seem to be common to experiments and simulations:

- comparisons between experiment or simulation results with already established data,
- parameter tuning, i.e., the process of tuning parameters to make model results better fit the database,
- the dynamical character of validation, namely, the fact that it is not established once and for all, but is gradual because it depends on the users' decisions (Morrison 2015).

Some differences are however important to be taken into account. First, among the various methods that contribute to validation, one is specific to simulations, namely, the use of validation experiments (Morrison 2015). They need to be performed under specific parameter values when existing experimental data is lacking for those values. Second, the sources of error are different. In experiments, these are measurement errors and measurement noise. Measurement errors can be due to a malfunctioning detection device, for instance, or to a biased data treatment method (see Parker 2008b and Mayo 1996). They are sometimes so large that no accurate information can be drawn from the measurements (see the example in Tal 2011, which illustrates that a model is sometimes more reliable than an experiment). Measurement noise is due to interference phenomena. If noise is too severe, it may be easier to run a simulation than an experiment. In simulations, there are at least three sources of error, i.e., computer round-off errors, discretization and truncation errors in the case of discretization-based numerical method. Because computers can only store a *finite* set of bits which represent the values of variables obtained at each computational step, computer simulations generate *computer round-off errors*. When the original differential equations have no explicit solution, they need to be discretized, i.e., turned into approximate discrete algebraic equations. Such a transformation generates discretization and truncation errors. *Discretization errors* are produced when initially continuous variables (such as time and space) are replaced by a discrete set of values because the intervals between these values—the "steps"—cannot be infinitely small. *Truncation errors* are introduced when the differential equations are transformed into approximate algebraic equations. For that, the development in a first-order Taylor series is sometimes used as a discrete approximation of the differential operators. It provides at every point a linear relation between the partial differentials of a function and the values of the same function; it becomes nonlinear though, if performed at a higher order. Truncation errors, therefore, correspond to the neglected terms in the development to some order (cf. Chap. 5 by Roy and Chap. 11 by Rider in this volume).

The most important difference between simulations and experiments does not relate to validation but to the epistemological specificity of experiments. This should be recalled to conclude this part: only experiments can refute a hypothesis (once the Duhem–Quine problem has been overcome), whereas simulations, when they provide us with predictions that are consistent with available data, can increase our trust in a hypothesis or a theory that the simulation implements but cannot, in itself, refute it.

38.3 Illustration: Validation of Experiments and Simulations in the Field of Evolution

Let us now turn to an example illustrating how validation in experiments and computer simulations can be compared. We have chosen the field of evolutionary studies to carry out this comparison because it is relatively new, as compared with physics, and thus exhibits a state of research in which many questions are still open and explicitly posed by researchers. Even though the field of experimental evolution is growing quickly (cf. Lenski 2017 for a survey), experiments within it cannot be seen as routine practice in anyway and are therefore accompanied by pressing epistemological and methodological questions. Examples of the former are: Can the simplified environment of the lab say anything meaningful about the complexity of ecological and genetic interactions outside the lab? For experiments only involving one species, are extrapolations to more realistic situations in anyway legitimate? Among the most pressing methodological difficulties of experimental evolution, one may mention the risk of contamination of the bacterial strains that are commonly used as “models”, i.e., as the organisms the generations of which constitute the domain on which the dynamic of evolution is observed and analyzed. As for computer simulations, which is more commonly labeled computer “experiments” or “artificial life” in this context, they began in the early 1990s but are still a minority practice in evolutionary studies. Their promoters thus face the obligation to defend this practice and its relevance to the field. Debates bearing on the validity and relevance of both experiments and computer simulations are thus more explicit than in physics, so that philosophical questions are openly discussed.

We will focus on Richard Lenski’s Long-Term Experimental Evolution (LTEE) setup, which began in 1998 and is still going on. This is undoubtedly *the* major experiment in evolutionary studies so far. In 1998, 12 populations were made out of the same ancestral strain of the bacterium *Escherichia coli*. Between 1998 and 2018, these 12 populations have yielded more than 66,000 generations whose physiological, ecological, and genetic characteristics have been carefully observed and processed by the researchers at Lenski’s lab (an overview of the general protocol can be read here: http://myxo.css.msu.edu/ecoli/overview.html). The choice of *E. coli* for this long-term experiment is easy to understand and participates in the upstream process leading to validation: *E. coli* has been a model organism for long, which means that it has already been used in multiple experiments in all domains of biology; its characteristics are well known and there is no doubt how to have them grow and reproduce or how to measure their ecological and physiological performance. Not only is the model organism in this experiment very well known, making data acquisition easy and robust, but another biotechnological facility has developed since the beginning of the experiment, namely, genetic sequencing. For sure, researchers could sequence small parts of genomes long before the late 1990s, but it was still a long and expensive process. By the 2000s, it had not only become possible to quickly sequence whole genomes, but the costs dropped considerably. LTEE has thus benefited from a major technological advance, which was unexpected at the beginning, but contributed a lot to the interest and validity of the experiment's outcomes, as the availability of genomic data enhances the possibility of hypothesis testing.

In order to analyze how LTEE experiments are validated, it is first necessary to briefly present the general scheme of the experimental protocol, following Lenski 2017 (also described at the *E. coli* long-term experimental evolution project website: http://myxo.css.msu.edu/ecoli). The populations are propagated in a glucose-limited medium by transferring 1% of the volume into fresh medium every day. The 100-fold dilution and resulting regrowth allow about seven generations each day. Samples from each population are periodically stored at −80 °C, where they are available for later study. The frozen cells remain viable (i.e., stay alive in the sense that they can recover their usual physiological capacities once defrosted), so that changes in performance can be analyzed at later times. From this experimental setup and dynamics, it is possible to easily measure the extent of adaption by natural selection by competing bacteria from a later-generation sample against the ancestral strain. Relative fitness is expressed as the ratio of the realized growth rates of the evolved and ancestral bacteria (which are distinguished by colors) as they compete with one another. Fitness measurements are by no means the only outcomes that LTEE enables. For instance, unexpected and interesting complexities have been observed as arising spontaneously: one population, called Ara-2, has diverged into two distinct lineages, called L and S, which have coexisted for over 50,000 generations. During the establishment of this polymorphism, the L ecotype lost its ability to use acetate whereas the S type improved that ability. As a result of the trade-off between growth on glucose and acetate, the two ecotypes can stably coexist (Lenski 2017). Another important observation is related to changes in the mutation rate. Six populations evolved hypermutability caused by mutations affecting either the DNA mismatch repair or the ability to remove certain oxidized bases. The most spectacular observation has been the completely unexpected emergence, after about 31,000 generations, of a population that is able to consume citrate (as opposed to glucose) as its sole carbon source. All these observations are both qualitative and completed with measurements.

The questions that motivated LTEE in the first place were very general ones about evolution that concerned researchers since Darwin's time like: Is the process of adaptation by natural selection invariably slow and gradual, or are there periods of rapid change and stasis? Does fitness eventually reach some maximum level, or can organisms continue to improve on fitness indefinitely, even in a constant environment? Will the replicate populations achieve the same fitness peaks, or will some discover better solutions than others? If fitness trajectories evolve in parallel, does that imply the same underlying genetic changes? How is phenotypic and genetic evolution coupled, both dynamically and functionally? (Lenski 2017) According to the criteria mentioned in Sect. 38.1, the LTEE experiments can be said externally "validated" when they are able to answer these questions, in the sense that if this happens, they do what they are meant to do, namely, provide researchers with the answers to the questions they are interested in. As the LTEE experiments provide multiple answers to questions about selection, adaptation, their dynamics, their genetic basis, etc., part of their external validation is abundantly manifested. This is so because they are the first systematic experimental, large-scale endeavor aiming at providing researchers with empirical evidence about natural selection: the very fact that natural selection occurs in the lab takes part in the external validation of *this* experiment, whereas usually, just getting answers is not enough to validate an experiment—the answers have to be good. But do the LTEE experiments provide *good* answers? This is the other part of external validation. (There are other senses of "validation" that we examine below). In order to examine whether the questions about selection, adaptation, etc., find *good* answers via the LTEE experiments, it is necessary to recall some historical elements about evolutionary studies. As emphasized by Lenski (2017), all these questions have found some answers within the last 30 years, which cohere with the ones provided by the LTEE experiments, but further questions have emerged that relate to the general epistemological concerns bearing on experimental evolution: Do the experiment's outcomes apply to other species? Even though 66,000 generations is the largest number ever reached in an evolutionary experiment, it is still a "drop in the bucket" (Lenski's words) compared with evolution in nature: is not the scope of LTEE's outcomes very limited, after all? It might be argued that the oversimplified and strictly controlled environment of the experiment is a major obstacle to transferring any conclusion outside the lab, although it is the very condition of its success. In order to assess both external and internal validation of the LTEE experiments, it is important to examine how these new questions have been answered. The very first answer is that the experiments' outcomes, briefly presented above, are already remarkable and important *in themselves*, even before trying to extend them outside the lab. Indeed, Lenski and his team were the first to experimentally establish the effects of natural selection (increase in fitness) that were predicted by Darwin. For sure, no evolutionary biologist doubted these effects at the end of the twentieth century, but LTEE provided researchers with no less than an empirical *proof* of the action of natural selection. Second, the apparition of the citrate-eating bacteria is astonishing evidence in favor of the role of contingency in evolution, with all its unexpected effects.

Following this interpretation of LTEE's outcomes, it might be suggested that it is not to be counted as an *experiment* but rather as an *illustration* of evolutionary dynamics, an experiment requiring that its results being extendable to other cases. However, what could be LTEE if not an experiment, with all its controlled conditions and careful measurements that are important for its internal validation, as we show below? LTEE *is* an experiment, not only an illustration of evolutionary dynamics. As suggested by an anonymous referee, this is a reason to count some illustrations as genuine experiments.

From the methodological point of view, the spectacular character of LTEE's results should not hide all the painstaking and minute details that allow for their internal validation: the boring, daily transfer of a tiny part of the bacteria, the freezing away of populations every 500 generations (75 days), the measurement of differential fitness, gene sequencing, genome sequencing, all that for decades, not to speak of the avoidance of contamination risk, mixing up Petri dishes, miscalculation of statistical quantities, errors in Polymerase Chain Reaction or in gene identification, etc. In most published papers, these details are omitted; however, some of them are described in the "Standard Protocols" part of the LTEE website (http://myxo.css.msu.edu/ecoli/standprot.html), which allows one to imagine how tedious internal validation may be in this case. The very existence of this public web page is in itself evidence of the care that is taken in LTEE to avoid error, mismatch, and malfunction. It allows for replication of (part of) the experiment in other labs. More generally, transparency is an often-recognized way to provide readers with evidence that internal validation has been taken care of.

Let us now turn to simulations of evolutionary processes. We shall focus on those developed on the digital platform Avida (Ofria and Wilde 2004), itself derived from the earlier Tierra platform. These computer programs are commonly referred to as "artificial life". They involve populations of simulated organisms (namely, programs) that replicate and are submitted to two sorting processes that simulate natural selection and genetic drift. They give rise to a variety of evolutionary dynamics, including the emergence of complexity (Lenski et al. 1999) and contingency (Lenski 2004).

The study of artificial life is motivated by a number of reasons that contribute to shaping what counts as validation in this domain. The first motivation is that evolution in nature is too slow for the average span of human life, which gives us a relevant scale for meaningful experimentations. Even LTEE can only offer a limited glimpse on evolutionary processes because (i) 66,000 generations, even though an impressive number, cannot be compared with the scales that are relevant for evolutionary processes in nature and (ii) it is confined to one level of genomic complexity (Lenski 2004). The second motivation is that in a simulation, every aspect, however minute, can be scrutinized, whereas on the field, or even in the lab, some data can be difficult or even impossible to obtain; moreover, it is often impossible to carry out enough replications that would allow for high statistical accuracy. By contrast, Avida provides the ability to perform a detailed control over experimental settings and protocols, as well as a variety of measurement tools, and sophisticated methods to analyze and post-process experimental data. For instance, Lenski writes that Avida simulations allow for running "replicated experiments to examine the statistical repeatability of evolutionary dynamics and outcomes" and "rewind[ing] an experiment to any particular point in time and restart the experiment, with replication, from that precise moment" (Lenski 2004). He adds that "the ability to rewind and restart the tape is critical for putting hypotheses that invoke historical contingency into an experimental context". This point is especially important with respect to validation as hypothesis testing is the main goal of artificial life. As mentioned above, evolutionary hypothesis testing is immensely difficult in the field, and even in the lab. The difficulty is increased when it comes to hypotheses involving contingency. Now, the easiness and speed of running Avida simulations allow for obtaining evidence about the role of contingency.

Before turning to validation of Avida, let us briefly mention a point that has been much discussed with respect to these simulations: in the same way, as it was necessary to analyze whether LTEE can actually be counted as an *experiment*, there is some ambiguity involved in the expression "simulation of evolution". Some might prefer to use the word "simulation" to exclusively designate computer models of population genetics involving the discretization of continuous equations, and refer to Avida programs as "experiments". However, "organisms" in Avida are simulated by computer programs; this is the main reason why artificial life may be called "simulations of evolutionary processes", as computer programs, being artifacts, can by no means be conceived as undergoing *natural* evolution—their evolution *simulates* evolution in nature. Let us further emphasize that the question of the very nature of Avida simulations (are they simulations or experiments?) may also raise in other cases, like agent-based simulations in social sciences. Both this question and the fact that many experiments, like those in LHC analyzed by Morrison (2015), include simulations at their very heart and cannot be dispensed with indicate that experiments and simulations are best conceived of as complementary, rather than competing, so that there is no reason to see their respective validation procedures as opposite. As illustrated below, they have much in common.

With respect to validation, Avida simulations both appear similar and different from simulations in physics. On the one hand, internal validation involves the same compound of context-dependent and well-grounded methodological rules as in other types of simulations, whose aim is to avoid errors and disfunctioning in the code. On the other hand, external validation is difficult to carry out for the very reason that triggers researchers to run Avida simulations in the first place: we do not have access to *precise* and controlled evolutionary processes in nature that would provide us with the possibility to clearly identify at which point contingency is explanatory of the presence of certain traits in a population. As a result, we lack satisfactory comparison basis for Avida outcomes, which threatens their soundness: because the prospects of comparison with experimental results are so thin, the question raises whether the simulations provide us with any evidence at all about evolutionary processes in nature.

As with other types of simulations, the very dynamics of validation has allowed for a partial answer to the question of whether Avida simulations' outcomes really bear on evolutionary processes in nature. LTEE's results indeed provide simulationists with a qualitative and quantitative basis that is small, for sure, but well established: external validation can thus rely on precise statistics, even though for a very limited number of cases. The recent availability of LTEE's outcomes has been an important step toward increased external validation of Avida simulations. The reason for this is that external validation of both LTEE and Avida relies on the same epistemological principles, if not on the same methodological rules. These principles themselves derive from the very nature of evolutionary theory, within which it is very difficult to sort effects of selection from effects of genetic drift or contingent events. This difficulty commands incredibly ingenious experimental and data processing design that is common to experiments and simulations in this field. The first step toward a spiraling dynamics of progressive validation for both LTEE (and other studies in experimental evolution) and Avida (and other simulations of evolutionary processes) will undoubtedly be followed by others coming from other in vivo evolutionary experiments. This dynamics is representative of the way both simulationists and experimenters proceed in practice: validation of experiments and simulations is not a short-term effort but evolves as new results and techniques become available.

38.4 Discussion and Conclusion

In this chapter we have argued that, even though, from the practical point of view, validation in experiments and in simulations is obtained with the help of different operations, it has to be analyzed as being based on the same epistemological principles, namely, the ones that allow researchers to take the outcomes of the experiment or simulation at hand as reliable in view of current knowledge in the field. These principles may depend on background theories, as it is the case in the example we have presented: reliable outcomes in evolutionary studies are assessed with respect to the current state of evolutionary theory (a theory that is very well confirmed, but still lacks predictive power), whereas outcomes in fluid dynamics are assessed with respect to the knowledge of the numerical solutions of Navier–Stokes equations that are currently available. There is thus no fundamental difference between validation in experiments and in simulations, only differences regarding practice, which of course require specialized actions and assessments.

With respect to materiality, for sure, it makes experiments and computer simulations different, but this difference is not so important from the epistemological point of view. Major differences remain nevertheless from the methodological point of view, as illustrated by our example in evolutionary studies.

References

Franklin, A. (1997). Calibration. *Perspectives on Science, 5*, 31–80.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J., Balaji, V., Duan, Q., et al. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society.*

Humphreys, P. (2004). *Extending ourselves. Computational science, empiricism, and scientific method*. OUP.

Jebeile, J., & Barberousse, A. (2016). Empirical agreement in model validation. *Studies in History and Philosophy of Science Part A, 56,* 168–174.

Lenhard, J. (2018). Holism, or the erosion of modularity–a methodological challenge for validation, to appear in *Philosophy of Science* (PSA 2016).

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Science Part B, 41*(3), 253–262.

Lenski, R. (2004). The future of evolutionary biology. *Ludus Vitalis, 12*(21), 67–89.

Lenski, R. (2017). Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *The ISME Journal, 11,* 2181–2194.

Lenski, R., Ofria, C., Collier, T., & Adami, C. (1999). Genome complexity, robustness and genetic interactions in digital organisms. *Nature, 400,* 661–664.

Mayo, D. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Morrison, M. (2009). Models, measurement and computer simulation: The changing face of experimentation. *Philosophical Studies, 143,* 33–47.

Morrison, M. (2015). *Reconstructing reality: Models, mathematics, and simulations*. USA: OUP.

Oberkampf, W. L., Trucano, T. G. (2002). Verification and validation in computational fluid dynamics. *Rapport Sandia*. SAND2002-0529.

Oberkampf, W. L., Trucano, T. G., & Hirsch, C. (2002). Verification, validation and predictive capacity in computational engineering and physics. *Applied Mechanics Review, 57*(5), 345.

Ofria, C., & Wilde, C. O. (2004). Avida: A software platform for research in computational evolutionary biology. *Artificial Life, 10*(2), 191–229.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science, 263*(5147), 641–646.

Parker, W. S. (2008a). Franklin, holmes, and the epistemology of computer simulation. *International Studies in the Philosophy of Science*, *22*(2), 165–183.

Parker, W. S. (2008b). Computer simulation through an error-statistical lens. *Synthese*, *163*(3), 371–84.

Roy, C. (2010). Review of discretization error estimators in scientific computing. In *48th AIAA Aerospace Sciences Meeting, Orlando, FL, January 4–7, 2010*.

Tal, E. (2011). From data to phenomena and back again: Computer-simulated signatures. *Synthese, 182*(1), 117–129.

Trucano, T. G., Swiler, L. P., Igusa, T., Oberkampf, W. L., & Pilch, M. (2006) Calibration, validation, and sensitivity analysis: What's what. *Reliability Engineering & System Safety*, *91*(10–11), 1331–1357.

Winsberg, E. (1999). Sanctioning models: The epistemology of simulation. *Science in Context, 12*(2), 275–292.

Winsberg, E. (2010). *Science in the age of computer simulation*. The University of Chicago Press.

Chapter 39
How Does Holism Challenge the Validation of Computer Simulation?

Johannes Lenhard

Abstract Designing and building complex artifacts like simulation models often rely on the strategy of modularity. My main claim is that the validation of simulation models faces a challenge of holism because modularity tends to erode over the process of building a simulation model. Two different reasons that fuel the tendency to erosion are analyzed. Both are based on the methodology of simulation, but on different levels. The first has to do with the way parameter adjustment works in simulation; the second comes from how different groups of software programmers work together. The chapter will conclude by drawing lessons about how holism challenges the validation of simulation and by discussing a corollary to conformational holism: the boundary between validation and verification tends to become blurred, thus undermining a strategy that insists on keeping them separate.

Keywords Kluge · Modularity · Parameterization · Tuning · Verification and validation

39.1 Introduction

Computer simulation models have much in common with mathematical models and it is useful to conceive simulation models as a type of mathematical models. Typical examples of this type differ from other mathematical models in the algorithmic nature, the complexity of internal model dynamics, and the role of visualization, among other factors. While the basic concept of validation is certainly applicable to simulation, those factors that make simulation models a special type of mathematical model create special challenges for validation.

The notion of validation is open to different uses, especially regarding the level of objectivity implied by validation (cf. the Chap. 2 by Beisbart in this volume). The holism challenge presented in this chapter does not depend on how objective and

J. Lenhard (✉)
HLRS, University of Stuttgart, Stuttgart, Germany
e-mail: johannes.lenhard@uni-bielefeld.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_39

final validation is taken to be. The following generic notion of validation suffices: The validation of a model is a process that includes (a) checking whether a model agrees in relevant respects to a target system and (b) modifying the model so that the results of (a) are improved. Validation thus can be a matter of degree, open to revision, and dependent on a purpose for which a model is supposed to be good enough.

Holism is best introduced via its counterpart, namely modularity. Designing and building complex artifacts, from skyscrapers to software, often relies on the strategy of modularity, i.e., breaking down the complex task into subtasks, starting from independent modules that show a certain functionality (solving a subtask) and then putting the modules together in a secured architecture for building the more complex entity. The strategy of modularity in design has a cousin in validation that bears the same name. Breaking down complexity is important for validation, too. How can a model (or another artifact) be improved given that validation step (a) signals that something is going wrong? If the model is modular, one can test modules separately, if not, one must address the model as a whole.

That models need often to be addressed as a whole during validation is not just a purely "philosophical" consideration, but a challenge of actual concern in many fields of simulation. Here is a teaser (before the argumentation starts in the next sections). A former colleague of mine worked for a company that pioneered digital television. He oversaw a development group that permanently updated the software, eliminating those errors and flukes that were detected after market release. The software then was uploaded to all TV users. Of course, the developers were eager to upload only software that they had validated in rigid procedures. Nothing is worse than complaining customers. But the entire system was so extensive that it was impossible to reliably check whether it would perform well in all circumstances. In particular, any newly developed software module would interact with a number of other components, multiplying the number of possible situations where the software has to function properly. Part of the validation procedure was so-called "monkey-testing." They put up a test system with the new software and then went through random series of commands by the remote control. Like a monkey pressing keys—a not so unlikely scenario if the remote is hidden under a sofa cushion and children are playing on the sofa. If the system gets hooked up, they could retrace the combination of keys that caused it and find a patch. The point of this little story is that these engineers had tested all software modules independently, but modularity was imperfect so that the (apparently) valid modules plus their (apparently) valid coupling did not result in a valid overall model. Therefore, the engineers had to move to the "global" level for validating the system as a whole. The story is not so exotic since it shows a typical tendency: Validation procedures cannot fully utilize the modularization strategy. Instead, they have to deal with the entire system at once, i.e., run against a problem of (confirmational) holism (the word "holism" derives from the Greek word "holon" meaning the whole).

My main claim is that the validation of simulation models faces a challenge of holism because modularity tends to erode over the process of building a simulation model. Consequently, the strategy of modularity is threatened to break down. But why is there such tendency to erosion? In fact, there are two different reasons that fuel this tendency. Both are based on the methodology of simulation, but on different levels. The first has to do with how parameter adjustment works in simulation; the second comes from how different groups of software programmers work together.

Section 39.2 introduces and discusses the counteracting concepts of modularity and holism. I will present two arguments in favor of the erosion claim, one from parameterization and tuning (Sect. 39.3), the other from kluging (Sect. 39.4). Both are, in practice, part and parcel of simulation modeling and both make modularity erode. The chapter will conclude by drawing lessons about how holism challenges the validation of simulation (Sect. 39.5) and by discussing a corollary to conformational holism: the boundary between validation and verification tends to become blurred, thereby undermining a strategy that insists on keeping them separate.

39.2 Holism and Modularity—Two Counteracting Concepts

39.2.1 Modularity—The Rational Picture

The term “modularity” is not a particularly philosophical notion. It features prominently in the context of complex design, planning, and building—from architecture to software. Modularity stands for first breaking down complicated tasks into small and well-defined subtasks and then reassembling the original global task with a well-defined series of steps. It can be argued that modularity is the key pillar on which various rational treatments of complexity rest.

The design of complex systems has a long tradition in architecture and engineering. At the same time, it has not been much covered in the literature, because the design was conceived as a matter for experienced craftsmanship rather than analytical investigations. The work of Pahl and Beitz (1984, revised editions 1996, 2007) gives a relatively recent account of design in engineering. The design of complex computer systems (hardware as well as software) is a related field where dealing with complexity became an issue very quickly. Here, one can find more explicit accounts, since researchers could not orient their work at existing traditions. A widely read example is Herbert Simon’s “Sciences of the Artificial” (Simon 1969). Still up to today, techniques of high-level languages, object-oriented programming, etc., make the practice of design change on a fast scale.

One original contributor to this discussion is Frederic Brooks, software and computer expert (and former manager at IBM) and also hobby architect. In his monograph “The Design of Design” (Brooks 2010), he describes the rational model of design that is much more often adopted in practice than explicitly formulated in the theoretical literature. The rational picture starts with assuming an overview of all options at hand. According to Simon, for instance, the theory of design is the general theory of search through large combinatorial spaces (Simon 1969, 54). The rational model

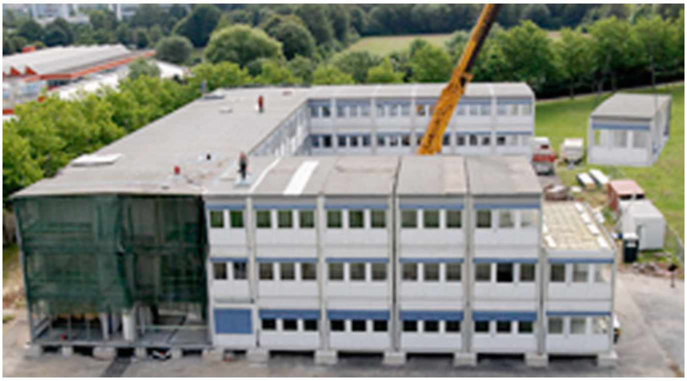


Fig. 39.1 A part of Bielefeld University is built from the container modules Courtesy by Norma Langohr

then presupposes an utility function and a design tree, which are exhausting the space of possible designs. Brooks rightly points out that this is an idealized picture as the space of possible designs is normally unknown. Nevertheless, the design is conceived as a systematic step-by-step process. Pahl and Beitz aim at detailing these steps in their rational order.

A hierarchical order is a key element of the Rational Picture of design and presumes modularity when a higher level task is achieved by putting together modules on a lower level. Let me illustrate this point. Consider first a simple brick wall. It consists of a multitude of modules (bricks), each with a certain form and static properties. These are combined into potentially very large structures. It is a strikingly simple example because all modules are similar.

A more complicated, though closely related, example is the one depicted in Fig. 39.1 where an auxiliary building of Bielefeld University is put together from container modules.

These examples illustrate how deeply ingrained modularity is in our way of building (larger) objects.

Some complex overall task is split up into modules that can be tackled independently and by different teams. The hierarchical structure shall ensure that the modules can be integrated to make up the original complex system; this requires top-down planning. Even if two tasks are achieved by modules from completely independent teams, there must be a (hierarchical) structure according to which the module tasks have been specified so that some higher task will be accomplished by putting together the modular (sub)tasks. Modularity not only plays a key role when designing and building complex systems, it is also of crucial importance when testing whether the system works or not. Validation is usually conceived in the very same modular structure: independently validated modules are put together in a controlled way for making up a validated bigger system. The standard account of how computational models are verified and validated gives very rigorous guidelines that are all based on the systematic realization of modularity (Oberkampf and Roy 2010, see also Fillion 2017). In short, modularity is the key element for designing as well as for validating complex systems. One instance is that a program is tested against a number of benchmark cases before it enters the shelf of approved building blocks for more complex systems.

This observation is paradigmatically expressed in Simon's parable of the two watchmakers which he expounds in his 1962 paper "The Architecture of Complexity" that later has been turned into a chapter in his immensely influential "The Sciences of the Artificial" (Simon 1969). There, Simon investigated the structure of complex systems. The stable structures, so Simon argued, are the hierarchical ones. He expressed his idea by narrating the parable of the two watchmakers named Hora and Tempus (Simon, 1969, 90–92). In his review of Simon, P. Agre describes the setting with the following words:

> According to this story, both watchmakers were equally skilled, but only one of them, Hora, prospered. The difference between them lay in the design of their watches. Each design involved 1000 elementary components, but the similarity ended there. Tempus' watches were not hierarchical; they were assembled one component at a time. Hora's watches, by contrast, were organized into hierarchical subassemblies whose "span" was ten. He would combine ten elementary components into small subassemblies, and then he would combine ten subassemblies into larger subassemblies, and these in turn could be combined to make a complete watch. (Agre 2003)

Since Hora takes additional steps for building modules, Tempus' watches need less time for assembly. However, it was Tempus' business that did not thrive because of an additional condition not yet mentioned, namely some kind of noise. From time to time, the telephone rings and whenever one of the watchmakers answers the call, all cogwheels and little screws fall apart and he has to restart the assembly. While Tempus has to start from scratch, Hora can keep all finished modules and work from there. In the presence of noise, so the lesson goes, the modular strategy is by far superior. Modularity—Agre speaks of the functional role of components—comes out as a necessary element when designing complex systems:

> For working engineers, hierarchy is not mainly a guarantee that subassemblies will remain intact when the phone rings. Rather, hierarchy simplifies the process of design cognitively by allowing the functional role of subassemblies to be articulated in a meaningful way in terms of their contribution to the function of the whole. Hierarchy allows subassemblies to be modified somewhat independently of one another, and it enables them to be assembled into new and potentially unexpected configurations when the need arises. A system whose overall functioning cannot be predicted from the functionality of its components is not generally considered to be well engineered. (Agre 2003)

In a well-engineered software system, one can replace single modules, like replacing the module for matrix inversion by a faster new version, without having to adapt other modules or their connections.

There is an obvious limit to the watchmaker picture, namely the fact that systems have to remain manageable by human beings (watchmakers). There are many systems of practical interest that are too complex—from the Earth's climate to the aerodynamics of an airfoil. Computer models open up a new path here since simulation models might contain a wealth of algorithmic steps far beyond what can be conceived in a clockwork picture. From this point of view, the computer appears as a kind of amplifier that helps to revitalize the rational picture. Do we have to look at a simulation model as a sort of gigantic clockwork? In the following, I will argue that this viewpoint is seriously misleading. Simulation models are different from watches in important ways and I want to focus on the dis-analogy.[1] The central notion for capturing this dis-analogy is holism.

39.2.2 Holism—A Multifaceted Challenge

Holism is a term that appears in different contexts and uses. Variants of holism are relevant in different fields like philosophy of language, or metaphysics. The Stanford Encyclopedia of Philosophy, for instance, includes (sub-)entries on methodological, metaphysical, relational, or meaning holism. Holism generically states that the whole is more than the sum of its parts, meaning that the parts of a whole are in an intimate interconnection, such that they cannot exist independently of the whole, or cannot be understood without reference to the whole. For more details, I refer the reader to the Stanford Encyclopedia. W. V. O. Quine has been especially effective in popularizing the concept in philosophy of science, where one speaks of the so-called Duhem–Quine thesis. This thesis is based on the insight formulated by scientist-philosopher Pierre Duhem: One cannot test a single hypothesis in isolation because any such test depends on "auxiliary" theories or hypotheses, for example, the way the measurement instruments work. Thus, any test addresses a whole ensemble of theories and hypotheses (holism about testing or confirmation).

Lenhard and Winsberg (2010) have discussed the problem of confirmation holism in the context of validating complex climate models. They argued that "due to interactivity, modularity does not break down a complex system into separately manageable pieces" (2010, 256). However, I consider that it is worth to put the thesis into a much more general context, i.e., pointing out a dilemma that is built on the tension between modularity and holism and that occurs quite generally in validating simulations. In fact, the notion of holism has made another appearance in philosophy of simulation, namely in the controversial debate about the philosophical novelty of simulation, see Humphreys (2009) versus Frigg and Reiss (2009) for an instructive example. The latter authors deny novelty in most aspects, but concede that issues of holism might be an exception. Surprisingly, this caveat went nearly unnoticed. In a sense, this chapter confirms both parties: holism is indeed a key concept when reasoning about simulation and it poses the problem of validation in a new way.

[1] There are several dis-analogies. One I am not discussing is that clockworks lack multi-functionality.

39.3 The Challenge Arising from Parameterization and Tuning

In stark contrast to the cogwheel picture of the computer, the methodology of simulation modeling erodes modularity in systematic ways. I want to discuss two separate—though related—aspects, first, parameterization and tuning and second, kluging (also called kludging). Both are, for different reasons, part and parcel of simulation modeling; and both make modularity of models erode.

Parameterization and tuning are key elements of simulation modeling that stretch the realm of tractable subject matter much beyond what is covered by theory. Furthermore, simulation models can make predictions even in fields that *are* covered by well-accepted theories only with the help of parameterization and tuning. In this sense, the latter are success conditions for simulations.

Before we start with discussing an example, let me add a few words about terminology. A parameterization scheme details which parameters (adjustable variables) are used and how they work together. Assigning concrete values to these parameters is often an additional step. Parameterization refers to both steps. There are different expressions that specify what is done with parameters. The four most common ones are (in alphabetical order): adaptation, adjustment, calibration, and tuning. These notions describe very similar activities, but also valuate differently what parameters are good for. "Calibration" is commonly used in the context of preparing an instrument, like calibrating a scale one time for using it very often in a reliable way. "Tuning" has a more pejorative tone, like achieving a fit with artificial measures, or fitting to a particular case. "Adaptation" and "adjustment" have more neutral meanings while the former suggests more strongly that simulation modeling is guided by something it should adapt to. There are interesting reasons why dealing with adjustable parameters counts as good or as bad practice. These reasons deserve a separate treatment. In what follows, I will ignore the differences in terminology.

A typical example for parameterization arises in the simulation of atmospheric circulation. The latter is modeled on the basis of accepted theory (fluid dynamics, thermodynamics) on a grand scale. Climate scientists call this the "dynamical core" of their models and there is more or less consensus about this part. Although the employed theory is part of physics, climate scientists mean a different part of their models when they speak of "the physics". It includes all the processes that are not completely specified within the dynamical core. These processes include convection schemes, cloud dynamics, and many more. The "physics" is where different models differ and the physics is what modeling centers regard as their achievements and try to maintain even if their models change into the next generation.

The physics acts like a specifying supplement to the grand scale dynamics. It is based on modeling assumptions, like which subprocesses are important in convection, what should be resolved in the model, and what should be treated via a parameterization scheme, i.e., a form that details how variables and measurements depend on each other but leave parameters open that control the quantitative details. Often, such subprocesses are not known in full detail, and some aspects (at least) depend on what happens on a sub-grid scale. The dynamics of clouds, for instance, depends on a staggering span of very small (molecular) scales and much larger scales of many kilometers. Hence even if the laws that guide these processes would be known, they could not be treated explicitly in the simulation model. Modeling the physics has to bring in parameterization schemes.[2]

How does moisture transport, for example, work? Rather than trying to investigate into the molecular details of how water vapor is entrained into air, scientists use a parameter, or a scheme of parameters, that controls moisture uptake so that their model fits to known observational data. Often, such parameters do not have a direct physical interpretation, nor do they need one, like when a parameter stands for a mixture of processes not resolved in the model. The important property rather is that they make the parameterization scheme flexible, so that the parameters of such a scheme can be changed in a way that makes the properties of the scheme (in terms of climate dynamics) match some known data or reference points.

From this rather straightforward observation—adequate flexibility is a virtue for parameterization—follows an important fact. A parameterization, including assignments of parameter values, makes sense only in the context of the larger model. Observational data are not compared to the parameterization in isolation, but rather to the parameterizations together with all interactions in the model. The Fourth Assessment Report of the IPCC acknowledges the point that “parameterizations have to be understood in the context of their host models” (Solomon et al. 2007, 8.2.1.3). In other words, a good parameterization for model one needs not—and likely is not—a good parameterization for model two.

The question of whether the parameter value that controls moisture uptake (in our oversimplified example) is adequate can be answered only by examining how the entire parameterization behaves and, moreover, how it behaves in the context of the larger simulation model. Answering such questions would require, for instance, looking at more global properties like mean cloud cover in tropical regions, or the amount of rain in some area. Briefly stated, parameterization is a key component of climate modeling, and tuning (adjusting parameter values) is part and parcel of parameterization.[3] Typically, parameter adjustments are meaningful only with respect to the model as a whole.

It is important to note that tuning one parameter takes the values of other parameters as given, be they parameters from the same scheme, or be they parts of other schemes that are part of the model. A particular parameter value (controlling moisture uptake) is judged according to the results it yields for the overall behavior (like cloud cover). In other words, tuning is a local activity that is oriented at global behavior. Researchers might try to optimize parameter values simultaneously, but for reasons

[2] Parameterization schemes and their more or less autonomous status are discussed in the philosophical literature, cf. Smith (2002), Gramelsberger and Feichter (2011), or Parker (2013).

[3] The studies of so-called perturbed physics ensembles convincingly showed that crucial properties of the simulation models hinge on exactly how parameter values are assigned (Stainforth et al. 2007).

of computational complexity, this is possible only with a rather small subset of all parameters.

Furthermore, the procedure of tuning parameters is not only oriented at the global model performance, it tends to blur the local behavior, like cloud dynamics. This is because every model will be importantly imperfect, since it contains technical errors, works with insufficient knowledge, etc.—which is just the normal case in scientific practice. Now, tuning a parameter according to the overall behavior of the model then means that the errors, gaps, and bugs get compensated against each other (if in an opaque way). Mauritsen et al. (2012) have pointed this out in their pioneering paper about tuning in climate modeling.

In climate models, cloud parameterizations play an important role, because they influence key statistics of the climate and, at the same time, cover major (remaining) uncertainties about how an adequate model should look like. Typically, such a parameterization scheme includes more than two dozen of parameters; most of them do not carry a clear physical interpretation, but rather are motivated from physical reasoning together with pragmatic considerations like inserting flexibility in adjustments. The simulation then is based on the combination of these parameters in the context of the overall model (including other parameterizations). Over the process of adjusting the parameters, these schemes become inevitably interdependent, i.e., parameter values of scheme one depend on the setting of parameters in scheme two and vice versa. I leave aside the fact that models of atmosphere and oceans get coupled, which arguably aggravates the problem.

Tuning is part and parcel of simulation modeling methodology. It poses great challenges, like finding a good parameterization scheme for cloud dynamics, which is a recent area of intense research in meteorology. But when is a parameterization scheme a good one? On the one hand, a scheme is sound when it is theoretically well motivated; on the other hand, the key property of a parameterization scheme is its adaptability. Both criteria do not point in the same direction. One cannot, therefore, optimize both at the same time; finding a balance is still considered as an art. I suspect that the widespread reluctance against publishing about practices of adjusting parameters comes from reservations against aspects that call for experience and art rather than theory and rigorous methodology.

I want to maintain that nothing in the above argumentation is particular to climate science. Climate modeling is just one example out of many. The point holds for simulation modeling quite generally. Admittedly, climate might be a somewhat peculiar case, because it is placed in a political context where some discussions seem to require that only ingredients of proven physical justification and realistic interpretation are admitted. Arguably, this expectation might motivate using the pejorative term of *tuning*. This reservation, however, ignores the very methodology of simulation modeling. Adjusting parameters is a necessary condition for obtaining prediction even in areas where theoretical knowledge is very strong.

Another example will document this. Adjusting parameters is also occurring in thermodynamics, an area of physics with a very high theoretical reputation. So-called equations of state (EoS) describe how, e.g., pressure and temperature depend on each other. The exact form of an equation of state contains much information about chemical and physical properties. The ideal gas equation is the most basic example; it is valid only in the low-pressure limit. However, in fact, using thermodynamics requires working with less idealized equations of state than the ideal gas equation. More complicated equations of state find wide applications also in chemical engineering. They are typically very specific for certain substances and require extensive adjustment[4] of parameters as Hasse and Lenhard (2017) describe and analyze. Clearly, being able to process specific adjustment strategies that are based on parameterization schemes is a crucial success condition. Simulation methods have made applicable thermodynamics in many areas of practical relevance, exactly because equations of state are tailored to particular cases of interest via adjusting parameters.

One further example is from quantum chemistry, namely the so-called density functional theory (DFT), a theory developed in the 1960s that won its originators the Nobel Prize in chemistry in 1998. Density functionals capture the information of the Schrodinger equation, but are much more computationally tractable than the latter. However, only many-parameter functionals brought success in chemistry. The more tractable functionals with few parameters worked only in simpler cases of crystallography, but were unable to yield predictions accurate enough to be of chemical interest. Arguably, being able to include and adjust more parameters has been the crucial condition that had to be fulfilled before DFT could be successfully adopted in computational quantum chemistry. This happened around 1990 when access to computation had become so easy and cheap that exploring and tentatively adjusting parameters became convenient procedures. The upswing of DFT is truly impressive. DFT is by now the most widely used theory in scientific practice, see Lenhard (2014) for a more detailed account of DFT and the development of computational chemistry.

Whereas, the adjustment of parameters—to use the more neutral terminology—is pivotal for matching given data, i.e., for predictive success, this very success condition also entails a serious disadvantage.[5] Complicated schemes of adjusted parameters might block theoretical progress. In our climate case, any new cloud parameterization that intends to work with a more thorough theoretical understanding has to be developed for many years and then has to compete with a well-tuned forerunner. Again, this kind of problem is more general. In quantum chemistry, many-parameter adaptations of density functionals have brought great predictive success but at the same time have rendered the rational reconstruction of why such success occurs hard, if not impossible (Perdew et al. 2005; discussed in Lenhard 2014). The situation in thermodynamics is similar, cf. Hasse and Lenhard (2017).

Let us take stock regarding the first argument for the erosion of modularity. Tuning, or adjusting, parameters is not merely an ad hoc procedure to refine a model, rather it is part and parcel of simulation modeling. Tuning convolutes heterogeneous parts that do not have a common theoretical basis. Tuning proceeds holistically on the basis of global model behavior. How particular parts function often remains opaque. Tuning destructs modularity because local and global considerations are interwoven,

[4] Walter Kohn (1923–2016) shared the price with John Pople (1925–2004). While Kohn received it for DFT, Pople was awarded it for building and promoting computational models in chemistry.

[5] There are other dangers, like over-fitting, that I leave aside.

and various parameter assignments are interdependent. Thus, adjusting parameters in some scheme of module does not depend on how it works in this very scheme, but rather on how it works on the whole model.

Looking back at Simon's clockmaker story, we see that its basic setting does not match the situation in a fundamental way. The perfect cogwheel picture is misleading, because it presupposes a clear identification of mechanisms and their interactions. The preceding examples show that building a simulation model, in contrast to building a clockwork, cannot proceed top-down, i.e., counteracts a hierarchical structure. Moreover, different modules and their interfaces get convoluted during the processes of mutual adaptation.

39.4 The Challenge from Kluging

The second argument for the erosion of modularity approaches the matter from a different angle, namely from a certain practice in developing software known as *kluging* (also spelled kludging).[6] "Kluge" is a term from colloquial language that became a term in computer slang. I remember a precise episode of my childhood when our family accompanied with another befriended one drove toward holidays in two cars. In the middle of the night, while crossing the Alps, the exhaust pipe of our friends before us broke, creating a shower of sparks where the pipe met the asphalt. There was no chance of getting the exhaust pipe repaired, but the father did not hesitate long and used his necktie to fix it provisionally.

The necktie worked as a kluge, which is in the words of Wikipedia "a workaround or quick-and-dirty solution that is clumsy, inelegant, difficult to extend ,and hard to maintain, yet an effective and quick solution to a problem."[7] The notion has been incorporated and become popular in the language of software programming and is closely related to the notion of bricolage.

Andy Clark, for instance, stresses the important role played by kluges in complex computer modeling. For him, a kluge is "an inelegant, 'botched together' piece of program; something functional but somehow messy and unsatisfying", it is—Clark refers to Sloman—"a piece of program or machinery which works up to a point but is very complex, unprincipled in its design, ill-understood, hard to prove complete, or sound and therefore having unknown limitations, and hard to maintain or extend". (Clark 1987, 278)

Kluges carried forward their way from programmers' colloquial language into the body of philosophy due to scholars like Clark and Wimsatt who are both inspired by computer modeling and evolutionary theory. The important point in the present context is that kluges may function for a whole system, i.e., for the performance

[6] Both spellings "kluge" and "kludge" is used. There is not even agreement of how to pronounce the word. In a way, that fits to the very concept. I will use "kluge," but will not change the habits of other authors cited with "kludge."

[7] See https://en.wikipedia.org/wiki/Kludge. Accessed July 10th, 2016.

of the entire simulation model, whereas they do not serve as fixes in relation to the submodels and modules:

> what is a kludge considered as an item designed to fulfill a certain role in a large system, maybe no kludge at all when viewed as an item designed to fulfill a somewhat different role in a smaller system. (Clark 1987, 279)

Since “kluging” stems from colloquial language and since kluging is not seen as a good practice anyway, examples cannot be found easily in the published scientific literature. This observation notwithstanding, kluging is a widely occurring phenomenon. Let me give an example that I witnessed when visiting an engineering laboratory. There, researchers (chemical process engineers) are working with simulation models of an absorption column, the large steel structures in which reactions take place under controlled conditions. The scientific details do not matter here, since the point is that the engineers build their model on the basis of a couple of already existing modules, including proprietary software that they integrate into their simulation without having access to the code. Moreover, it is common knowledge in the community that this code is of poor quality. Because of programming errors and because of ill-maintained interfaces, using this software package requires modifications on the part of the remaining code outside the package. These modifications are there for no good theoretical reason, albeit for good practical reasons. They make the overall simulation run as expected (in known cases); and they allow working with existing software. These modifications thus are typical kluges.

Again, kluging occurs in virtually every site where large software programs are built. Simulation models hence are prime instances, especially when the modeling steps of one group build on the results (models, software packages) of other groups. One common phenomenon is the increasing importance of “exception handling”, i.e., of finding effective repairs when the software, or the model, perform well most of the time, but at rare instances behave in unanticipated and undesired ways. In this situation, the software might include a bug that is invisible (does not affect results) most of the time, but becomes effective under certain conditions. Often extensive testing is needed to find out about unwanted behavior that occurs in rare and particular situations that are conceived of as “exceptions.” The very fact that researchers speak of “exception handling” indicates that they do not aim at a major reconstruction, but at a local repair, counteracting (suppressing) this particular exception. The television developers mentioned in the introduction hunted after cases of this kind. Exception handling can be part of a sound design process, but an increased use of exception handling is symptomatic of excessive kluging.

Presumably, all readers who ever contributed to a large software program know about experiences of this kind. It is commonly accepted that the more comprehensive a piece of software gets, the more energy new releases will require for exception handling. Operating systems of computers, for example, often receive weekly patches. Many scientists who work with simulations are in a similar situation, though not obviously so.

If, for instance, meteorologists want to work on, say, hurricanes, they will likely take a meso-scale (multipurpose) atmospheric model from the shelf of some trusted modeling center and add specifications and parameterizations relevant for hurricanes. Typically, they will not know exactly in what respects the model had been tuned, and also lack much other knowledge about strengths and weaknesses of this particular model. Consequently, when preparing their hurricane modules, they will add measures into their new modules that somehow balance out undesired model behavior. These measures can also be conceived as kluges.

Why should we see these examples as typical, common practice and not as instances where researchers went astray? Because such situations arise from the practices of developing software, and because philosophy should accept these practices as a core part of simulation modeling. Software engineering is a field that was envisioned as the "professional" answer to the increasing complexity of software. And, I frankly admit that there are well-articulated concepts that would in principle ensure that the software is clearly written, aptly modularized, well maintained, and superbly documented. However, the problem is that science *in principle* is different from science *in practice*.

In practice, there are strong and constant forces that drive software development into resorting to kluges. Economic considerations are always a reason, be it on the personal scale of research time, be it on the grand scale of assigning teams of developers to certain tasks. Usually, software is developed "on the move", i.e., those who write it have to keep up with changing requirements and a narrow timeline, in science as well as industry. Of course, in the ideal case the implementation is tightly modularized. A virtue of modularity is that it is much quicker to incorporate "foreign" modules into a software system than developing such a software from scratch.

If these modules have some deficiencies, however, the developers will usually not start a fundamental analysis of how the unexpected behavior occurred, but rather spend their energy adapting the interfaces so that the joint model will work as anticipated in the given circumstances. In common language: repair, rather than replace. Examples reach from integrating a module of atmospheric chemistry into an existing general circulation model up to implementing the new version of the operating system of your computer. Working with complex computational and simulation models seems to require a certain division of labor. Software traveling easily is a major factor that supports such division of labor. At the same time, this will provoke kluges on the side of those who try to connect software modules.

Kluges thus arise from unprincipled reasons: The throw-away code, which has been made for the moment, is nevertheless not replaced later but becomes forgotten, buried in more code, and eventually comes to be permanent. This will lead to a cascade of kluges. Once there, they prompt more kluges, tending to become layered and entrenched.[8]

Foote and Yoder, prominent leaders in the field of software development, give an ironic and funny account of how attempts to maintain a rationally designed software architecture constantly fail in practice.

[8] Wimsatt (2007) writes about "generative entrenchment" when speaking about the analogy between software development and biological evolution, see also Lenhard and Winsberg (2010).

> While much attention has been focused on high-level software architectural patterns, what is, in effect, the de facto standard software architecture is seldom discussed. This paper examines this most frequently deployed of software architectures: the BIG BALL OF MUD. A big ball of mud is a casually, even haphazardly, structured system. Its organization, if one can call it that, is dictated more by expediency than design. Yet, its enduring popularity cannot merely be indicative of a general disregard for architecture. (…) Even systems with well-defined architectures are prone to structural erosion. The relentless onslaught of changing requirements that any successful system attracts can gradually undermine its structure. Systems that were once tidy become overgrown as piecemeal growth gradually allows elements of the system to sprawl in an uncontrolled fashion. (Foote and Yoder 2000, 3)

I would like to repeat the statement from above that there is no necessity in the corruption of modularity and rational architecture. It is rather a tendency. Again, this is a question of science *in practice* versus science *in principle*. "A sustained commitment to refactoring can keep a system from subsiding into a big ball of mud," Foote and Yoder concede (2000, 3). There are even directions in software engineering that try to counteract the degradation into Foote's and Yoder's big ball of mud. The movement of "clean code," for instance, is directed against what Foote and Yoder describe. Robert Martin, the pioneer of this school, proposes to keep code clean in the sense of not letting the first kluge slip in. And surely, there is no principled reason why one should not be able to avoid this. However, even Martin accepts the diagnosis of current practice.

Similarly, Gabriel (1996), another guru of software engineering, makes the analogy to housing architecture and Alexander's concept of "habitability", which intends to integrate modularity and piecemeal growth into one "organic order". Anyway, when he diagnoses the current state, he more or less duplicates what we heard above from Foote and Yoder.

Finally, I want to point out that the matter of kluging is related to what is discussed in philosophy of science under the heading of opacity (like in Humphreys 2009). A highly kluged software becomes opaque. One can hardly disentangle the various reasons that led to particular pieces of code, because kluges are sensible only in the particular context at the time. Furthermore, layered kluges solidify themselves. They make code hard or impossible to understand; modifying pieces that are individually hard to understand will normally lead to a new layer of kluges—and so on. In this important sense, simulation models are historical objects. They carry around—and depend on—their history of modifications. There are interesting analogies with biological evolution that have become a topic when complex systems had become a major issue in discussing computer use. Winograd and Flores, for instance, come to a conclusion that also holds in our context here: "each detail may be the result of an evolved compromise between many conflicting demands. At times, the only explanation for the system's current form may be the appeal to this history of modification." (Winograd and Flores 1991, 94)

Thus, the brief look into the somewhat elusive field of software development has shown us that two conditions foster kluging. First, the exchange of software parts that is more or less motivated by flexibility and economic requirements. This practice thrives, where infrastructure is networked. Second, iterations and modifications are easy and cheap. Due to the unprincipled nature of kluges, their construction requires repeated testing which examines whether they actually work in the factual circumstances. Kluges hence fit to the exploratory and iterative mode of modeling that characterizes simulations (according to Lenhard 2016). Kluging makes modularity erodes because a kluge is oriented at the behavior of the whole model. This is the second argument why modularity tends to erode in simulation modeling.

39.5 The Limits of Validation

We have seen that the power and scope of simulation come with a tendency toward the erosion of modularity. Holism and the erosion of modularity are two sides of the same coin. Hence, holism is driven by the very features that make simulation so widely applicable! It is through adjustable parameters that simulation models can be applied to systems beyond the control of theory (alone). What does the tendency toward holism mean for the validation of computer simulations?

First and foremost, the strategy of modularity breaks down in validation. Step (a), testing the model, is demanding for any complex model regardless of modularity. In extreme cases, like mathematics, one can prove complex theorems by putting together proven simpler theorems. In most usual situations, however, this bottom-up strategy gets into troubles. Assuming that validated modules plus secured architecture would guarantee a valid complex entity appears to be a risky strategy. In any case, step (b), modifying the model appropriately, would be much more feasible given a modular structure, because researchers could limit their analysis to particular modules. Without the modular structure, the search for appropriate modification has to take into account the entire model.

Second, there is a corollary to holism that challenges prominent concepts of validation. In the context of simulation models the community speaks of verification and validation, or “V&V” (cf. Chap. 4 by Murray-Smith and Chap. 41 by Beisbart in this volume). Both are related, but the unanimous advice in the literature is to keep them separate. While verification checks the model internally, i.e., whether the software indeed captures what it is supposed to, validation checks whether the model adequately represents the target system. A standard definition states that “verification [is] the process of determining that a model implementation accurately represents the developer’s conceptual description of the model and the solution to the model.” While validation is defined as “the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model”. (Oberkampf and Trucano 2000, 3) Though there is some leeway of defining V&V, the gist of it is entailed in the saying: verification checks whether the model is right,[9] while validation checks whether we have the right model.

[9] This sloppy saying should not obscure that the process of verification comprises an entire package of demanding tasks, see Chap. 10 by Rider in this volume.

Due to the increasing use and growing complexity of simulations, the issue of V&V is itself a growing field in simulation literature, witnessed by this volume. One example is the voluminous monograph by Oberkampf and Roy (2010) that meticulously defines and discusses the various steps to be included in V&V procedures. The first move in this analysis is to separate model form from model parameters. Each parameter then belongs to a particular type of parameter that determines which specific steps in V&V are required. Oberkampf gives the following list of model parameter types:

- measurable properties of the system or the surroundings,
- physical modeling parameters,
- ad hoc model parameters,
- numerical algorithm parameters,
- decision parameters,
- uncertainty modeling parameters. (Oberkampf and Roy 2010, Sect. 13.5.1, p. 623)

My point is that the adjustable parameters we have discussed in this chapter are of a type that is evading the V&V fencing. These parameters cannot be kept separate from the model form, since the scheme prior to adjusting parameters does not aim at representational (nor behavioral) adequacy. A cloud parameterization scheme makes sense only with parameter values already assigned and the same holds for a many-parameter density functional. Before the process of adjustment, the mere form of the functional does not offer anything to be called adequate or inadequate. In simulation models, as we have seen, (predictive) success and adaptation are entangled.

It is not possible to first verify that a simulation is “right” in view of a model before tackling the “external” question, namely whether the model is right. The separation of verification and validation thus cannot be fully maintained in practice. Performance tests hence become the main handle for confirmation. This is a version of confirmation holism that points toward the limits of analysis. This does not lead to a complete conceptual breakdown of verification and validation. Rather, holism comes in degrees[10] and is a pernicious tendency that undermines the verification–validation divide.[11]

Holism leaves intact all performance tests that work on the whole computer model or code. But holism challenges the Rational Picture of design. This challenge works, if you want, from “within,” It is a central part of simulation modeling, and the way in which it works in practice challenges the Rational Picture by making modularity erode.

Acknowledgements This chapter looks back on a history of improvements. I would like to thank the reviewers of this volume for their helpful suggestions and the active audiences at the “How to

[10] I thank Rob Moir for pointing this out to me.

[11] My conclusion about the inseparability of verification and validation is in good agreement with Winsberg’s more specific claim in Winsberg (2010) where he argues about model versions that evolve due to changing parameterizations, which has been criticized by Morrison (2015). As far as I can see, her arguments do not apply to the case made in this paper, which rests on a tendency toward holism, rather than a complete conceptual breakdown.

Build Trust…" conference, Hanover 2015, and the PSA 2016 (whose proceedings include a shorter version of this paper). I also thank the DFG for funding under SPP 1689.

References

Agre, P. E. (2003). Hierarchy and history in Simon's "Architecture of complexity", *Journal of the Learning Sciences, 3*, 413–426.

Brooks, F. P. (2010). *The design of design*. Boston: Addison-Wesley.

Clark, Andy. (1987). The kludge in the machine. *Mind and Language, 2*(4), 277–300.

Fillion, N. (2017). The vindication of computer simulations. In J. Lenhard, & M. Carrier (Eds.), *Mathematics as a tool*, 137–56. Boston studies in history and philosophy of science 327. Cham: Springer.

Foote, B., & Joseph, Y. (2000). Big ball of mud. In H. Neil, F. Brian & R. Hans (Eds.), *Pattern Languages of Program Design 4* (= *Software Patterns*. 4). Addison Wesley, 2000. Retrieved July 25, 2018, from http://laputan.org/pub/foote/mud.pdf.

Frigg, R., & Reiss, J. (2009). The philosophy of simulation. Hot new issues or same old stew? *Synthese*, *169*(3), 593–613.

Gabriel, R. P. (1996). *Patterns of software. Tales from the software community*. New York and Oxford: Oxford University Press.

Gramelsberger, G., & Johann F. (eds.). (2011). *Climate change and policy. The calculability of climate change and the challenge of uncertainty*. Heidelberg: Springer.

Hasse, H., & Lenhard, J. (2017). On the role of adjustable parameters. In J. Lenhard, & M. Carrier (Eds.), *Mathematics as a tool*, Boston Studies in History and Philosophy of Science, forthcoming.

Humphreys, Paul. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*(3), 615–626.

Lenhard, J. (2016). Computer simulation. In P. Humphreys (Ed.), *Oxford handbook in the philosophy of science* (pp. 717–737). New York: Oxford University Press.

Lenhard, Johannes. (2014). Disciplines, models, and computers: The path to computational quantum chemistry. *Studies in History and Philosophy of Science Part A, 48,* 89–96.

Lenhard, Johannes, & Winsberg, Eric. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Modern Physics, 41,* 253–262.

Mauritsen, T. et al. (2012). Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, 4.

Morrison, Margaret, & Reality, Reconstructing. (2015). *Models, mathematics, and simulations*. New York: Oxford University Press.

Oberkampf, William L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge: Cambridge University Press.

Oberkampf, W. L., & Trucano, T. G. (2000). Validation methodology in computational fluid dynamics. In *American Institute for Aeronautics and Astronautics* (2000–2549).

Pahl, G., & Beitz, W. (1984). *Engineering design: A systematic approach*. Revised editions in 1996, 2007. Heidelberg: Springer.

Parker, W. (2013). Values and uncertainties in climate prediction, revisited. *Studies in History and Philosophy of Science*.

Perdew, J. P., Ruzsinsky, A., Tao, J., Staroverov, V., Scuseria, G., & Csonka, G. (2005). Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *The Journal of Chemical Physics*, 123.

Simon, Herbert A. (1969). *The Sciences of the Artificial*. Cambridge: The MIT Press.

Smith, Leonard A. (2002). What might we learn from climate forecasts? *Proceeedings of the National Academy of Sciences USA, 4*(99), 2487–2492.

Solomon, S., et al. (eds.). (2007). *Contribution of working group i to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Stainforth, D. A., Downing, T. E., Washington, R., & New, M. (2007). Issues in the interpretation of climate model ensembles to inform decisions. *Philosophical Transactions of the Royal Society, 365*(1857), 2145–2161.

Wimsatt, William C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA and London: Harvard University Press.

Winograd, T., & Flores, F. (1991) *Understanding computers and cognition*. Reading, MA: Addison-Wesley.

Winsberg, Eric. (2010). *Science in the age of computer simulation*. Chicago: University of Chicago Press.

Chapter 40
What Types of Values Enter Simulation Validation and What Are Their Roles?

Gertrude Hirsch Hadorn and Christoph Baumberger

Abstract Based on a framework that distinguishes several types, roles and functions of values in science, we discuss legitimate applications of values in the validation of computer simulations. We argue that, first, epistemic values, such as empirical accuracy and coherence with background knowledge, have the role to assess the credibility of simulation results, whereas, second, cognitive values, such as comprehensiveness of a conceptual model or easy handling of a numerical model, have the role to assess the usefulness of a model for investigating a hypothesis. In both roles, values perform what we call first-order functions. In addition, cognitive values may also serve an auxiliary function by facilitating the assessment of credibility. As for a third type of values, i.e. social values, their legitimate role consists in specifying and weighing epistemic and cognitive values with respect to practical uses of a simulation, which is considered a second-order function. Rational intersubjective agreement on how to specify and weigh the different values is supposed to ensure objectivity in simulation validation.

Keywords Epistemic values · Cognitive values · Social values · Credibility · Relevance · Practicability

40.1 Introduction

In validating computer simulations, scientists typically appeal to values such as empirical accuracy and robustness when they assess how well simulation results fit observation-based data and results from other simulations and thus how credible the results are. Scientists use additional values to also assess the usefulness of a simulation. When, for instance, a representational model is supposed to simulate the regional climate 20 years from now, they also consider comprehensiveness, complexity, and spatio-temporal resolution as well as how easily the model can be implemented and

G. Hirsch Hadorn (✉) · C. Baumberger
ETH Zurich, Institute for Environmental Decisions, Zürich, Switzerland
e-mail: hirsch@retired.ethz.ch

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_40

run on a computer so as to calculate results that are not only credible but also useful (Flato 2011, p. 797).

There is an ongoing debate about the various values applied in scientific practice, including the influence of interests and goals pursued in society at large on decisions about the application of those values. While some defend the position that societal considerations should not be used in the validation of a simulation (Betz 2013; Parker 2014), others argue that their influence cannot be eliminated (Winsberg 2010) or is even desirable and justified under certain conditions (Elliott and McKaughan 2014; Intemann 2015). In this chapter, we systematize the ongoing debate on the diverse types of values and specifically focus on simulation validation and on the assessment of scientific representations more generally with the help of a framework that determines the legitimate roles of values (Hirsch Hadorn 2018). This framework provides a conceptual basis for discussing the different types of values and their roles in simulation validation.

The chapter is structured as follows: We introduce our framework in Sect. 40.2. Section 40.3 deals with epistemic values. Since their role in simulation validation is addressed in many other chapters of this volume, we focus on the sociological critique that questions the possibility of a reasonable distinction between epistemic and social values. In Sect. 40.4, we review proposals that argue for legitimate roles of cognitive and social values in the context of credibility assessments. In Sect. 40.5, we move on to discussing proposals that argue for an application of cognitive and social values for assessing the usefulness of a simulation, including its relevance to the hypothesis under investigation and its practicability for the user. Thereafter, Sect. 40.6 deals with proposals from decision theory and alternatives for evaluating the credibility and the usefulness of simulations in multi-criteria assessments. Section 40.7 summarizes the results of our application of the evaluative framework to computer simulations and closes with open questions and suggestions for further research.

40.2 The Framework

As a crucial preliminary, we have to clarify the pertinent terminology. First, the term "computer simulation" is sometimes restricted to the execution of a computer program that explores the (approximate) behavior of a numerical model. We, by contrast, use the term (or "simulation" for short) in a broader sense for denoting a series of steps and their products, which includes the development of a conceptual model, the construction or selection and adaptation of a numerical model, its implementation and execution on a computer, the interpretation of the outcomes, and the drawing of inferences with respect to the target system (Winsberg 2015). Second, we use the terms "evaluation" and "validation" interchangeably for the assessment of the products (or parts or aspects of them) that follow from the steps of a simulation in the broad sense introduced above; these products include the outcomes, which we call "simulation results" (even though "simulation" is used in the narrow sense here). Finally, we adopt a broad understanding of "value" that includes any kind of consideration that can be used as a criterion in an assessment (Laudan 1984, p. xii; McMullin 1983, pp. 4–6).

The framework for systematizing the debate on values in the assessment of scientific representations is based on Hirsch Hadorn (2018). It distinguishes three types of values that are also applied in the validation of simulations: epistemic, cognitive, and social values. We hold that these values fulfill at least one of two possible roles (i.e. assessing the credibility of simulation results or the usefulness of simulation models), and that they do this in three distinct functions (i.e. a first-order, a second order, or an auxiliary function).

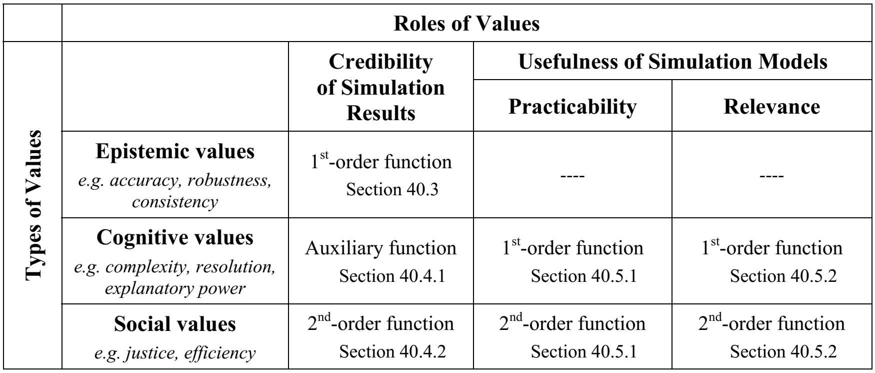
To begin with, the performance of simulation results in terms of empirical accuracy and consistency (partly) determines the warranted degree of belief or confirmation, i.e., the credibility of the simulation results. The term "epistemic value" refers to the values used in such a way for assessing credibility. Besides the traditional empiricist criteria, epistemic values include e.g. robustness of results and coherence with background knowledge and thus values that are frequently applied while assessing the credibility of simulation results. Other values such as comprehensiveness, complexity, resolution, and easy handling are properties of scientific representations that result from idealizations, that is to say, deliberate simplifications and distortions made in designing, adapting, and implementing a representational model (McMullin 1985; Weisberg 2007, 2013). These properties can serve as values in assessing how useful a simulation model is. Usefulness includes the questions of whether the model is relevant, for instance whether its degree of resolution is appropriate for investigating a hypothesis in a given context, and how practicable the simulation model is, for instance regarding the handling of the code. We use the term "cognitive value" (Douglas 2013, p. 800; Lacey 2004, p. 27; Laudan 1984, p. 42, 2004, p. 19; Levi 1986, pp. 36–46) to distinguish properties of scientific representations that serve to assess their usefulness from epistemic criteria for assessing the credibility of simulation results. Idealizations in the process of constructing and implementing a simulation model have consequences for the credibility of simulation results; a more complex simulation model may lead to a higher degree of accuracy, for instance. Since it is not the degree of complexity as such that is indicative of the credibility of simulation results but rather the degree of accuracy that is affected by the degree of complexity, complexity is not used as an epistemic value in such a case. Other authors do not distinguish between these roles of values and use "epistemic value," "cognitive value," and "scientific value" interchangeably (Kuhn 1977; Hempel 1988/2000). Finally, the term "social value" refers to moral and prudential principles and considerations that are typically employed to evaluate goals, decisions, and actions in society at large. Thus, while "social value" is often used as a synonym for "value", we understand the term in a narrower sense. To be specific, we claim that social values are legitimately used for specifying and weighing the epistemic and cognitive values with respect to a given application such as determining what degree of accuracy of results is required or whether it is acceptable to trade off a higher fit of simulation results for an easier handling of the simulation model in a given case. Legitimate uses of social values thus relate to both roles, i.e. assessing the credibility of simulation results and assessing the usefulness of models, but, as we will see, their functions differ from those of epistemic and cognitive values.

The performance of the products of various simulation steps in terms of epistemic and cognitive criteria is used for justifying evaluative judgments, for example for deciding that simulation A is better than simulation B because the results are more accurate in the case of A than in the case of B, or because the mathematical model can be more easily turned into a computer code in the case of A than in the case of B. The extent to which (the products of) A and B instantiate a particular criterion determines the evaluative ordering of A and B regarding this criterion. Such an ordering is called a "value relation" (Rabinowicz 2008). This use of "value" that refers to the (numerical) degree to which (the products of) A and B instantiate a criterion needs to be distinguished from our use of the term for evaluative criteria. Because simulation validation rests on several criteria, it inevitably faces the challenges of multi-criteria assessments. In consequence, improving the performance of a simulation with respect to criteria like simplicity or easy handling of model equations may come at the cost of a lower degree of empirical accuracy, for example. Philosophers of science disagree on whether this is a critical issue regarding the objectivity of the assessment of scientific representations at all and under which conditions trading off performance in terms of epistemic values against performance in terms of other types of values is legitimate (Levi 1986, pp. 40–42; Kuhn 1977, p. 322).

Both epistemic values that are used to assess the credibility of simulation results, and cognitive values that are used to assess the usefulness of simulation models, have what we call a first-order function because performance regarding these criteria directly indicates the credibility of a simulation result or the usefulness of a simulation model. Cognitive criteria can also be helpful in determining the extent to which simulation results instantiate epistemic criteria. If, for instance, a hypothesis is to be tested by implementing a simulation model with the intention of calculating results for specified initial conditions, researchers appreciate when the simulation model can be easily implemented on a computer and running the simulation does not require much time. While these advantages are due to properties of the simulation model like simple functions in the numerical model, the performance of the model in terms of simplicity is not in itself an indication of the credibility of the simulation results. This legitimate role of cognitive values in assessments of the credibility of simulation results is an auxiliary rather than a first-order function.

Social values that include moral and prudential considerations such as justice and efficiency are also important when it comes to assessing the credibility of simulation results and the usefulness of simulation models. Their role consists in specifying and weighing epistemic and cognitive values in order to account for the purpose and the context of a simulation. Since they operate on values with a first-order function, social values perform a second-order function. Figure 40.1 provides an overview of the various types, roles, and functions of values, and shows how this framework is reflected in the structure of this chapter.

The current debate on values and their legitimate roles in simulation validation rests on a broader philosophical debate on science and values. We refer to this broader debate when we systematize the types, roles, and functions of values in simulation



		Roles of Values		
		Credibility of Simulation Results	**Usefulness of Simulation Models**	
			Practicability	**Relevance**
Types of Values	**Epistemic values** *e.g. accuracy, robustness, consistency*	1st-order function Section 40.3	----	----
	Cognitive values *e.g. complexity, resolution, explanatory power*	Auxiliary function Section 40.4.1	1st-order function Section 40.5.1	1st-order function Section 40.5.2
	Social values *e.g. justice, efficiency*	2nd-order function Section 40.4.2	2nd-order function Section 40.5.1	2nd-order function Section 40.5.2

Fig. 40.1 Matrix for structuring legitimate roles of different values in the evaluation of computer simulations

validation since most issues about the role of values in the evaluation of hypotheses, theories, and models directly relate to simulation validation. This is no surprise because a simulation model is typically grounded in basic theories or empirical regularities of a given field, and it can be used for hypothesis testing when it is implemented under specific initial conditions. One might ask, however, whether the focus on the investigation of simulations specifies the scope of the discussion by centering on particular perspectives on values. We nevertheless maintain that turning from hypotheses, theories, and models to simulations does not require any modifications of the proposed framework. It might be the case that additional criteria come into play, e.g. in the evaluation of simulation steps like discretizing model equations and turning them into a computer code. Whether this is actually the case needs to be examined by studying different values in detail, which is beyond the scope of this chapter. Still, we suspect that even additional values are likely to fit the suggested framework.

40.3 A Defense of Epistemic Values that Assess the Credibility of Simulation Results

Simulation results are typically evaluated for their credibility through an assessment of their empirical accuracy, i.e., their distance to observation-based data, and, in the case of ensemble modeling, the robustness of simulation results, i.e., the degree of agreement between simulation results of different models or model versions. Performance of simulation results regarding empirical accuracy and robustness, both of which serve as epistemic values, is measured using elaborated quantitative metrics (often called validation metrics, see Chap. 12 by Marks and Chap. 17 by Saam in this volume). Since the application of epistemic values in simulation validation is addressed in many chapters of this volume, this section focuses on a defense of the distinction between epistemic and social values.

This distinction and the very idea of epistemic assessment, i.e., the thesis that performance in terms of epistemic values—as distinguished from non-epistemic values—can justify credibility, has been criticized by radical social epistemologists, as Longino (2002) has called them. Machamer and Osbeck (2004, pp. 77–78), for instance, claim "that epistemic, or cognitive, values are ineliminably social, and that this is so in many important ways." They see "little practical or theoretical worth in demarcating these by kind, that is, as social or epistemic." The idea of dismissing epistemic justifications for scientific theories altogether can be traced back to the macro-sociological "Strong Programme" (Barnes and Bloor 1982). The aim of this program is to "demystify" the epistemic normativity of empiricist standards like empirical accuracy and consistency by revealing that their use in scientific research is causally determined by powerful interests that are at work in society at large. Goldman and Blanchard (2016) call such approaches the "debunking form" of social epistemology and distinguish them from approaches of "positive social epistemology" that are compatible with epistemic normativity. Radical social epistemology is faced with serious problems. One is that radical social epistemology itself is subject to its own claim, which means that it has to conceive its own claim as being causally determined. Another problem is that radical social epistemologists use "social" in the sense of being shared by all or at least many members of a community. As Laudan (2004, p. 22) reminds us, "this sense of the term social is so broad as to be vacuous."

Longino, a moderate social epistemologist, uses "social" in the sense of being the result of interactions between diverse individuals of a community. She argues that what is required if individuals are to constitute a community is "not a set of shared substantive beliefs, but a set of public standards to which community members appeal in critical discursive interaction" (Longino 2002, p. 148). These standards address, among others, publicly recognized forums for discussion, uptake of criticism in a discussion, public criteria for evaluation, and tempered equality of intellectual authority of the participants (Longino 2002, pp.128–135). Longino's notion of the social thus points to the fact that the normative status of considerations that function as criteria in scientific assessments is based on social interaction and agreement in accordance with explicit standards. Thus, for moderate or positive social epistemologists, something can only legitimately be used as value if it is socially justified. This is compatible with a distinction of values into different kinds. More specifically, the use of the term "social" for qualifying procedures and standards of justification does not preclude the use of a different notion of "social" for distinguishing between social, epistemic, and cognitive values in scientific assessments, as we propose in our framework. It is necessary, however, to be explicit about the sense in which "social" is used.

40.4 Roles of Cognitive and Social Values in Assessing the Credibility of Simulation Results

Throughout the research process, from identifying and framing a problem to considering the implications of the results for further scientific inquiry and possible practical uses, scientists need to decide between alternative options for how to proceed. Social values (i.e., goals, moral and prudential principles of society) may enter research at any stage by figuring among the criteria that are considered in taking these decisions (Machamer and Wolters 2004, pp. 4–5). This is uncontroversial except when it comes to the justification of scientific results. Traditional empiricism adheres to its principles of empirical accuracy and logical consistency as the only legitimate criteria for assessing credibility. There is, however, a debate on whether and, if so, how other values can, or even must, legitimately contribute to an epistemic assessment.

In Sect. 40.4.1, we discuss the proposal that cognitive values can be instrumental in determining the extent to which a simulation instantiates epistemic values (Douglas 2013; Laudan 2004; Steel 2010). This instrumental use of cognitive values is compatible with an empiricist position, but this position has been challenged since the 1940s. It has been claimed that further values are needed for deciding whether hypotheses, theories, and models should be accepted because assurance provided by the instantiation of epistemic values is limited for principle reasons. In Sect. 40.4.2, we discuss the problem of inductive risk that is involved in generalizing empirical findings from a sample (Rudner 1953) and the corresponding problem for simulations that concerns the uncertainty in inferences from fit between simulation results and data or results of other simulations to hypotheses about the target.

40.4.1 Assistance in the Assessment of Performance in Terms of Epistemic Values

The extent to which cognitive values are instantiated does not as such count as warrant for belief in, or as a confirmation of, a model or simulation results (Hempel 1988/2000, p. 223; Laudan 2004, pp. 16–18; van Fraassen 1980, p. 88). Nevertheless, properties like broad scope or simple handling may facilitate the assessment of the performance of a result or a claim regarding epistemic values. Douglas (2013, p. 800), for instance, argues that simpler claims and broad-scope claims “are easier to work with. Simpler claims are easier to follow through their implications. Broadly scoped claims have more arenas (and more diverse arenas) of application to see whether they hold.” In this type of use, cognitive values relate to credibility but not in a first-order function. A first-order function would require that instantiating a cognitive value to a higher degree—e.g., a claim of broader scope—indicated a higher degree of credibility, which is not the case. Since the instantiation of certain cognitive values may only facilitate the determination of performance in terms of epistemic values, the function of cognitive values in assessing credibility is an auxiliary function.

40.4.2 Determining Minimal Probabilities for Accepting or Rejecting a Hypothesis

In empirical research, the practice of drawing inductive inferences from the findings in a sample to support the claim that an empirical hypothesis holds in general entails the twofold risk of error that has been called "inductive risk" (Hempel 1965, p. 92). The risk of false positives consists in accepting a hypothesis that does not hold in general, whereas the risk of false negatives consists in rejecting a hypothesis that does actually hold in general. Rudner (1953, p. 2) and others have claimed that accepting or rejecting a hypothesis—e.g. that electromagnetic pollution increases the risk of suffering from cancer—is a task for scientists qua scientists, and that in doing so they ought to anticipate and consider possible social consequences of an erroneous decision. Because the inclusion of ethical criteria is indispensable when acceptance or rejection of scientific hypotheses with foreseeable social consequences are to rest on an appropriate level of risk, scientists qua scientists are required to make ethical value judgments. Hence, science cannot be value-free. An analogue of inductive risk in the context of assessments of simulations is the uncertainty inherent in inferences from the degree of empirical accuracy and robustness of simulation results to hypotheses about the target. As a background for discussing the role and the function of social values in the evaluation of such inferences, we first review the extensive debate on inductive risk.

In the initial stage of the debate, "the argument from inductive risk against value-free science," as it is typically referred to, was questioned in two respects. Jeffrey (1956) claimed that scientists qua scientists should characterize the uncertainty of a hypothesis by referring to the evidence available but not accept or reject hypotheses. Levi (1960), by contrast, contended that scientists qua scientists have to accept or reject hypotheses based on their minimal probability, but he argued that this does not imply that the criteria for minimal probability must be ethical in nature and relate to social risks. Instead of having a practical objective, decisions on the acceptance or rejection of hypotheses can also have a theoretical objective like, for instance, arriving at statements with desirable characteristics such as simplicity or explanatory power.

The more recent debate has started with the paper "Inductive risk and values in science" by Douglas (2000) who defends Rudner's position. Availing herself of studies on dioxin and its potential effect on cancer as an example, she argues that scientists need to rely on social values in various methodological decisions as, for example, in considering what should count as relevant evidence and how to structure and classify the data in order to account for the social consequences of error (p. 559). Douglas calls this use of social values an indirect role, which is to be distinguished from the direct role of epistemic values (p. 564). Elliott and others have rightly asked for further clarification of this distinction (Elliott 2011, p. 305; Elliott and McKaughan 2014, p. 2). Douglas refers to Hempel (1965, p. 92) who distinguishes between (i) rules of confirmation that determine the degree of evidential support and (ii) rules of acceptance that determine the requisite strength of evidential support for acceptance. This distinction between different levels at which values operate can be adduced for clarification. It leads to the distinction between two types of functions that we have already introduced in Sect. 40.2: epistemic values have a first-order function if the performance of a simulation in terms of these values determines how well a hypothesis is confirmed. Social values operate on epistemic values because they serve as rules for specifying minimal probabilities for accepting or rejecting the hypothesis in question. Thus, they operate on a meta-level and perform, in this sense, the second-order function of specifying thresholds. We use "second-order function" since "indirect role" may also be used for the auxiliary functions that cognitive values take in epistemic assessments (see Sect. 40.4.1).

In the assessment of simulations, the uncertainty affecting inferences from the degree of empirical accuracy and robustness of simulation results to hypotheses about the target—as, for instance, discussed for the case of climate predictions by Intemann (2015, pp. 225–226), Steele (2012) and Winsberg (2010, pp. 93–102)—is an analogue of inductive risk. Risks of error arise from difficulties in modeling and in particular from the need for parameterizations that account for the net effect of climate processes (e.g. cloud formation) that cannot be explicitly modeled and may have major consequences for the outline of climate policy. Some take Jeffrey's (1956) position and restrict the tasks of scientists to characterizing the probability (Parker 2014, pp. 24–27) or the possibility (Betz 2013, p. 213) of simulation-based predictions. A case in point is the approach of the Intergovernmental Panel on Climate Change (IPCC) to assessing uncertainty in findings on climate change by means of a metric that consists of seven categories for characterizing the probability of an outcome, ranging from 0–1% probability (= exceptionally unlikely) to 99–100% probability (=virtually certain) (Mastrandrea et al. 2010, p. 3). In order to justify the position of the IPCC, one may refer to Bayesian confirmation theory that dispenses with the notion of acceptance of a hypothesis by conceiving confirmation in a purely quantitative and dynamic way, i.e., the increase or decrease in the epistemic probability of a hypothesis over the course of Bayesian updating (Strevens 2006; see Chap. 7 by Beisbart in this volume).

Still, already Rudner had argued that assigning a probability to a statement is "nothing more than the acceptance by the scientist of the hypothesis that the degree of confirmation is p" (Rudner 1953, p. 4). Thus, refraining from accepting or rejecting an empirical hypothesis only sets the problem one step back instead of eliminating it. This last move may be resisted by claiming that scientists do not need to determine precise probabilities for their beliefs to inform decision makers. However, Steele (2012) has shown that there is a more general problem that supports Rudner's conclusion. If scientists translate the uncertainty of their beliefs into broader categories—e.g. those of the IPCC's confidence scale that rest on the five qualifiers "very high," "high," "medium," "low," and "very low" for expressing confidence in a finding (Mastrandrea et al. 2010, p. 3)—for the purpose of informing policy makers, the codification of the uncertainty of their complex beliefs in terms of the broader categories is not fully determined. Referring to the IPCC's uncertainty rating, Steele argues that "scientists cannot avoid making value judgments, at least implicitly, when deciding how to match their beliefs to the required scale" (Steele 2012, p. 899). Steele's argument speaks against proposals that defend value-free science by using hedged hypotheses, i.e. hypotheses that make the uncertainty explicit (Betz 2013, p. 212), or by classifying predictions as epistemic possibilities (Betz 2016, p. 139), if these characterizations and classifications are not fully determined projections.

If social values cannot be eliminated, they need to be made explicit and justified. Some social epistemologists argue that scientists can legitimately avail themselves of social values in decisions on methodology if they are democratically endorsed (Intemann 2015, p. 219; Kitcher 2001). Wilholt (2009, pp. 94–99) claims that relying on democratically endorsed values is likely to prevent researchers from having a bias that arises when they just follow their personal preferences in dealing with problems like inductive risk. It is an open question, however, whether an agreement on goals and values in a democratic procedure can be achieved in reasonable time.

40.5 Roles of Cognitive and Social Values in Assessments of the Usefulness of Simulation Models

So far, we have discussed the roles of cognitive and social values in assessments of the credibility of simulation results. In what follows, we turn to broader conceptions of validation that justify further functions of cognitive and social values. Such a broader conception has famously been proposed by Kuhn (1977) in his discussion of theory choice and has also become common practice in the validation of simulations. In "Objectivity, Value Judgment and Theory Choice," Kuhn lists five characteristics of what constitutes a good scientific theory: accuracy, consistency, broad scope, simplicity, and fruitfulness. He stresses that these characteristics are not exhaustive but "individually important and collectively sufficiently varied to indicate what is at stake" (Kuhn 1977, p. 321). Kuhn replaces the empiricist goal of assessing how well-confirmed theories are by the goal of assessing how good theories perform with regard to a range of criteria. While the empiricist position takes epistemic values to be the only scientifically legitimate criteria for theory choice, Kuhn conceives epistemic criteria as belonging to a larger set of evaluative criteria with first-order functions as we call them. The extent to which all these criteria are met counts in the assessment of a theory. Levi (1960, 1986) and Hempel (1988/2000) likewise proposed conceiving the principal goal of science as characterized by both cognitive and epistemic values.

What are the characteristics of a good simulation? Flato's answer, to which we referred in the introduction, stating that simulations should "provide useful and reliable results" (Flato 2011, p. 797), is not restricted to simulations of earth system models, which are the subject of his review. Besides assessing whether results are credible, or "reliable", in Flato's terms, the usefulness of a simulation needs to be assessed as well. Usefulness consists of two aspects. First, a simulation is only useful if it represents the target in those respects that are relevant to answering the question under investigation. Second, the practicability of a simulation for its users is a further aspect of its usefulness. In Sect. 40.5.1, we discuss the use of cognitive and social criteria for assessing the practicability of a simulation. In Sect. 40.5.2, we turn to the use of values for assessing the relevance of a model to the hypothesis under investigation.

40.5.1 Accounting for the Practicability of Simulation Models

Practicability issues typically come to the fore when simulations are evaluated for uses in applied contexts. It has been proposed that simulations in such contexts ought to be assessed with regard to questions like "'Is it easy enough to use this model?', 'Is this hypothesis accurate enough for our present purposes?', 'Can this theory provide results in a timely fashion?' and 'Is this model relatively inexpensive to use?'" (Elliott and McKaughan 2014, p. 5). The characteristics of simulations that increase their practicability for the users perform a first-order function in usefulness assessments, which contrasts with their auxiliary function in epistemic assessments (Sect. 40.4.1).

Taking account of the needs of users requires consideration not only of aspects of practicability but also of aspects of credibility. In this vein, Elliott and McKaughan (2014, pp.15 and 19) contend that in applied contexts, performance in terms of values that relate to the needs of the users of a simulation may legitimately trump performance in terms of empirical accuracy if this is in accordance with two individually necessary and collectively sufficient conditions: these values (i) must be explicit as criteria that govern the appraisal, and they (ii) should get priority only to the extent to which they advance the goals associated with the assessment. Expedited Risk Assessment of hazardous substances (Cranor 1995) is typically used as a case in point in this regard (e.g., Elliott and McKaughan 2014, Steel 2010). Cranor found that the social costs of fairly accurate but very slow procedures are greater than those that arise from Expedited Risk Assessment methods. The latter are less accurate but much faster methodologies for the assessment of risks, not least because the more accurate procedures cannot keep up with newly emerging information on hazardous substances.

We have two comments regarding Elliott and McKaughan's conditions for legitimately trading performance in terms of values for credibility for performance in terms of values for practicability. First, these conditions do not imply that credibility of simulation results can always be legitimately traded off for practicability of simulation models. They are compatible with there being cases in applied contexts such as warnings about extreme weather events that require high performance regarding, e.g., empirical accuracy and robustness of predictions, while empirical accuracy might be less important to other goals such as fundamental understanding in basic research. Second, Cartwright (2006, 2012) rightly argues that appropriate "evidence for use" is not simply a question of the degree of accuracy. Because idealizations are indispensable in standardized controlled trials and basic research in general, an appropriate conception of epistemic assessment for applied contexts has to account for causal complexity and the variability of conditions at work in the given context of use.

The social values held by users, by contrast, function as second-order criteria for how to specify cognitive and epistemic values. Economic interest in an efficient use of a simulation, for instance, specifies and prioritizes properties like easy handling, low demand on resources in running the simulation, and receiving results in a short time. Thus, societal interests and goals may guide decisions on idealizations in various steps of a simulation. This is a second-order function that operates on cognitive values and is analogous to the function of ethical values in the argument from inductive risk, where ethical values specify the rules for the acceptance of a hypothesis in relation to the kind of practical use to which it is put. Winsberg (2010, p. 131) concludes from this that simulations cannot be value-free in the sense of being free from social values.

Considerations on practicability are not restricted to applied research, however, but are important in research in general since decisions on idealizations need not be guided by social values. Parker (2014), for instance, contends that the choice as to which physical processes should be included in a climate model, i.e., its comprehensiveness, can be a relevant aspect and thus argues against Winsberg's (2010, p. 131) claim that science cannot be free of social values. She highlights practicability issues when she points out "that such choices can also be influenced or even determined by pragmatic factors. [...] For instance, the scientist may already have in hand some computer code for process P but not for processes Q, R, or S. Or, they might judge that it will be much easier to incorporate P than to incorporate Q or R or S" (Parker 2014, p. 27). Although Parker's argument defeats the claim that social values necessarily enter simulation evaluation in a first-order function, it does not imply that using social values as second-order values is illegitimate in every case, however.

40.5.2 Accounting for the Relevance of Simulation Models

Simulations are representational tools that are used for answering specific questions about their target by providing suitable hypotheses. Hence, they should be assessed with regard to whether they are appropriate for their purpose (Parker 2009; Frigg et al. 2015). Hypotheses can be distinguished according to their kind (such as prediction or explanation) , the specific variables or phenomena investigated, the temporal or spatial scales of interest, their specificity, or the allowed margin of error (Baumberger et al. 2017, p. 4). Whether a simulation is appropriate for answering a specific question does not depend on the empirical accuracy and the robustness of the simulation results alone. What is also required is that the model represents the target in a way that is relevant to answering the question at issue. For example, if a simulation generates empirically accurate results but does not represent the causal structure of the target, it is not possible to infer from the accuracy of the results that the simulation provides adequate explanations.

Relevance is typically discussed in terms of what features of the target system need to be represented and what data ought to be considered as evidence for or against the investigated hypothesis (Douglas 2000, pp. 569–572; Intemann 2015, pp. 220–221; Peschard and van Fraassen 2014). Since simulations are idealized representations of their target, questions of relevance also arise with regard to how the represented is represented. Climate simulations provide a case in point. Properties like simplicity of model structure, elegance of equations (e.g. symmetric equations), and explanatory power of functions that describe basic mechanisms are relevant to understanding the dynamics of the global climate system. If, however, the purpose consists in predicting future regional climate change, relevant properties include comprehensiveness with respect to processes, complexity of their representation, high spatio-temporal resolution, and explanatory power of functions that describe sub-grid processes (Held 2005; Knutti 2008; Schmidt and Sherwood 2015).

Which features of the target system need to be represented and which data ought to be considered as evidence for or against the investigated hypothesis depends on the context in which a specific hypothesis is investigated. This needs to be taken into consideration for specifying and weighing cognitive values of simulations as well (van Fraassen 1980, p. 89). There are scientific and societal goals that are connected with the execution of a simulation. The scientific goal is typically conceived as improving the performance of scientific representations with respect to cognitive and epistemic values like explanatory power, broad scope, and empirical accuracy. Hempel, for instance, argues that together these values "reflect a profound and widely shared human concern whose satisfaction is the principal goal of scientific research—namely, the formation of a general account of the world which is as accurate, comprehensive, systematic and simple as possible and which affords us both understanding and foresight" (Hempel 1988/2000, p. 216). Hence, the use of cognitive values as first-order criteria in assessing simulations is justified if they are part of what characterizes the scientific goal to which the scientific community is committed. These abstract values are typically regarded as universal criteria of good science (Hempel 1988/2000, p. 216; Kuhn 1977, p. 321; Laudan 2004, p. 16; van Fraassen 1980, p. 88). Levi (1986), by contrast, argues for a pluralistic account of scientific goals and does so by asserting that an application of the criteria in question requires that they are properly specified for the purpose at issue. Without doubt, interpreting the specified criteria used in the assessment of simulations as different specifications and weightings of a universal set of ambiguous and vague criteria would be a difficult task. Still, even if this were done successfully, this would still not provide a strong argument for their legitimate use since the way in which the criteria have been specified and weighted for the purpose of the simulation needs to be justified. A telling example in this regard is Rochefort-Maranda's analysis of "simplicity" in the context of model selection, which distinguishes five concepts of simplicity and shows "that the importance that we give for a particular notion of simplicity will depend on the goal that we pursue when we select a model" (Rochefort-Maranda 2016, p. 269).

Since the different values need individual specification and weighing in accordance with the problem to which a simulation is applied, a pluralistic account of values seems reasonable for the purpose of assessing simulations. Up to now, there are only a few systematic analyses of standards for relevance available, however. One is Weisberg's (2007, 2013) account of representational ideals. Weisberg develops this account in the context of the justification of different kinds of idealizations in science. His thesis holds that not only different intended uses like explanation or prediction (Weisberg 2007, p. 635) but also the state of the art in the field and considerations of practicability like easy handling (Weisberg 2007, p. 641) require different representational ideals that, in turn, guide different kinds of idealization.

Societal goals can also contribute to the justification of using specific cognitive values. We have already seen this in connection with practicability issues, but it is also the case if the performance of a hypothesis in terms of cognitive criteria is relevant to using the simulation results for pursuing a societal objective. This has been pointed out by Intemann: "Social and ethical aims are also relevant to determining the sorts of features that adequate models will have. For example, some argue that adequate Integrated Assessment Models (IAMs) must not only provide information about the aggregate impacts to be expected from climate change, but also information about the distribution of those impacts to ensure that costs and benefits can be distributed equitably [...] Thus, models that fail to account for the distribution of effects will be inadequate for developing ethical policies" (Intemann 2015, p. 220). Still, even within such an account, the societal goals have to be made explicit (Elliott and McKaughan 2014), and must be justified by democratic bodies and procedures in society (Kitcher 2001; Longino 2002).

40.6 Simulation Validation as a Multi-criteria Assessment

The application of a broad variety of criteria for credibility and usefulness in scientific assessments faces two difficulties, namely the ambiguity inherent in each criterion and the need to define trade-offs between the criteria. Kuhn argues that the criteria should be specified and weighed by using informal considerations, based on expert judgment and discussion: "The considerable effectiveness of such criteria does not [...] depend on their being sufficiently articulated to dictate the choice of each individual who subscribes to them. Indeed, if they were articulated to that extent, a behavior mechanism fundamental to scientific advance would cease to function" (Kuhn 1977, p. 330; see also Hempel 1988/2000, p. 221). Levi criticized what he takes to be an implication of Kuhn's position, namely that controversies about values in the assessment of hypotheses are settled "through persuasion or coercion" (Levi 1986, p. 41) because this is not the only alternative to the algorithmic procedures that Kuhn rejects. However, a general hierarchical ordering of values is no reasonable option even for authors who criticize Kuhn by arguing for lexical priority of epistemic over cognitive values (Douglas 2016, p. 619).

As an alternative to Kuhn's approach, Levi proposed to regard controversies concerning cognitive values as cases of "decision making under unresolved conflicts" (Levi 1986, p. 46). He suggests settling unresolved conflicts through further inquiry into how to specify the values. In the case of cognitive values, for instance, it needs to be determined "what is to count as simple or as explanatory powerful" (Levi 1986, p. 39). Further suggestions that rest on the framework of decision theory have recently been put forward. Okasha (2011), for instance, reads Kuhn's thesis that there is no algorithm for choices with multiple criteria as the claim "that there are many algorithms, all equally acceptable" (Okasha 2011, p. 110)—a reading which Kuhn himself considers in a hypothetical dialogue with a Bayesian, though without subscribing to it (Kuhn 1977, 227–330). In order to criticize Kuhn as he reads him, Okasha reconstructs theory choice as a social choice with each individual representing one of Kuhn's five criteria for theory assessment and shows that Arrow's impossibility theorem holds for theory choice. This means that there is no algorithm whatsoever that determines the way of choosing and consequently no rational choice among theories if multiple criteria are to be applied. Okasha explores several strategies for avoiding this result, which is possible, for instance, if performances regarding the criteria can at least be measured on a ratio scale. While this is feasible as regards empirical accuracy, it seems less so with respect to some cognitive values like explanatory power. Gaertner and Wüthrich (2015) take a different approach to model theory choice as a rational choice. They propose scoring rules for measuring performance in terms of each of the criteria. This allows for inter-criteria comparability while aggregation of scores can be used to determine a weak ordering of alternative theories. These scoring rules can be applied so as to account for the usefulness of a simulation for investigating the hypothesis in a given context. Hence, it seems that there is a way of formally determining which of the considered theories or simulations works best for a given purpose and context, all criteria considered.

However, when performances are aggregated with the intention of ranking alternative simulations, information about the individual performances regarding the various criteria gets lost or becomes hidden. Sometimes, this is a disadvantage. For controlling for, and improving on, both relevance and accuracy, for instance, it may be required to keep an eye on the extent to which a simulation meets each criterion over the course of the various steps of a simulation. Doing this provides a basis for improving the simulation in an iterative assessment procedure (Diekmann and Zwart 2014; Winsberg 2010).

40.7 Summary and Conclusion

Our approach to systematizing the debate on values in simulation validation rests on a distinction between different types, roles and functions of values. Types of criteria that are used in the assessment of simulations are epistemic values, cognitive values, and social values. The legitimacy of their use depends on their function in the assessment of a simulation. Against this background, we suggested distinguishing between three formal types of functions, namely first-order functions, second-order functions, and auxiliary functions. Values with first-order functions are applied to simulations if information about the credibility and the usefulness of simulations and their results is to be generated. Values with second-order functions specify and weigh the values with first-order functions with respect to the purpose of a simulation, i.e., testing the hypothesis about the target to be investigated by means of the simulation, and with respect to its context. Values with auxiliary functions, in turn, take an instrumental role in assessments of how well a simulation performs in terms of values with a first-order function.

Epistemic criteria like accuracy, robustness, and consistency have a first-order function. The extent to which these values are instantiated indicates how credible simulation results are. Cognitive values are a consequence of idealizations in the various steps of conducting a simulation. They can perform several functions. Values like comprehensiveness, complexity, scope, explanatory power, and easy handling are used in a first-order function if they act as criteria for the usefulness of a simulation model. Usefulness involves, on the one hand, the question as to whether a simulation model is relevant, i.e., whether the target is appropriately represented with respect to investigating the hypothesis about the target, and, on the other hand, the question as to how practicable the simulation model is for its users. Values with first-order functions need to be specified in relation to the sort of hypothesis and the context of investigation. If the practicability of a simulation model that instantiates cognitive values is simply of help in assessing the performance in terms of epistemic values, cognitive values serve an auxiliary function for assessing the credibility of simulation results. Legitimate use of social values is restricted to a second-order function, i.e., to specifying and weighing the values with a first-order function, if simulation results are intended to be useful for pursuing a societal objective, as in scientific policy advice.

An evaluation of how credible simulation results are and of how useful a simulation model is by means of a broad range of criteria often requires a trade-off between performances regarding epistemic and cognitive criteria with first-order functions. Normative principles for deciding on legitimate trade-offs refer to the goals of implementing a simulation. In the case of societal goals as, for instance when simulations are conducted with the intention of informing policy-makers about the effectivity or possible risks of certain instruments so as to enable them to address these risks appropriately, it is legitimate to use social values for specifying and weighing the pertinent values with a first-order function. In the case of scientific goals, such as promoting the state of the art in the field by improving the complexity, resolution, simplicity, accuracy, and efficiency of a simulation, these goals serve to specify and weigh the values with first-order functions.

Neither determining the degree of performance in terms of particular criteria nor weighing criteria against each other for deciding on acceptable trade-offs needs to be a matter of subjective, that is to say arbitrary, individual preferences. Instead, simulation validation understood as a multi-criteria assessment is objective in the relevant sense if it is based on rational intersubjective agreement on how to specify and weigh the values with a first-order function (Spohn 2004). Various avenues have already been explored for answering the question of how rational agreement could be achieved, but much further work in this direction is still necessary. Elementary questions to be addressed are the following: Under which conditions would rational choice provide an appropriate framework? What models of decisions under uncertainty are applicable? Are there alternative frameworks such as procedural principles for explicit deliberation that could be used to reach agreement on the specification and weighting of values with first-order functions? The avenue that will be followed will also frame the stance on how to conceive scientific rationality since it seems no longer tenable to build only on empiricist principles.

Acknowledgements We thank Claus Beisbart, Johannes Lenhard and Nicole J. Saam for valuable comments and suggestions to improve this manuscript.

References

Barnes, B., & Bloor, D. (1982). Relativism, rationalism and the sociology of knowledge. In M. Hollis & S. Lukes (Eds.), *Rationality and relativism* (pp. 21–47). Oxford, UK: Blackwell.

Baumberger, C., Knutti, R., & Hirsch Hadorn, G. (2017). Building confidence in climate model projections: An analysis of inferences from fit. *WIREs Climate Change*, e454, https://doi.org/10.1002/wcc.454.

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science, 3,* 207–220. https://doi.org/10.1007/s13194-014-0095-4.

Betz, G. (2016). Accounting for possibilities in decision making. In S. O. Hansson & G. Hirsch Hadorn (Eds.), *The argumentative turn in policy analysis: Reasoning about uncertainty* (pp. 135–169). Cham: Springer. https://doi.org/10.1007/978-3-319-30549-3_6.

Cartwright, N. (2006). Well-ordered science: Evidence for use. *Philosophy of Science, 73,* 981–990. https://doi.org/10.1086/518803.

Cartwright, N. (2012). Presidential address: Will this policy work for you? Predicting effectiveness better: Philosophy helps. *Philosophy of Science, 79,* 973–989. https://doi.org/10.1086/668041.

Cranor, C. (1995). The social benefits of expedited risk assessments. *Risk Analysis, 15,* 353–358. https://doi.org/10.1111/j.1539-6924.1995.tb00328.x.

Diekmann, S., & Zwart, S. D. (2014). Modeling for fairness: A Rawlsian approach. *Studies in History and Philosophy of Science, 46,* 46–53. https://doi.org/10.1016/j.shpsa.2013.11.001.

Douglas, H. E. (2000). Inductive risk and values in science. *Philosophy of Science, 67,* 559–579. https://doi.org/10.1086/392855.

Douglas, H. E. (2013). The value of cognitive value. *Philosophy of Science, 80,* 796–806. https://doi.org/10.1086/673716.

Douglas, H. E. (2016). Values in science. In P. Humphreys (Ed.), *The Oxford handbook of philosophy of science* (pp. 609–630). New York: Oxford University Press.

Elliott, K. C. (2011). Direct and indirect roles for values in science. *Philosophy of Science, 78,* 303–324. https://doi.org/10.1086/659222.

Elliott, K. C., & McKaughan, D. J. (2014). Nonepistemic values and the multiple goals of science. *Philosophy of Science, 81,* 1–21. https://doi.org/10.1086/674345.

Flato, G. M. (2011). Earth system models: An overview. *WIREs Climate Change, 2,* 783–800. https://doi.org/10.1002/wcc.148.

Frigg, R., Thompson, E., & Werndl, C. (2015). Philosophy of climate science, part II: Modelling climate change. *Philosophy Compass, 10,* 965–977. https://doi.org/10.1111/phc3.12297.

Gaertner, W., & Wüthrich, N. (2015). Evaluating competing theories via a common language of qualitative verdicts. *Synthese, 193,* 3293–3309. https://doi.org/10.1007/s11229-015-0929-4.

Goldman, A., & Blanchard, T. (2016). Social Epistemology. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), https://plato.stanford.edu/archives/win2016/entries/epistemology-social/.

Held, I. M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society, 86,* 1609–1614. https://doi.org/10.1175/BAMS-86-11-1609.

Hempel, C. G. (1965). Science and human values. In C. G. Hempel (Ed.), *Aspects of scientific explanation and other essays in the philosophy of science* (pp. 81–96). New York: The Free Press.

Hempel, C. G. (1988/2000). On the cognitive status and the rationale of scientific methodology. In R. Jeffrey (Ed.), *Selected philosophical essays* (pp. 199–228). Repr. Cambridge: Cambridge University Press.

Hirsch Hadorn, G. (2018). On rationales for cognitive values in the assessment of scientific representations. *Journal for General Philosophy of Science*. https://doi.org/10.1007/s10838-018-9403-6.

Intemann, K. (2015). Distinguishing between legitimate and illegitimate values in climate modeling. *European Journal for Philosophy of Science, 5,* 217–231. https://doi.org/10.1007/s13194-014-0105-6.

Jeffrey, R. C. (1956). Valuation and acceptance of scientific hypotheses. *Philosophy of Science, 23,* 237–246. https://doi.org/10.1086/287489.

Kitcher, P. (2001). *Science, truth, and democracy*. Oxford: Oxford University Press.

Knutti, R. (2008). Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society A, 366,* 4647–4664. https://doi.org/10.1098/rsta.2008.0169.

Kuhn, T. (1977). Objectivity, value judgment, and theory choice. In T. Kuhn (Ed.), *The essential tension* (pp. 320–339). Chicago: University of Chicago Press.

Lacey, H. (2004). Is there a significant distinction between cognitive and social values? In P. Machamer & G. Wolters (Eds.), *Science, values and objectivity* (pp. 24–51). Pittsburgh, PA: University of Pittsburgh Press.

Laudan, L. (1984). *Science and values. The aims of science and their role in scientific debate*. Berkeley, CA: University of California Press.

Laudan, L. (2004). The epistemic, the cognitive, and the social. In P. Machamer & G. Wolters (Eds.), *Science, values and objectivity* (pp. 14–23). Pittsburgh, PA: University of Pittsburgh Press.

Levi, I. (1960). Must the scientist make value judgments? *The Journal of Philosophy, 57,* 345–357. https://doi.org/10.2307/2023504.

Levi, I. (1986). *Hard choices: Decision making under unresolved conflicts*. Cambridge, UK: Cambridge University Press.

Longino, H. (2002). *The fate of knowledge*. Princeton, NJ: Princeton University Press.

Machamer, P., & Osbeck, L. (2004). The social in the epistemic. In P. Machamer & G. Wolters (Eds.), *Science, values and objectivity* (pp. 78–89). Pittsburgh, PA: University of Pittsburgh Press.

Machamer, P., & Wolters, G. (2004). Introduction. In P. Machamer & G. Wolters (Eds.), *Science, values and objectivity* (pp. 1–13). Pittsburgh, PA: University of Pittsburgh Press.

Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Frame, D. J., Held, H., Kriegler, E., Mach, K. J., Ebi, K. L., Matschoss P. R., Plattner, G.-K., Yohe, G. W., & Zwiers, F. W. (2010). Guidance note for lead authors of the IPCC Fitfth Assessment Report on consistent treatment of uncertainties. *Intergovernmental Panel on Climate Change (IPCC)*. Retrieved from http://www.ipcc.ch.

McMullin, E. (1983). Values in science. *Proceedings of the Biennial meeting of the philosophy of science association, 2,* 3–28. http://www.jstor.org/stable/192409.

McMullin, E. (1985). Galileian idealization. *Studies in the History of Philosophy of Science, 16,* 247–273. https://doi.org/10.1016/0039-3681(85)90003-2.

Okasha, S. (2011). Theory choice and social choice: Kuhn versus Arrow. *Mind, 120,* 83–115. https://doi.org/10.1093/mind/fzr010.

Parker, W. S. (2009). Confirmation and adequacy-for-purpose in climate modeling. *Proceedings of the Aristotelian Society, Supplementary, 83,* 233–249. https://doi.org/10.1111/j.1467-8349.2009.00180.x.

Parker, W. S. (2014). Values and uncertainties in climate prediction, revisited. *Studies in History and Philosophy of Science, 46,* 24–30. https://doi.org/10.1016/j.shpsa.2013.11.003.

Peschard, I., & van Fraassen, B. C. (2014). Making the abstract concrete: The role of norms and values in experimental modeling. *Studies in History and Philosophy of Science, 46,* 3–10. https://doi.org/10.1016/j.shpsa.2013.11.004.

Rabinowicz, W. (2008). Value relations. *Theoria, 74,* 18–49. https://doi.org/10.1111/j.1755-2567.2008.00008.x.

Rochefort-Maranda, G. (2016). Simplicity and model selection. *Euro Journal for Philosophy of Science, 6,* 261–279. https://doi.org/10.1007/s13194-016-0137-1.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science, 20,* 1–6.

Schmidt, G. A., & Sherwood, S. (2015). A practical philosophy of complex climate modeling. *European Journal for Philosophy of Science, 5,* 149–169. https://doi.org/10.1007/s13194-014-0102-9.

Spohn, W. (2004). On the objectivity of facts, beliefs, and values. In P. Machamer & G. Wolters (Eds.), *Science, Values and Objectivity* (pp. 172–189). Pittsburgh, PA: University of Pittsburgh.

Steel, D. (2010). Epistemic values and the argument from inductive risk. *Philosophy of Science, 77,* 14–34. https://doi.org/10.1086/650206.

Steele, K. (2012). The scientist qua policy advisor makes value judgments. *Philosophy of Science, 79,* 893–904. https://doi.org/10.1086/667842.

Strevens, M. (2006). The Bayesian approach to the philosophy of science. In D. M. Borchert (Ed.), *Encyclopedia of Philosophy* (2nd ed., pp. 495–502). New York, NY: Macmillan Reference.

van Fraassen, B. C. (1980). *The scientific image*. Oxford: Clarendon Press.

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy, 104,* 639–659. https://doi.org/10.5840/jphil20071041240.

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford: Oxford University Press.

Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Science, 40,* 92–101. https://doi.org/10.1016/j.shpsa.2008.12.005.

Winsberg, E. (2010). *Science in the age of computer simulation*. Chicago, IL: Chicago University Press.

Winsberg, E. (2015). Computer simulations in science. In E. Edward (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2015 Edition). Retrieved from http://plato.stanford.edu/entries/simulations-science/.

Chapter 41
Calibration, Validation, and Confirmation

Mathias Frisch

Abstract This chapter examines the role of parameter calibration in the confirmation and validation of complex computer simulation models. I examine the question to what extent calibration data can confirm or validate the calibrated model, focusing in particular on Bayesian approaches to confirmation. I distinguish several different Bayesian approaches to confirmation and argue that complex simulation models exhibit a predictivist effect: Complex computer simulation models constitute a case in which predictive success, as opposed to the mere accommodation of evidence, provides a more stringent test of the model. Data used in tuning do not validate or confirm a model to the same extent as data successfully predicted by the model do.

Keywords Predictivism · Bayesian epistemology · Problem of old evidence · Tuning · Climate models

41.1 Introduction

Many complex computer simulations involve semi-empirical parameterizations. Parameterizations represent complex processes through simplified, approximate equations involving parameters that often are only poorly constrained through a theoretical understanding of the phenomena modeled. Since parameter values are not determined by any underlying theory, the values need to be *calibrated* or *tuned*. Yet modelers often express reservations about the need for calibration in model development, suggesting that this need presents a problem for the degree of trust we can have in a model's predictions. In particular, it is common to maintain that data used in calibration cannot unproblematically also be used to evaluate a model's performance (Mauritsen et al. 2012; Intergovernmental Panel on Climate Change 2014, Box 9.1). Calibration data, that is, cannot also be used to validate a model.

M. Frisch (✉)
Institute for Philosophy, Leibniz Universität Hannover, Hanover, Germany
e-mail: mathias.frisch@philos.uni-hannover.de

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_41

My aim in this chapter is to examine these concerns and ask how the need for parameter calibration affects the validation of a simulation. The core epistemological issue presented by calibration—to what extent calibration data can also be used to evaluate a model's performance and to validate the model—is a special case of a problem that has a long history in the philosophy of science, the problem of predictivism: is the ability successfully to predict new evidence more highly confirmatory of a theory or model than the successful accommodation of existing evidence?

My discussion will concentrate on the calibration of climate models (where the practice is usually referred to as *tuning*), since the issue of parameter calibration has received a fair amount of attention in the climate modeling literature (Hourdin et al. 2016; Baumberger et al. 2017; Bellprat et al. 2012; Masson and Knutti 2012; Golaz et al. 2013; Kennedy and O'Hagan 2001; Gleckler et al. 2008) (see also Chap. 29 by Rood in this volume). But the issues I will discuss apply to the calibration of complex computer simulations more generally. And while I will examine calibration mainly through the lenses of the philosophical problem of predictivism, I will also briefly touch on several other conceptual issues concerning calibration.

In the next section I will provide a brief overview of parameter calibration in climate modeling and of some of the conceptual problems that arise for calibration in this context. Then I will provide a brief survey of the philosophical debate concerning predictivism, focusing in particular on Bayesian confirmation theory and the so-called *problem of old evidence*. Finally, I will discuss what various strategies for responding to the problem of old evidence in a Bayesian framework entail for the epistemological status of calibration and its role in model validation. My conclusion in this somewhat opinionated survey will be that there is at least one kind of argument in support of the concern frequently expressed by climate modelers that data used in tuning do not validate a model or do not confirm a model to the same extent as data successfully predicted by the model. Complex computer simulation models constitute a case in which predictive success, as opposed to the mere accommodation of evidence, can provide a more stringent test of the model.

41.2 Computer Simulations, and Calibration

41.2.1 Calibration, Verification, and Validation

We can broadly (and roughly) distinguish two types of computer simulation: equation-based simulation, and agent-based simulation (Parker 2013). Equation-based simulations involve dynamical equations, which often will be differential equations, reflecting our theoretical understanding of the processes modeled. Climate models, for example, involve equations from fluid dynamics that allow us to model the flow of mass and energy in the atmosphere. The dynamical equations of agent-based simulations, by contrast, represent—often very simple—rules of behavior for individual agents. My focus will be on equation-based simulations and in particular on climate modeling.

Two central concepts in the evaluation of computer simulations are the *verification* and the *validation* of a simulation. Oberkampf et al. (2004) distinguish the two concepts as follows:

> Verification: The process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model.
>
> Validation: The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model. (347)

That is, verification is concerned with the relation between a simulation model and the conceptual and mathematical model from which it is derived. "Verification deals with mathematics" (Roache quoted in Oberkampf et al. 2004) and not with the relation between a simulation and the world. Validation, by contrast, concerns the relation between a simulation and the world; or as Roache put it somewhat metaphorically "validation deals with physics." Validation, thus, concerns what philosophers call the empirical confirmation of a simulation model. (cf. Chap. 2 by Oberkampf and Chap. 3 by Murray-Smith in this volume.)

Some of the equations in complex climate models express basic and well-understood physical principles or are approximations to such principles. Among these principles are the principles of mass, energy, and angular momentum conservation or basic equations of thermodynamics or fluid dynamics. These equations together comprise a mathematical climate model. To obtain numerical solutions, the equations are discretized on a spatiotemporal grid, which for global climate models has a grid size of 25–300 km. The question of a climate simulation's verification concerns the question as to whether numerical solutions to the discretized system of equations accurately represents the systems defined by the original set of equations and whether the discretized equations approximate the original system of equations sufficiently closely.

One of the complexities of global climate models consists in the fact that they need to integrate processes taking place at different scales. Here *parameterizations* that constitute sub-models for processes not explicitly represented in the complex model play an important role. Parameterizations allow modelers to represent complex processes involving scales smaller than the models' grid size through simplified, approximate equations. Parameterizations are chosen when more fully resolving processes in the model may be numerically too costly, or when a process may be too ill understood physically to be represented more fully (Hourdin et al. 2016).

A distinction that is important in this context is that between a *family of models* and *model instances*. A family of models is characterized by a common set of equations, which may include parameterizations. A family of models consists of different model instances, each of which is characterized by a set of specific parameter values. For example, the linear equation $y = ax+b$ defines a family of models comprising different model instances. Each model instance is defined by a tuple $(a_n; b_n)$ of specific values for the parameters a and b. Thus, one or several parameterized equations help to define a family of models, whereas a particular set of parameter values serves to pick out a specific model instance.

Parameter calibration can take the form of conducting a so-called "perturbed physics experiment" (Parker 2010), in which a model with the same model-structure is run several times with different parameter values to determine the values with the best observational fit. Due to the computational costs of running complex models, however, modelers generally cannot explore the full space of parameter values. Instead, ranges of plausible parameter values that can be computationally explored need to be determined by expert judgment (Bellprat et al. 2012). That is, models are not calibrated to the parameter values that we know to provide the best fit among all possible values, but to the values that provide the best fit among the small set of values modelers find plausible or worth exploring and do in fact test.

Moreover, there may be more than one parameter set that is equally compatible with the observational evidence used to calibrate the data (Knutti et al. 2008) and values of different parameters may depend on one another. Thus, the data used to calibrate a model may underdetermine possible parameter values and observational constraints may, in this sense, be too weak. At the same time observational constraints may also be too strong in that it may not be possible to calibrate a given model equally well to different data sets (*Ibid.*). Instead of rejecting a model completely in such cases, modelers often argue that the model is adequate for some purposes but not for others. This raises the question how much we can trust a model that is calibrated with respect to one type of data set (such as twentieth century changes in global mean surface temperature) to be predictively adequate with respect to some other data set, such as temperature changes or precipitation changes in the second half of the twenty-first temperature.

Examples of calibrated parameters in climate models include parameters related to the behavior of clouds, which involve sub-grid processes that also remain relatively poorly understood physically. Other examples are parameters related to Earth's albedo (that is, the reflectivity of the oceans or land surfaces). Cloud parameters are, for instance, tuned against the top of the atmosphere (TOA) energy balance. As a consequence of discretization many models do not satisfy the principle of energy conservation: in control runs simulating preindustrial climate in which atmospheric constituents are fixed at preindustrial levels models leak energy and have a positive radiation imbalance. This problem can be addressed by calibrating a cloud homogeneity factor to adjust the TOA net shortwave flux, and hence the TOA energy flux imbalance (Mauritsen et al. 2012).

Cloud parameters are also calibrated to provide a better fit with changes in the global mean surface temperature (GMST). For example, the threshold radius at which cloud droplets fall as rain—the so-called auto-conversion threshold radius—can be calibrated so as to produce a better fit with observed changes in the GMST (Golaz et al. 2013, 2010). The threshold radius determines the size of cloud water droplets at which droplets turn into rain. Increasing the radius delays rainfall, increases cloudiness, and thereby affects the radiation balance, which in turn affects surface temperatures.

In this example, a model has been explicitly and intentionally calibrated to better match twentieth-century warming. Some authors worry that climate models may also be tuned *inadvertently* to match twentieth-century warming. If we broaden the notion of calibration to include not only the process of adjusting the value of a free parameter in a model but also decisions on whether to include parameterizations of certain processes in the model and on how to model these processes, then it is not implausible that tuning to twentieth-century warming occurs unintentionally during the construction of climate models (Hourdin et al. 2016). Whether this raises epistemological problems is a question to which I will return below.

Both examples I mentioned can be used to illustrate a potentially problematic feature of parameter calibration: the fact that calibration sometimes amounts to a process of trying to find a successful balance among compensating errors. Ocean and surface sub-models in a coupled climate model often use different grid-sizes. The energy leakage in global climate models seems to be largely due to the mismatching grids between the sub-models (Mauritsen et al. 2012, 8). Calibrating cloud parameters to adjust the TOA does not correct these structural problems of the model, but rather compensates for these problems by making adjustments to the model elsewhere. As Mauritsen et al. explain: "adjusting cloud parameters involves a process of error compensation, as it is well appreciated that climate models poorly represent clouds and convective processes. Tuning aims at balancing the Earth's energy budget by adjusting a deficient representation of clouds, without necessarily aiming at improving the latter." (2012, 2)

That calibration can involve compensating errors is also illustrated by the second example. Here, the droplet radius that achieves the best fit with observed changes in the GMST is one that is smaller than appears physically reasonable. Thus, the droplet radius is tuned to a value that is in tension with our best physical understanding of the parameterized processes. This apparently incorrect representation of cloud droplets is justified by the fact that it appears to compensate for unknown other, perhaps structural problems of the model.

What role does parameter calibration play in verification and validation of simulations? In fact, it can play a dual role: on the one hand, parameter values are picked with an eye on a model's *verification* in order to compensate for errors introduced in implementing the underlying dynamical equations in a discretized computer simulation. An example of this is the calibration of parameters to compensate for the fact that a model's discretization does not satisfy energy-momentum conservation. On the other hand, parameter values are chosen with an eye on a model's empirical fit, for example, when a simulation is calibrated with respect to changes in twentieth-century global temperatures. Now, intuitively calibrating a model to increase its empirical fit—or, more carefully, choosing a particular model instance over others because of its superior empirical fit—concerns the model's *validation*. Yet there is some disagreement in the literature on whether parameter calibration can in fact be part of the process of validating a model or not.

Oberkampf and Barone (2006) explicitly distinguish between the task of evaluating models with the help of a validation metric and the process of parameter calibration: "Our emphasis in validation metrics is in blind assessment of the predictive capability of a computational model (how good is the model?), as opposed to optimizing the agreement between a given model and experimental measurements." (10) In fact, Oberkampf and Barone explicitly exclude parameter calibration, which amounts to adjusting parameter values in the light of known evidence, from the process of properly validating a model. Thus, when discussing the interaction between computer modelers and experimentalists, who supply empirical data, they say: “What should not be provided to the computationalists [by the experimentalists] in a rigorous validation activity is the measured SRQ [the system response quantity measured]. Stated differently, it is our view that a blind computational prediction be compared with experimental results so that a true measure of predictive capability can be assessed in the validation metric.” (9) Proper validation requires a “true measure of the predictive capability” of a simulation, which is only possible through a “blind” prediction without prior knowledge of what the correct empirical values of the quantity of interest are.

Oberkampf et al. (2004) explain in more detail why successful prediction as part of a model’s validation is important. Oberkampf, Trucano, and Hirsch maintain that our confidence in the accuracy of predictions may be low for predictions that are far, in some sense, from the validation database. While we infer the accuracy of a model’s predictions from its performance with respect to the validation database, the reliability of this inference can be quite weak in the case of complex computer simulations. For unlike in the case of “traditional scientific theories” computational simulation “also relies on many additional mathematical issues, e.g., discretization algorithms and grid quality, and practical implementation issues, e.g., computer hardware, operating-system software, source-code reliability, and analyst skill, that are not present in classical scientific theories.” (351) All these additional features of computer simulation serve to reduce the strength of the inference from predictive accuracy in one domain—the validation base—to that in another: the domain of the model’s intended application.

Whereas in the case of engineering there is either a complete or at least a partial overlap between the validation and application domain, they maintain, in the case of many complex computer simulations there is no overlap between the two domains. In the latter case, we face an inference problem: how can we justify our belief in the model’s predictions in the application domain from its success in the validation domain? And while, in contrast with (Oberkampf and Barone 2006), Oberkampf, Trucano, and Hirsch allow calibration data to be part of the validation database, they stress that the inference from successful validation to successful prediction is stronger for data in the validation database that have not been used in calibration: “The need to perform this extrapolation reinforces our need for models to be judged on the basis of achieving the right answers for the right reasons in the validation regime. [… It] is not convincing that model calibration provides a starting point for the inference process” from validation to application domain. (Oberkampf et al. 2004, 352) Below I will propose a Bayesian defense of Oberkampf, Trucano ond Hirsch’s claim.

How are the different domains in question delineated and what determines the closeness among different domains? What the relevant criteria of similarity and identity for domains are depends on the details of modeling framework at issue. The domains of models constructed purely with the help of physical laws or “traditional scientific theories” will be much larger, due to the laws’ known inductive stability, than the domains of complex computer simulation models. The identity conditions for a domain of application depend on the range of invariance of a model across changes of initial or boundary conditions. In the case of models built with the help of physical principles that we know to have a large range of invariance, changes of initial or boundary conditions within the invariance range will not take us to a new domain of application. By contrast, in the case of complex simulation models, where we do not know to what extent the model's predictions are sensitively dependent on specific initial or boundary conditions, domains have be distinguished in a much more fine-grained manner.

41.2.2 Adequacy for Purpose

The features of calibration I surveyed suggest that the aim of calibration cannot be to arrive at a "true" model in which parameters have their true values. Rather, calibrations have to satisfy a range of different, sometimes conflicting constraints on the model and on parameter values. Because these different constraints may be weighted in different ways, and since different modeling aims may suggest different weightings, tuning arguably is a skill requiring expert judgment. As Hourdin et al. put it, tuning is an "art" as much as it is a scientific or "engineering calibration exercise" (Hourdin et al. 2016, 598).

Some argue that type of expert judgment required in deciding on plausible parameterizations for complex simulation models and settling on a preferred set of parameter values introduces a "subjective element" into model calibration (see Hourdin et al. 2016), even when, as in the case of climate models, the models are ultimately based on well-understood physical principles. Yet it is important not to overemphasize the subjectivity of calibration and to characterize carefully what the subjective or pragmatic element introduced by model calibration consists in. The need for expert judgment in tuning does in no way undermine the basic predictions made with the help of climate models and does not license a "climate skeptical" conclusion.

At least some of the features of parameterized climate models which are discussed in the literature are features of scientific models much more generally. Still speaking mainly of climate models, Wendy Parker (Parker 2009) has argued that the aim of climate modeling cannot be to arrive at a true model of the phenomena but only at a model that is adequate to some specific purpose. Some models, such as simple energy balance models of the Earth that treat the Earth as a simple black body, are obviously highly idealized, abstract away from many messy yet important features of Earth's climate, and represent only a few basic physical processes. Yet even highly complex models, such as general circulation models and Earth System Models contain idealizations and abstraction and arguably also structural errors. These errors or misrepresentations are partly due to a particular discretization chosen, partly due to the fact that certain processes are too complex to be included even in ESMs in complete detail, and partly due to the fact that some of the processes modeled are still not fully understood. Thus, the aim in climate modeling is not (and arguably cannot be) to arrive at a true model of the climate system—a representation that is correct in all details but is to construct models that capture different aspects of the climate system in ways that are adequate for different representational purposes.

Different models will be best suited for different purposes and a particular model may be adequate for some purpose but not for others. Climate scientists speak of the "skill" of a model and point out that there is no unique overall metric of skill and that different models can be more or less skillful at different tasks (Gleckler et al. 2008; Knutti et al. 2008). For example, a particular model might be skillful at adequately reproducing twentieth century temperature trends in the Northern Hemisphere—that is, predicted trends might be within what for a given purpose are acceptable margins of error for measured (and reconstructed) average temperature trends—while the model might be less skillful at representing precipitation patterns in the Southern Hemisphere. Or, to take another example, simple energy balance models may be the models best suited for providing a qualitative understanding of the greenhouse effect and of temperature trends, since these models make the physical mechanisms at work perspicuous in ways in which highly complex general circulation models (GCMs) or Earth system models (ESMs) do not. The latter models, by contrast, are better suited for quantitative projections. When a climate model is evaluated, we do not confirm the correctness *tout court* of the model but rather what Parker has called its "adequacy-for-purpose" (Parker 2009).

Purposes to which models may be put can be distinguished along different dimensions (Frisch 2015; Baumberger et al. 2017) (see also Ch. 39 by Hirsch Hadorn and Baumberger in this volume). Differences in purpose do not only concern predictive skills in different domains, but also whether a model can offer precise quantitative predictions or merely predicts qualitative trends. Alternatively we may be interested in an explanation of a certain relationship between climate variables or may be looking to understand the mechanism that is responsible for a certain phenomenon rather than be looking for a numerical prediction. Different models at different levels in the modeling hierarchy ranging from simple energy balance models to GCMs and ESMs may be better suited for different purposes (Held 2005).

While Parker discusses her claim that evaluating models always has to consist in evaluating a model's adequacy for some particular purpose only in the context of climate models, this point arguably applies to models in the sciences much more generally and is a consequence of Nancy Cartwright's influential argument (Cartwright 1983) for the view that all models in the sciences idealize, contain abstractions, and hence partly misrepresent their intended targets.[1] Parameterizations are often characterized as offering coarse-grained representations of processes at length scales below the grid scale. But arguably every model of physical processes, except, perhaps processes at the most fundamental physical level (if there is such a level), is coarse-grained with respect to levels more fundamental than the level at which the process is modeled.

Thus, the putatively "subjective element" in calibration is characteristic of all modeling and better described as a *context- or purpose-relativity* of scientific models and as expressing the fact that *constructing* the right model in a given context

[1] See also (Box 1979, p. 202): "All models are wrong but some are useful."

requires skills that may be difficult to characterize explicitly. What distinguishes parameterizations from many other aspects of climate models is that the relationships they posit are not based on well-established and robust physical principles. This is a point that has important consequence for the validation of climate models, as we will see below. In what follows, I will simply speak of the confirmation or validation of simulation models, but always intend this to be purpose relative. Which quantities are relevant to a model's confirmation or validation and what degree of fit between a simulation and empirical data is required will be understood as a context- and purpose-dependent question.

41.3 Predictivism

Does the fact that a model can accommodate observational evidence through calibration or tuning confirm the specific model instance resulting from the calibration to the same extent as it confirms a model instance that can adequately predict data not used in calibration? Or does the fact that a model can be calibrated to a given data set offer less confirmation of the calibrated model instance than if a model instance had successfully predicted the data? Many climate scientists suggest that data used in tuning cannot confirm or validate a model instance to the same extent that successfully predicted data do. Thus, the Fifth IPCC Assessment Report writes in its discussion of "Climate Model Development and Tuning": "Model tuning directly influences the evaluation of climate models, as the quantities that are tuned cannot be used in model evaluation. Quantities closely related to those tuned will provide only weak tests of model performance." (IPCC 2014 Box 9.1)

41.3.1 The Paradox of Predictivism

The underlying issue—the problem of the epistemic status of a theory's accommodation of evidence compared to that of successful prediction—has a long history in the philosophy of science (Glymour 1980; Glymour 2010; Maher 1988; Howson and Franklin 1991; Eells and Fitelson 2002; Eric C Barnes 2008; Howson 1991). There are both philosophers defending a privileged status for predictive successes and those arguing that there is no epistemic difference between accommodation and successful prediction. And, indeed, there exist seemingly powerful intuitions supporting *both* the view that predictive success is more highly confirmatory of a theory than its ability successfully to accommodate empirical data *and* the view that predictive and accommodating successes are equally confirmatory. The philosopher Imre Lakatos famously distinguished progressive and degenerating theory changes; the former are characterized by their predictive successes, whereas the latter merely succeed in accommodating data after they have been collected (Worrall 1980). For Lakatos, predictive successes are the hallmark of a successful scientific research program.

Accommodation can seem *ad hoc* if a theory is modified only for the purposes to fit a particular type of evidence with which it otherwise would have been in conflict. An oft-cited example is the defense of phlogiston theory. Phlogiston was believed to be a substance that is given off during combustion. When it was observed that some substances gained weight when they were burned, defenders of the theory tried to save phlogiston theory by postulating that phlogiston had negative weight. Despite its success in accounting for the evidence, this accommodation strikes many as an illegitimate ad hoc maneuver to rescue the theory.

Yet, as many have argued in response to Lakatos (see, e.g., Worrall 2014 and references therein) even if successful prediction makes a difference, the *temporal* order between the development of a hypothesis and the discovery of evidence does not appear to be epistemically relevant. If there is a predictivist effect at all, then this can at most concern the question whether evidence has been *used* in the construction of a theory or not: what matters, it seems, is only whether evidence is accommodated in the construction of a theory or is what is known as *use-novel*. A famous example that is often invoked to support the intuitions that temporal order cannot make a difference to confirmation is the confirmation of the theory of General Relativity (GR) by observations of the perihelion of Mercury. Even though the precession of the perihelion of Mercury had been observed well in advance of Einstein's development of GR, the theory is confirmed by these observations, since Einstein apparently did not use the relevant data in the construction of his theory. Similar intuitions may be behind the demand in (Oberkampf and Barone 2006) that the computer modeler does not know what the experimental values of the quantities are she is trying to simulate.

While there are strong intuitions in support of predictivism, at least as far as use-novel evidence is concerned, there are perhaps equally powerful intuitions suggesting that a theory's or model's historical development should not have any bearing on how well it is confirmed vis-à-vis the existing evidence. What should matter, one might think, is only the relation between a model and the relevant data and not how the model was constructed. Philosophers of science who argue against the epistemic importance of novelty point to examples like Deborah Mayo's example of average scores in the scholastic aptitude test (SAT): If we calculate the average SAT scores for a groups of students, then our "theory t" may be that the average score $S_{average}$ has a certain value: t says that $S_{average} = s$. But it seems that t is as well confirmed by the individual scores when we derive t from these scores as when we assume that the theory has been concocted by some other means (see Barnes 2008). This suggests that at the very least predictivism cannot be true in general: a hypothesis is not always better confirmed by evidence successfully predicted than by evidence successfully accommodated.

Eric Barnes, in his authoritative examination and careful defense of predictivism (Barnes 2008), calls this conflict of intuitions *the paradox of predictivism*. But are there situations in which some version of predictivism is true? As Eric Barnes maintains, predictivism is best thought of as a comparative thesis—the thesis that a theory t is more highly confirmed by the successful prediction of evidence e than by successfully accommodating e. Are there conditions under which this thesis is true and, if yes, does the calibration of complex computer models satisfy these conditions?

41.3.2 Bayesian Confirmation Theory V and the Problem of Old Evidence

Many discussions of predictivism take place in the context of formal decision frameworks, in particular Bayesian decision theory. But while formal accounts of confirmation may go some way toward clarifying the conflict of intuitions, they bring with them their own set of difficulties.

According to the Bayesian framework, a hypothesis *h* is confirmed by evidence *e* just in case its posterior probability *p'*(*h*) exceeds its prior probability *p*(*h*). The posterior probability is defined by conditioning *h* on the evidence *e*: *p'*(*h*) = *p*(*h*|*e*). Here "*p*(*h*|*e*)" is the conditional probability of *h* given *e*. Hence *h* is confirmed just in case *p*(*h*|*e*) > *p*(*h*). All probabilities here and in what follows have to be understood as relative to a set of accepted background beliefs *b*. That is, more explicitly, *h* is confirmed by *e* just in case *p*(*h*|*e.b*) > *p*(*h*|*b*). Now, if the evidence *e* is already known then *p*(*e*|*b*) = 1, since *e* is part of our accepted background beliefs. It then follows from Bayes's Theorem that the posterior probability of *h* is equal to its prior probability: *p*(*h*|*e.b*) = *p*(*e*|*h.b*)x *p*(*h*|*b*)/ *p*(*e*|*b*) = *p*(*h*|*b*).

This is the Bayesian *problem of old evidence*, which was first discussed by Clark Glymour (Glymour 1980). According to the most straightforward application of the Bayesian framework, known evidence can never confirm a hypothesis. In particular, for the naïve Bayesian it makes no difference whether evidence is used in the construction of a hypothesis or not. If evidence is known and its probability is equal to one, conditionalizing *h* on the evidence cannot affect the probability of *h*. According to a straightforward application of Bayesian reasoning, data used in the successful calibration of a climate model do not confirm the model—but neither do data on the precession of the perihelion of Mercury confirm Einstein's theory. Thus, before we can hope for any insight from Bayesian reasoning into the issue of model tuning, we need to confront the problem as to how Bayesians may allow for at least some types of old evidence to have positive evidential impact.

One might try to avoid the problem of old evidence by insisting that we should never set probabilities concerning empirical facts strictly equal to one: only logical or mathematical truths ought to be assigned probability one. Thus, even for known evidence *p*(*e*|*b*) < 1. But while this move might solve what has been called *the qualitative problem of old evidence* (since it allows for old evidence to be confirmatory) it does not solve *the quantitative problem*—the problem to what degree old evidence can be evidence for a hypothesis (Barnes 1999). If we assume (as presumably, we should) that *p*(*e*|*b*) is very close to one for known evidence, then no hypothesis can ever be more than incrementally confirmed by known evidence.[2]

[2] This can be seen as follows. By Bayes's Theorem *p*(*h*|*e.b*) = *p*(*e*|*h.b*)x *p*(*h*|*b*)/*p*(*e*|*b*). If *p*(*e*|*b*) = 1-ε, for some small number ε, then *p*(*h*|*e.b*)/*p*(*h*|*b*) ≈ *p*(*e*|*h.b*) x (1 + ε). That is, the posterior probability *p'*(*h*|*b*) = *p*(*h*|*e.b*) cannot be appreciably larger than *p*(*h*|*b*). A version of the problem also arises if we replace the Principle of Conditionalization with Jeffrey Conditionalization, which presupposes that observations result in non-inferential changes in the probability of an evidential statement *e*.

There are several different proposals on how best to solve the problem of old evidence, but there seems to be near consensus that the problem of old evidence does indeed present a difficult challenge for the Bayesian and that, applied to known evidence, a straightforward application of the Bayesian machinery often gives the wrong answer. While a straightforward application of Bayesian reasoning suggests that evidence used in calibrating a model does not confirm the model—thus providing support for what sometimes is claimed in the climate modeling literature—we should not simply accept this conclusion, since the equivalent conclusion in the case of Einstein's theory and the perihelion of Mercury is obviously problematic.[3]

Philosophers have explored several types of response to this problem. One reply is to give up the assumption of omniscience for a Bayesian agent and argue that what Einstein learned when he derived the perihelion data from the theory is that GR implies the data. GR was confirmed when Einstein learned that a certain relation r holds between the theory and the evidence (Garber 1983; Sprenger 2015). This relation might be deductive entailment or may be some other, perhaps weaker, explanatory relationship. That is, a theory or hypothesis is confirmed when we learn that a certain explanatory relationship r holds between the theory and the previously known evidence. Then, according to one way of spelling this out, r confirms h relative to old evidence e when $p(h|e.r) > p(h|e)$.

A second type of response is to evaluate confirmation in terms of a counterfactual probability function $p_{\{b\}\setminus e}$ that is obtained by somehow subtracting the old evidence e from the set of background beliefs. The challenge for this strategy is to specify a procedure for subtracting e in a way that is well enough defined to result in a reasonably definite answer to how well h is confirmed by e. (Howson 1991; Barnes 1999)

One promising idea is to take both strategies on board and argue that the two strategies provide answers to two different problems. (Eells and Fitelson 2000; Sprenger 2015). When we learn of an explanatory or inferential relationship between a hypothesis or model and old evidence, then this *confirms* the model in the sense that it

As in the traditional formulation, the problem is that the probability of evidential statements does not change for old evidence.

[3] Steele and Werndl (2016) are among the very few dissenters from this consensus and suggest that the Bayesian formalism can be applied to the case of climate-model calibration directly and without any modification. Yet curiously they argue that a direct application of the Bayesian formalism implies that successful calibration against known data *can* confirm a model. The argument they give, however, is mistaken. Their discussion focuses on the case of comparative confirmation. Whether one hypothesis h_1 is confirmed with respect to another hypothesis h_2 is given by the following ratio (where conditionalization on background beliefs is left implicit): $p(h_1|e)/p(h_2|e) = p(e|h_1)/p(e|h_2)$ x $p(h_1)/p(h_2)$. Steele and Werndl maintain that this ratio can change as a consequence of calibrating our models against known evidence and hence that one model can be incrementally confirmed or disconfirmed with respect to another model: "For the Bayesian, calibration is not really distinct from confirmation." Yet they also (as is standard) assign known evidence probability one: "When new data are learnt, the relevant evidence proposition is effectively assigned a probability of one." (*Ibid.*) But then in the case of calibration against data e that have been previously known the likelihoods $p(e|h_1)$ and $p(e|h_2)$ are both equal to one and hence $p(h_1|e)/p(h_2|e) = p(h_1)/p(h_2)$. Thus, a direct application of Bayesian reasoning yields exactly the opposite conclusion from the one Steele and Werndl want us to reach.

increases our credence in the hypothesis or increases our credence in the adequacy of the model for the purpose at issue. Thus, this strategy offers an answer to what Sprenger calls the *dynamic* problem of old evidence: how can old evidence affect our degree of belief in a hypothesis *h* when the explanatory relationship between *e* and *h* has been discovered? The strategy answers this problem by giving up the idealized assumption that Bayesian agents are logically omniscient.

By contrast, the second strategy is concerned with the *evidential* relation between a piece of evidence and a hypothesis or model. Here we engage in counterfactual reasoning and ask to what extent a certain data set (in light of a specific set of background assumptions) provides evidence for a model or hypothesis. At issue here is not our degree of belief in the hypothesis, but an evidentiary relationship between hypothesis and evidence. The second strategy, thus, may be proposed as an answer to the *static* problem of old evidence: the problem as to how old evidence *e* can be evidentially relevant to a hypothesis *h* even after it has been discovered that *h* accounts for *e* (Eells and Fitelson 2000). An advantage of taking the counterfactual probability functions employed in the second strategy not as expressing *confirmation* but as *evidentiary* relations is that the problem that the counterfactual probabilities are ill-defined does not arise. If the probabilities express evidentiary relations, we are not concerned with what the full set of background beliefs of an agent is and how one might counterfactually subtract a commitment to some piece of evidence from that set. The probabilities do not express the degrees of belief of an agent with a rich set of background beliefs. Rather we are interested in relations of evidential support that can be evaluated in terms of a reasonably precisely delineated set of counterfactual background assumptions. As Sprenger notes, this type of counterfactual judgment is a standard component of scientific reasoning.

There are other Bayesian solutions or responses to the problem of old evidence, but there is no room to discuss these here. However, the very existence of the multitude of different and often conflicting Bayesian answers to the problem of old evidence suggests that appeals to Bayes's theorem, as abstract template for arguments about evidence and confirmation, cannot alone settle the problem of the epistemic status of parameter calibration. Indeed, as one commentator has pointed out, all possible positions with respect to the relation between Bayesianism and predictivism have their defenders: "philosophers have defended all four possible positions: Bayesian analysis is (i) valid because it favors novel prediction, (ii) valid because it does not favor novel predictions, (iii) invalid because it favors novel predictions, and (iv) invalid because it does not favor novel predictions" (Brush 1994; see also Douglas and Magnus 2013). No formal Bayesian argument on its own can settle the question whether old evidence in general and tuning-evidence, in particular, can be used in theory-evaluation. At best we can hope that there exists a Bayesian formalization that further illuminates philosophical commitments supported by another route. *The* Bayesian logic of confirmation simply does not exist.

41.3.3 Validation and Confirmation

Just as our intuitive concept of empirical confirmation is ambiguous between the static concept of evidential support and the dynamic concept of changes in our degree of belief in a hypothesis, the concept of validation similarly appears to conflate these two dimensions. Thus, validation is often described as part of a dynamical process of modeling and simulation, consisting of different temporal phases, which can be represented in flow charts. This suggests a dynamic understanding of validation: validating a simulation increases our confidence in the simulation, which is then used to predict novel evidence. Given such a dynamic understanding, a special role for novel predictions and the demand that experimentalists not share their data with computer modelers may seem plausible. Yet modelers also develop validation metrics (see, e.g., Oberkampf and Barone 2006) capturing the agreement between computational results and experiment and experimental uncertainties. Such metrics suggest a static understanding of validation and a view of validation as being concerned with evidentiary relations between a simulation and experimental evidence.

41.4 The Problem of Old Evidence and Model Calibration

41.4.1 The Static Problem of Old Evidence

How do the last two Bayesian strategies treat parameter calibration? As far as the static problem of old evidence and relations of evidential support are concerned, it would appear that the fact that a climate model is calibrated against a particular data set does not affect the data set's evidential status *vis a vis* the model. If the strength of evidential support between model and evidence is independent of when the evidence was discovered historically, then it should also be independent of whether the evidence was used in the construction of the model. What we are assessing, in this case, is what amounts to a counterfactual probability: the probability of the evidence given the model and some reasonably clearly delineated set of background assumptions concerning the phenomenon of interest. These assumptions need not—and in general will not—accurately mirror our actual background beliefs.

For example, in the case of climate models that are tuned against twentieth-century warming we are interested in how probable a particular model (with calibrated parameter. values) is, given the observed warming and relevant background assumptions, such as the absence of any feedback factors not included in the model. By Bayes's theorem, this probability is related to the model's likelihood, which is the probability of the observed warming conditional on the model's adequacy together with the set of background assumptions. The probabilities in question do not directly represent credences or our degree of belief in the model's adequacy but rather evidential relations between model and data in what amounts to a counterfactual setting—a setting in which the background assumptions not only do not already contain the evidence used to tune the model but also in other ways are not intended to capture our full set of beliefs about the world. Background beliefs represent assumptions made in a certain modeling context. As far then as the static problem of old evidence is concerned, data used to tune the model provide evidence for the model's adequacy in the same way in which fit to new data would. Examples of this are comparisons of climate models with different forcings to determine which combination provides the best fit with the observed twentieth century warming. Models that include both natural and anthropogenic forcings fit the evidence much better than models with natural forcings alone (see, e.g., IPCC 2007 9.4.1.2). That is, when we use observational data on twentieth century temperature changes to validate climate simulation models, we find that models that include both natural and anthropogenic forcings score much higher on any reasonable validation metric than simulations that include only natural forcings.

41.4.2 The Dynamic Problem of Old Evidence

What the solution to the dynamic problem of old evidence says about calibration is somewhat more ambiguous. On the one hand, the strategy seems to allow for an obvious distinction between different uses of old evidence: cases in which the evidence is used in the construction of a theory and cases in which it is not used. In the latter case, our credence in the adequacy of a model increases, when we discover that the model does in fact explain the data. Thus, GR was confirmed when Einstein succeeded in deriving the perihelion data from the theory. By contrast, if evidence e is used in the construction of a model, then the fact that the model accounts for e appears to be built into the model and the model is, thus, not confirmed by deriving e from it. As Sprenger argues, in this case it is certain, conditional on e, that h explains e. That is, $p(r|e) = 1$ and hence $p(h|e.r) = p(h|e)$. Thus, r fails to confirm h.

On the other hand, in the case of complex simulation models modelers arguably do not know in advance of running a model which parameterization provides the best fit with the evidence. Due to the models' complexity, modelers often do not know how changes in parameter values will affect the models' performance. Thus, analogous to the case of Einstein's derivation do we only find out that a particular model (with a specific parameterization) is adequate after the model is run. If this is right, then according to the Bayesian strategy we are currently considering it makes a difference to confirmation *how* data are used in the construction of a model. If a model is constructed using existing evidence so that it could not fail but to account for the evidence, then the evidence does indeed not confirm the model. Yet if a model is sufficiently complex, so that we only learn whether a certain choice of parameterization is successful after we have run the model (or have calculated the model's empirical consequences), then $p(r|e) < 1$ before the model is run and the parameterized model is confirmed by a successful calibration.

Let us briefly take stock. I distinguished three Bayesian approaches to confirmation with old evidence and discussed how these apply to the calibration of climate models. According to a direct application of "the Bayesian method", data used to calibrate a model do not confirm the model, since the probability of the evidence and the likelihood are equal to one. This is commonly taken to be a problem for any straightforward application of Bayesian reasoning to confirmation. If we apply a counterfactual subtraction method and examine relations of evidential support between a theory or model and evidence, then, in contrast to "straightforward Bayesianism", old data, including data used in model calibration, can provide evidential support for a model. But this relation of support is best thought of not as a *dynamic* relation of confirming the model in the sense of raising our degree of belief in the model's adequacy but rather as a *static* relation of evidential support. Validation metrics constitute examples of how to capture such relations of evidential support. Finally, we can give up the idealized assumption of omniscience, which is generally made in the Bayesian framework, and take confirmation by old evidence to consist in the learning of certain inferential relationships between model and data. On this last account, calibration data arguably confirm the calibrated model, at least in the case of complex climate models, since we only learn that a calibration is successful when the model is run.

41.4.3 An Argument for Predictivism

There is one further aspect concerning model confirmation I have not discussed so far and that suggests that data not used in model calibration can play a different role in model confirmation than data used in calibration.[4] Consider John Worrall's version of Patrick Maher's well-known coin toss example (Worrall 2014). Two investigators I_1 and I_2 both make a prediction about the 100th toss of a deterministically operating coin-tossing machine, which produces a sequence of 100 seemingly random outcomes. I_1, the accommodator, announces a hypothesis about the 100th outcome and the 99 prior outcomes after having observed the first 99 outcomes. I_2, the predictor, formulates a hypothesis about the results of the first 100 tosses before having observed any outcomes, correctly predicting the outcome of the first 99 tosses. If we ourselves are ignorant about the mechanism of the coin toss machine, then we should trust the prediction for 100th toss made by I_2 more strongly than the prediction of I_1. For, as Worrall argues, the fact that I_2 was able correctly to predict the first 99 tosses is strong evidence that she has insight into the mechanism of the coin toss machine, while we have not such evidence in the case of I_1.

Predictive success, thus, can play a role in situations involving multiple epistemic agents in that it can provide reasons for one agent to take another agent to possess a certain expertise. Predictive success is a means for an agent to certify her epistemic credentials to another agent: it allows another agent to infer that the successful predictor possesses a certain kind of knowledge that also renders her further predictions trustworthy. This does not imply that the predictor is necessarily more credible than the accommodator. If I_1 can convincingly explain to us why she understands the

[4] The discussion below follows closely my presentation in (Frisch 2015).

tossing mechanism, then the predictive effect in favor of I_2 arguably disappears. Predictive success, according to the present argument, is not epistemically significant in itself but is significant only as a symptom of the presence of some other epistemically relevant feature–in this case the investigator's knowledge of the machine's mechanism.

Worrall's thought experiment provides an argument for the claim that predictive success can be epistemically significant when multiple agents are involved. Yet predictive success can also be significant in the case of a single agent, since a single agent can be in a situation *vis a vis* complex models or theories that is analogous to that of an agent *vis a vis* another agent, whose expertise the former needs to judge. Just as we can use an agent's earlier predictive successes as a sign that she has certain qualities that make her novel predictions credible, we can use the predictive successes of a complex simulation model as a sign that the model possesses certain features that make it predictively successful. As in the case of multiple agents, the model's prior predictive successes are a symptom that the model possesses some property that accounts for its success. And analogous to the case of multiple agents, prediction is epistemically significant precisely in cases where the good-making properties of the model are not otherwise accessible to us.

Complex simulations arguably possess this feature. Complex models, such as complex climate models are to some degree *epistemically opaque* in that modelers cannot fully track analytically how the different model components (such as physical principles, initial and boundary conditions, and parameterizations) interact and contribute to a model's success along various performance metrics (Frisch 2015; Baumberger et al. 2017). While we have great confidence in the fundamental physical principles that underlie the construction of climate models, it is often not known how the more principled components of the model interact with parameterized model components to result in the model's outputs for different climate variables. Moreover it is often difficult to know in advance of running the simulation whether a model that is skillful at one specific task also is skillful at another task. In particular, one of the aims of GCMs or ESMs is to provide adequate predictions of how the climate system will evolve in the future. Our only evidence, of course, consists in the models' skill with respect to past and present datasets. Thus, climate scientists are faced with the task of evaluating to what extent a model's success with respect to some performance metric concerning past or present climate should increase our confidence in the model's skill with respect to future climate states.

The foregoing discussion, thus, supports the view in (Oberkampf et al. 2004) discussed above: While calibration data can be part of a model's validation dataset, successful calibration does not provide us with strong reasons for trusting a simulation's predictions in domains far from the calibration and validation domain. Validation that involves successful "blind" predictions (Oberkampf and Barone 2006), by contrast, can increase our confidence in the inference from successful validation to a simulation's successful application in another domain. The reason is that in the latter case we can have greater confidence that the model's successful predictions are not the result of calibrating the simulation to features characteristic of the validation context but rather are evidence of the fact that the model successfully represents the relevant underlying physical mechanisms.

41.4.4 A Novel Bayesian Argument for Predictivism

Predictive success, I have argued, can play an epistemically significant role. As I want to show now, this view, too,can be supported by a formal Bayesian argument.

We want to compare models that are tuned to account for certain data with models that successfully predict the data without having been tuned to these data. Let e be the statement that a model M is empirically adequate with respect to evidence E. Let f be the statement that M has a good-making feature F that allows it to be robustly adequate across different contexts. For example, F might be that M latches on to underlying physical principles or successfully represents the underlying physical mechanism. And let t be the statement that M is calibrated against the evidence E. Now let us make the following assumptions:

(1) $p(e|t) = 1$.

That is, we are restricting our attention to the set of models that can successfully account for E, either through tuning or without tuning. This assumption is legitimate in contexts in which we wish to compare models that predictively can account for E with ones that do so after being calibrated with respect to E.

(2) $p(f|t) = p(f)$

According to (2), whether M has the feature F is independent of its being tuned. This assumption, too, is plausible. We are assuming that we do not know if our models have the good-making feature F and thus, plausibly, modelers' decisions to tune a given model against a specific data set are independent of whether the model possesses F.

(3) $p(e|\neg f.\neg t) = \partial < p(e|f.\neg t)$

According to (3), the probability that a model can account for evidence E, if it is neither tuned nor possesses the good-making feature F is extremely small and is smaller than the probability that an untuned model with feature F can account for the evidence. In effect, I am assuming that there are two independent ways in which a model might account for E: either by being tuned against E or by possessing F. (3) states that the probability of a fluke—that is, that M accounts for E through neither of these two ways—is extremely small.

The three assumptions are jointly sufficient to prove the following theorem:

Theorem: Let e, f, t, be three elements of an algebra $\boldsymbol{A}$ with associated probability measure p, and let the following three conditions be satisfied:

(1) $p(e|t) = 1$,

(2) $p(f|t) = p(f)$,
(3) $p(e|\neg f.\neg t) = \partial < p(e|f.\neg t)$

Then $p(f|e.t) < p(f|e.\neg t)$, that is, that M is adequate with respect to E is more highly confirmatory of M's possessing F if M is not tuned with respect to E than if M is tuned with respect to E.

The proof is given in the appendix.

Thus, when we are not certain, whether a model possesses certain good-making features representing projectable correlations, predictive successes can be more confirmatory of the presence of such features than mere accommodation. And arguably this is the situation we find ourselves in with respect to some of the more detailed future projections made with the help of complex climate models.

It is instructive to compare the fourth Bayesian argument to the Bayesian answer to the static problem of old evidence I discussed above. Both strategies focus on evidential relations between models and data. How can it be, then, that the former strategy does not recognize a predictive effect while the latter strategy does? There is no predictive effect, if we already know that the models under considerations either all possess a feature that makes them potentially inductively reliable or none possess such a feature. Consider Cavendish's experiment to determine the value of the gravitational constant G. We can think of this experiment as type of tuning. The gravitational law contains a free parameter G, which has to be tuned against existing data. In this case, there arguably is no predictivist effect: the same data confirm the gravitational law with the correct value for G equally when the data are used to tune G or when the data are used to test a correct guess for G's value. In this example, however, we are testing different versions of a putatively lawlike relationship.

There is equally no predictivist effect in situations involving true randomness. Consider a version of the coin flip experiment, in which we know that the coin tossing mechanism is truly random. In this example too, the evidence given by the first n tosses confirms the predictor's and the accommodator's hypotheses equally, if both hypotheses are correct with respect to the first n tosses. The fit of their hypotheses with the evidence so far provides no reason to trust their predictions and assign anything but probability ½ to the two possible outcomes of any future toss. There is no reason to think of the predictor's success as anything but a fluke.

Thus, there is a predictivist effect neither in situations in which we strongly believe to be considering inductively reliable correlations nor in contexts in which we know there are no projectable correlations. The case of complex computer simulation models differs from both these situations. Here we believe there to be underlying robust physical principles governing the climate system, yet due to the complexity of the system we do not know whether our models adequately latch onto these principles and whether the correlations exhibited in our models are projectable in the right way. Recall our discussion above. Even though climate models are built with the help of well-confirmed physical principles, the tunable parameterized equations are not derived from physical theory and tuning often seems to have the character of finding the right balance of compensating errors. Thus, the worry is that some of the correlations that result from tuning our models are highly domain dependent and might, for example, hold only for the specific boundary conditions characterizing twentieth century climate.

In situations such as this—in which it is an open question to what extent the relationships posited by our models are projectable from one context to another–successful prediction makes a difference. Just as the predictive success in Worrall's coin toss example, in which we initially do not know if the coin toss mechanism results in a projectable regularity, provides us with evidence for the existence of such a mechanism and the predictor's skill in having understood the mechanism, predictive successes of a climate model are evidence for the models' success in representing the underlying physical mechanism in a projectable manner. Predictive success is evidence that a model has in fact latched onto the underlying physical mechanism and successfully represents projectable correlations between variables of interest.

Nevertheless, in all three cases the Bayesian reply to the static problem of old evidence will posit relations of evidential support between successfully accommodated data and the models or hypotheses in question. Newton's law of gravity with successfully tuned gravitational constant is confirmed by the calibration data, as is a successfully tuned complex climate simulation and even a coin toss hypothesis that is successfully fitted to past outcomes. Yet in cases where we do not know if the model is projectable in the right way (or in cases in which we strongly suggest a hypothesis not to be projectable) a model's confirmation amounts to mere "content cutting". The model is confirmed because it adequately can account for data in its domain, yet there may be no (or only very little) reason to infer from this that its predictions with respect to other datasets are reliable.

Distinguishing carefully between confirmation that amounts to mere content cutting and confirmation that increases our confidence in the inductive projectability of a model can perhaps help explain the ambivalent attitude some computer modelers have toward calibration. On the hand, one may want to allow that calibration data can be part of the validation database (see, e.g., Oberkampf et al. 2004). On the other hand, successful calibration may amount to achieving the right answers for what appear to be the wrong reasons (see, e.g., Oberkampf et al. 2004, 352), and thus one may be inclined to exclude calibration data from what ought to count as a model's proper validation database (see Oberkampf and Barone 2006). As we have seen, successful calibration can be part of validating a simulation in that it can show that the simulation successfully accounts for a certain subset of the data in its validation domain. But unlike successful genuine prediction, successful calibration does little to increase our confidence in the models success in its intended domain of application.

41.5 Conclusion

This chapter discussed epistemological problems associated with parameter calibration focusing on calibration of complex climate models. I examined the question to what extent calibration data can confirm or validate the calibrated model, in particular through the lenses of Bayesian approaches to confirmation theory. I distinguished four different Bayesian strategies. According to a straightforward application of the Bayesian formalism, tuning data cannot confirm a model. But straightforward Bayesianism is faced with the problem of old evidence and, therefore, is of limited usefulness in discussing the epistemic import of calibration data.

Evidence-deletion strategies offer a solution to the static problem of old evidence and allow us to capture evidential relations between a model and data. These relations are insensitive to whether the evidence in question has been used in the construction of the model. Calibration data provide evidence for a model in the very same way in which data successfully predicted by a model do. Evidence-deletion strategies agree with validation metrics as developed in (Oberkampf and Barone 2006) that use-novelty is irrelevant to evaluating a simulation.

The third Bayesian strategy focuses on the dynamic problem of old evidence. According to this strategy, a model can be confirmed by old evidence when we derive that the model can account for the existing data. This strategy seems to allow a distinction between use-novel evidence, for which we can learn that it can be accounted for by the model in question, and evidence used in the construction of a model, for which this type of confirmation does not seem possible. Yet I argued that when a model is sufficiently complex—as complex climate simulation models are—then confirmation can occur even from calibration data used in the construction of the calibrated model, since we do not know prior to running the model that the calibrated model is in fact compatible with the data.

Finally, I discussed a sense in which successful tuning nevertheless provides less evidence for a model than successful prediction. In the case of correlations among variables that are not known to be robustly inductively projectable to domains beyond those for which we have data, successful prediction can provide evidence for the fact that a complex model has latched onto an inductively relevant underlying mechanism.

Appendix

We want to show that $p(f|e.t) < p(f|e.\neg t)$. (C)

$$\begin{aligned} p(f|e.t) &= p(e|f.t)p(f|t)/p(e|t) && \text{Bayes's Theorem} \\ \text{Proof:} \quad &= p(f|t) && \text{premise (1)} \\ &= p(f) = 1 - p(\neg f) && \text{premise (2)} \end{aligned}$$

On the other hand:

$$\begin{aligned} p(f|e.\neg t) &= 1 - p(\neg f|e.\neg t) = 1 - p(e|\neg f.\neg t)p(\neg f|\neg t)/p(e|\neg t) \\ &= 1 - p(e|\neg f.\neg t)p(\neg f)/p(e|\neg t) \end{aligned}$$

Thus, (C) is equivalent to:

$$1 - p(\neg f) < 1 - p(e|\neg f.\neg t)p(\neg f)/p(e|\neg t)$$

or:

$$p(e|\neg t) > p(e|\neg f.\neg t) \quad \text{(C')}$$

But (C') can be shown to follow from premise (3) as follows:

$$\begin{aligned} p(e|\neg t) &= p(f)p(e|f.\neg t) + p(\neg f)p(e|\neg f.\neg t) \\ &> p(f)p(e|\neg f.\neg t) + p(\neg f)p(e|\neg f.\neg t) \text{ premise (3)} \\ &= p(e|\neg f.\neg t). \end{aligned}$$

References

Barnes, E. C. (2008). *The paradox of Predictivism*. Cambridge, New York: Cambridge University Press.

Barnes, E. C. (1999). The quantitative problem of old evidence. *The British Journal for the Philosophy of Science, 50*(2), 249–264.

Bellprat, O., Kotlarski, S., Lüthi, D., & Schär, C. (2012). Objective calibration of regional climate models. *Journal of Geophysical Research: Atmospheres, 117*(D23). https://doi.org/10.1029/2012JD018262.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201–36). Academic Press. https://www.sciencedirect.com/science/article/pii/B9780124381506500182.

Brush, S. G. (1994). Dynamics of theory change: The role of predictions. In *PSA Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1994 (January) (pp. 133–45).

Christoph, B., Reto, K., & Gertrude, H. H. (2017). Building confidence in climate model projections: An analysis of inferences from fit. *Wiley Interdisciplinary Reviews: Climate Change, 8*(3), e454. https://doi.org/10.1002/wcc.454.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford University Press.

Douglas, H., & Magnus, P. D. (2013). State of the field: Why novel prediction matters. *Studies in History and Philosophy of Science Part A, 44*(4), 580–589.

Ellery, E., & Fitelson, B. (2000). Comments and criticism: Measuring confirmation and evidence. *Journal of Philosophy, 97*(12), 663–72.

Ellery, E., & Fitelson, B. (2002). Symmetries and asymmetries in evidential support. *Philosophical Studies, 107*(2), 129–42.

Frisch, M. (2015). Predictivism and old evidence: A critical look at climate model tuning. *European Journal for Philosophy of Science, 5*(2), 171–190. https://doi.org/10.1007/s13194-015-0110-4.

Garber, D. (1983). Old evidence and logical omniscience in Bayesian confirmation theory. http://conservancy.umn.edu/handle/11299/185350.

Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres, 113*(D6), D06104. https://doi.org/10.1029/2007JD008972.

Glymour, C. (2010). Why I Am Not a Bayesian. In *Philosophy of Probability: Contemporary Readings*. Routledge.

Glymour, C. N. (1980). *Theory and evidence*. Princeton, N.J.: Princeton University Press.

Golaz, J.-C., Horowitz, L. W., & Levy, H. (2013). Cloud tuning in a coupled climate model: impact on 20th century warming. *Geophysical Research Letters, 40*(10), 2246–2251. https://doi.org/10.1002/grl.50232.

Golaz, J.-C., Salzmann, M., Donner, L. J., Horowitz, L. W., Ming, Y., & Zhao, M. (2010). Sensitivity of the aerosol indirect effect to subgrid variability in the cloud parameterization of the GFDL atmosphere general circulation model AM3. *Journal of Climate, 24*(13), 3145–3160. https://doi.org/10.1175/2010JCLI3945.1.

Held, I. M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society, 86*(11), 1609–1614. https://doi.org/10.1175/BAMS-86-11-1609.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., et al. (2016). The art and science of climate model tuning. *Bulletin of the American Meteorological Society, 98*(3), 589–602. https://doi.org/10.1175/BAMS-D-15-00135.1.

Howson, C. (1991). The ‘Old Evidence’ problem. *The British Journal for the Philosophy of Science, 42*(4), 547–555. https://doi.org/10.1093/bjps/42.4.547.

Howson, C., & Franklin, A. (1991). Maher, Mendeleev and Bayesianism. *Philosophy of Science, 58*(4), 574–585.

Intergovernmental Panel on Climate Change, ed. (2014). Evaluation of climate models. In *Climate Change 2013—The Physical Science Basis* (pp. 741–866). Cambridge: Cambridge University Press. http://ebooks.cambridge.org/ref/id/CBO9781107415324A028.

Intergovernmental Panel on Climate Change. (2015). *Climate Change 2014: Mitigation of Climate Change: Working Group III Contribution to the IPCC Fifth Assessment Report*. Cambridge University Press.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 63* (3), 425–64.

Knutti, R., Allen, M. R., Friedlingstein, P., Gregory, J. M., Hegerl, G. C., Meehl, G. A., et al. (2008). A Review of Uncertainties in global temperature projections over the twenty-first century. *Journal of Climate, 21*(11), 2651–2663. https://doi.org/10.1175/2007JCLI2119.1.

Maher, P. (1988). Prediction, accommodation, and the logic of discovery. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1988 (January): (pp. 273–85).

Masson, D., & Knutti, R. (2012). Predictor screening, calibration, and observational constraints in climate model ensembles: An illustration using climate sensitivity. *Journal of Climate, 26*(3), 887–898. https://doi.org/10.1175/JCLI-D-11-00540.1.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H. et al. (2012). Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems, 4*(3), M00A01. https://doi.org/10.1029/2012MS000154.

Oberkampf, W. L., Trucano, T. G., & Hirsch, C. (2004). Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews, 57*(5), 345–384. https://doi.org/10.1115/1.1767847.

Oberkampf, W. L., & Barone, M. F. (2006). Measures of agreement between computation and experiment: Validation metrics. *Journal of Computational Physics, Uncertainty Quantification in Simulation Science, 217*(1), 5–36. https://doi.org/10.1016/j.jcp.2006.03.037.

Parker, W. S. (2009). Confirmation and adequacy? for? purpose in climate modelling. *Aristotelian Society Supplementary Volume, 83*(1), 233–249.

Parker, W. S. 2010. Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in History and Philosophy of Science Part B* 41 (3): 263–272.

Parker, W. S. (2013). Computer simulation. In S. Psillos & M. Curd (Eds.), *The Routledge Companion to Philosophy of Science*, 2nd Edition. Routledge.

Sprenger, Jan. (2015). A novel solution to the problem of old evidence. *Philosophy of Science, 82*(3), 383–401. https://doi.org/10.1086/681767.

Steele, K., & Charlotte, W. (2016). Model-selection theory: The need for a more nuanced picture of use-novelty and double-counting. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axw024.

Worrall, J. (1980). *001: The methodology of scientific research programmes: Philosophical papers* (Vol. 1). Cambridge: Cambridge University Press.

Worrall, J. (2014). Prediction and accommodation revisited. *Studies in History and Philosophy of Science Part A, 45*(March), 54–61. https://doi.org/10.1016/j.shpsa.2013.10.001.

Chapter 42
Should Validation and Verification be Separated Strictly?

Claus Beisbart

Abstract Verification and validation are methods with which computer simulations are tested. While many practitioners draw a clear line between verification and validation and demand that the former precedes the latter, some philosophers have suggested that the distinction has been over-exaggerated. This chapter clarifies the relationship between verification and validation. Regarding the latter, validation of the conceptual and of the computational model are distinguished. I argue that, as a method, verification is clearly different from validation of either of the models. However, the methods are related to each other as follows: If we allow that the validation of the computational model need not include the comparison between simulation output and measured data, then the computational model may be validated by validating the conceptual model independently and by verifying the simulation with respect to it. This is often not realistic, however, because, in most cases, the conceptual model cannot be validated independently from the simulation. In such cases, the computational model is verified with the aim to use it as an appropriate substitute for the conceptual model. Then simulation output is compared to measured data to validate both the computational and the conceptual model. I analyze the underlying inferences and argue that they require some prior confidence (i) in the conceptual model and (ii) in verification. This suggests that verification precede validation that proceeds via a comparison between simulation output and measured data. Recent arguments to the effect that the distinction between verification and validation is not clear-cut do not refute these results, or so I argue against philosopher E. Winsberg.

Keywords Computational model · Conceptual model · Reality · Accuracy · Comparison simulation output vs. data · Inductive inference

C. Beisbart (✉)
Institute of Philosophy, University of Bern, Bern, Switzerland
e-mail: claus.beisbart@philo.unibe.ch

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_42

42.1 Introduction

Among simulation scientists from many fields, the term “V & V” is known to denote activities of verification and validation. V & V does not only require that researchers verify their computer simulations (or their software, more generally), i.e., that they show their simulations to provide approximate solutions to the model chosen. Scientists are also expected to validate their simulations in view of the intended applications, i.e., to show that the simulation represents the target system with sufficient accuracy for the chosen range of applications. But how are verification and validation related to each other? What is the reason to combine both in what is called V & V? Is there are a neat and tidy separation between verification and validation? And if so, should verification always precede validation?

The last two questions have recently been a matter of discussion. Many practitioners draw a clear distinction between verification and validation. As Murray-Smith (Chap. 4 in this volume, Sect. 42.4.1) reports, “there is general agreement that the process of simulation model testing involves two issues one of which concerns the correctness, or otherwise, of the process of translating the conceptual, mathematical and logical basis of a model into the description implemented on a computer. The other issue is concerned with potential errors and uncertainties within the structure and logic of the underlying model, along with limitations of that description in terms of accuracy.” It is further often recommended that verification be completed before validation starts (see e.g., Fig. 3.3 in Chap. 3 by Oberkampf in this volume). Some philosophers, by contrast, notably Winsberg (2010, pp. 19–25; 2018a, Sect. 4.3 and 2018b, pp. 155–160), have argued that the distinction between verification and validation is not as clear-cut as some might have thought. In this vein, Lenhard argues in his Chap. 39 in this volume that “[t]he separation of verification and validation [...] cannot be fully maintained in practice”.

The aim of this chapter is to clarify the relationship between verification and validation. Our focus is on computer simulation, although most of the points made carry over to other software that is used to represent real-world systems. The chapter is written by a philosopher first because it is philosophers who have challenged the idea that verification and validation are distinct activities, and second, because the task of the chapter is to clarify concepts and to assess arguments, which is typical of philosophical inquiry.

The chapter is organized as follows: Sect. 42.2 addresses philosophical preliminaries and explains, for instance, what a clear distinction between verification and validation would amount to. Section 42.3 turns to the possibility of a conceptual distinction between verification and validation and to the relationship between both methods. I argue that verification and a suitable sort of validation can be combined in interesting ways and analyze the underlying inferences. A critical discussion of philosophical arguments challenging the distinction can be found in Sect. 42.4. I draw my conclusions in Sect. 42.5. I hope to obtain a picture of V & V that integrates the most important insights about V & V so far.

42.2 Preliminaries

What would it mean to say that there is (not) a clear distinction between verification and validation?

42.2.1 Scientific Methods

Validation and verification are *scientific methods*. Computer simulation or experimentation are other examples of such methods. In general terms, scientific methods are *types of activities* that researchers can choose to pursue the aims constitutive of science, e.g., to gain new knowledge. A specific series of activities carried out by a researcher can thus be classified as a *token* of a method, say, verification. One such token of applying a method often consists of various sub-activities (which are often called steps). For instance, to verify a simulation, a scientist may first compare the output of the simulation program with a known analytical solution and then apply the method of manufactured solutions (see Chap. 11 by Rider and Chap. 12 by Roache in this volume).

How are methods identified and individuated? That is, what makes it the case that a certain series of activities counts as application of one method rather than another? What is essential for a specific method is often an overarching goal. Consider, for instance, the following statement by Oberkampf, in his Chap. 3 in this volume (Sect. 3.4.1):

> Model validation, as defined here, is focused on the assessment of the error due to the approximations and assumptions made in the formulation of the conceptual and mathematical models.

It is natural to read this as a claim about the purpose of model validation: Model validation is aimed at assessing certain errors.

Methods are often too constrained by conditions on what is done during the application of the method. Frequently, a certain step is demanded for the execution of the method. In this way, the method of experimentation crucially involves the observation of the object that is experimented on. And when Oberkampf, later in his Chap. 3 in this volume (Sect. 3.4.2.1) writes: “Model validation […] is the activity of quantitatively assessing model accuracy by way of comparison of simulation results with experimental measurements”, he does not only specify an aim (quantitative assessment), but further requires that this assessment be done in a particular way, namely by comparing simulation output with measured data.

Methods can be related to each other in various ways. For instance, two different methods may be used one after the other to help accomplish the same super-ordinate goal. Alternatively, one method may be used during the application of another method, either because this is required or because the former method just proves suitable in a specific context. It is plausible to think, for instance, that the method of computer simulation requires validation and verification as necessary steps. If a person carries out one method as a way to at least partly apply another method, she can be described as applying both methods at the same time. Thus, a certain activity may count both as a step of an experiment and of a simulation. The upshot is that methods and action types, more generally, are not always exclusive categories (cf. Anscombe 2000, Sect. 23).

42.2.2 Verification and Validation

Let us now turn to verification and to validation, more specifically. Clearly, whether there is a sharp distinction between the methods depends on what the methods are. In the literature, we find various definitions that try to explain what verification and validation are. In what follows, we will mainly rely upon the definitions provided by the AIAA guide (AIAA 1998), but we will also explain connections to other definitions that are quite similar. We will later discuss to what extent our results are affected when some very different definitions are assumed.

The AIAA guide defines *verification* as follows:

> *Verification:* the process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model (AIAA 1998, here quoted after Oberkampf and Roy 2010, p. 26).

According to this definition, the proper object of verification is a *model implementation*. It is natural to think of the model implementation as a *computer code*. The idea that such a code is verified is also implicit in the so-called SCS-definition of verification (Schlesinger et al. 1979, pp. 103–104).[1] This definition takes it that a *computer program* is verified, which we take to be the same as a computer code. The ASME guide (ASME 2006) seems to suggest a different object of validation, because a *computational model* is said to be subject to verification. But for our purposes, the code does not relevantly differ from the computational model. True, a model is something different from a code because the former is, e.g., a set of equations, while a code is a set of instructions. However, in practice, this difference in category between a code and a model is irrelevant. Each computer simulation code can be said to define a unique model as follows: Carrying out the instructions from a code yields outputs that are interpreted as values of characteristics such as air pressure, precipitation, etc., at various times. The time-ordered series of output numbers can be interpreted as solutions to equations. Running the computer simulation program solves these equations exactly. In a source code written in, e.g., C++, the equations can in fact easily be seen from the program. In a simple example, the equations are difference equations that arise if the so-called Euler method is used to approximate differential equations. Now we can say that the equations solved by running the computer simulation code form the core of a dynamical model; and this model is aptly called the computational model. For this reason, a model is implicit in the program, and we can gloss over the difference between the code and the computational model.

[1] The definition was recommended by the Society for Modeling and Simulation International (SCS).

As far as the code is concerned, we will focus on the machine code, as run on a particular hardware, when we talk about the computer simulation code. This ensures that the program and the corresponding computational model imply what is actually output. By contrast, the instructions of the source code (and a model corresponding to it) do not yield the output, strictly speaking, because following the instructions would, e.g., not produce any round-off errors, which typically affect the actual output. The output may also differ from the one that would arise by strictly carrying out the instructions from the source code if there is a compiler error.

According to the AIAA definition of verification, the implementation of the model is supposed to represent the *conceptual description of a model*. This presupposes a clear contrast between the computational model, as we have called it, and what is here named the conceptual description of a model. But what exactly is the latter? When researchers are asked to describe their simulations, they usually refer to a model, e.g., a fluid-dynamical model that has the Navier–Stokes equations as its mathematical core. This model is different from the computational model: Since the computer cannot solve differential equations such as the Navier–Stokes equations, the computational model approximates the latter using difference equations.

The model that is used for a natural description of simulation is often called *conceptual model*, e.g., by Schlesinger et al. (1979, p. 103), see also Oberkampf and Roy (2010, p. 38). This model can be characterized as a prior model that is typically independent of the computer simulation, motivated by the best knowledge of the system under scrutiny and thus the focus of the primary interest on the part of researchers. For instance, in fluid dynamics, the Navier–Stokes differential equations are supposed to express our best physical knowledge about certain types of fluids, if considered for a certain range of scales.[2] It is admittedly sometimes not entirely clear what the conceptual model is for a given computer simulation, because it is not unambiguous which model is the locus of primary interest or because different researchers using the program may differ in their interests. But in what follows we will assume that the conceptual model has been unambiguously identified.

Following the AIAA definition, the goal of verification then is to show that the computational model is an appropriate representation of, or substitute for, the conceptual one. In saying this, we assume that models, or representations, are substitutes for other systems.[3] What it here means to say that a representation is *appropriate* may be explained in terms of accuracy (cf. Schlesinger et al. 1979, p. 104). The idea here is that, for a certain range of applications, the computational model represents the conceptual one appropriately if, and only if, their solutions coincide within certain bounds, or to a certain accuracy. If the solutions to both models can be expressed in terms of numbers, then accuracy can easily be measured by taking distances between solutions to both models. This is sometimes not possible for qualitative (aspects of) solutions, i.e., that there is a thunderstorm this night. Even in such situations, we can say that a simulation is more accurate than another if its output agrees more often

[2] It is, of course, allowed that the conceptual model is in some way simplified, e.g., idealized. Which simplifications are appropriate depends on the intended uses of the model.

[3] See e.g., Suárez (2004) and Weisberg (2007) for related views of modeling and representation.

with what would be expected for the conceptual model (see Chap. 2 by Beisbart in this volume for more on accuracy).

Verification is often divided into code and solution verification. While the former is supposed to show that the approximate solutions obtained by the simulation program converge to some known solutions to the conceptual model, the latter is focused on determining the uncertainties about errors due to numerical approximation schemes in a simulation output (see Rider, Chap. 11 in this volume). Code verification is supposed to include software quality assurance (Oberkampf and Roy 2010, p. 32).

Summarizing we can say that verification is supposed to show that the conceptual model is appropriately implemented in the computer code. The criterion of appropriateness is accuracy, meaning that the deviations between solutions to the conceptual and the computational model are small.

Turn now to *validation* (see Chap. 2 by Beisbart for a comparison between various definitions of validation). As far as the object of validation is concerned, it is sometimes said that specific results from a computer simulation need validation (see Chap. 2 by Beisbart in this volume for evidence and a related discussion). But, during validation, it is more interesting (and more challenging) to consider all possible results that may be obtained using a computer simulation program by varying the initial conditions and parameter values within some realm of intended applications. When this plurality of possible results is considered, we are effectively talking about a model associated with the simulations, or, more precisely, about certain aspects of this very model, e.g., about its accuracy in predicting, say, precipitation. In what follows, we will thus take it that the proper object of validation is a whole simulation, and not just a small set of specific results.

The AIAA guide defines validation of a model as follows:

> Val-AIAA The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model [here quoted after Oberkampf and Trucano 2008, p. 719]

The aim of validation thus is to show that a model accurately represents the real world, or more specifically, the target system, in some respects that depend on the intended uses. The accuracy of the representation depends on how accurate relevant predictions of the model are as compared to the truth, i.e., to the true values that obtain in reality. Here, "prediction" is used as an umbrella term for all sorts of implications that the model has in terms of its solutions. The accuracy of the relevant quantitative model predictions can in principle be quantified by taking and aggregating the distances between model output and corresponding true values in reality for the relevant characteristics.

It is arguable that validation need not only show that the model *predictions* are accurate enough for the intended uses, but also that underlying model *assumptions* are correct or accurate (see Chap. 2 by Beisbart in this volume). However, in what follows, we will focus on the accuracy of the model predictions.

Talk of an "accurate representation of the real world" presumes that a computer simulation has a target in the real world. Many simulations do have such a target, e.g., the climate system of our planet. But this is not always the case; for instance, some computer simulations are aimed at tracing the dynamical evolution of merely imagined systems, e.g., point particles that interact via forces that do not work in the real world. For such simulations, the question of an accurate representation of a real-world system does not arise, so validation is not an issue anymore. In what follows, we will concentrate on simulations for which validation is a concern.[4]

So far, we have been talking about the validation of models quite generally, as do many definitions of validation (e.g., AIAA 1998, ASME 2006). However, for computer simulation, the question arises whether their validation should focus on the conceptual or the computational model. The SCS-definition of validation takes a clear stance on this question by defining the validation of a computational model (Schlesinger et al. 1979, p. 104). The reason is presumably that, during validation, often output from the computer model is compared to measurements from reality.

But there are also reasons to focus on the conceptual model during validation of a simulation. First, the conceptual model is at least typically the model that scientists are primarily interested in and that they have chosen on the basis of their scientific knowledge. The computational model, by contrast, is a crutch that is only used because the conceptual model cannot be solved exactly. Second, the conceptual model is typically well-known to scientists and easy to express, e.g., in terms of differential equations. Things are different regarding the computational model implicit in the machine code. The latter is often not inspected. If there is an unknown error, e.g., in the compilation of the source code, then the computational model implicit in the machine code is not really known. Also, the computational model implicit in the source code is often difficult to express, for instance, because it cannot be cast in terms of differential equations. We will thus leave open whether the proper object of validating a simulation is a conceptual or a computational model. Instead, we distinguish between the validation of the conceptual model and that of the computational model.[5]

42.3 The Distinction Between Verification and Validation

After these preliminaries, we can address the questions of whether a clear distinction between verification and validation can be drawn and whether the former should always precede the latter.

As methods, verification and validation are clearly different, because they have different aims. Verification is supposed to show that the computational model is an

[4] There are interesting boundary cases, e.g., simulations that study a real-world target system under very counterfactual circumstances. The question to which extent such simulations can be validated is beyond the scope of this chapter.

[5] The AIAA definitions of verification and validation each assume that a degree of accuracy be determined. As it happens, particularly in validation, it is very difficult to *prove* that a certain extent of accuracy is present. What is realistic is at most to make it very *credible* or likely that a certain accuracy holds (see Chap. 2 by Beisbart in this volume). In what follows, we need not always make this explicit.

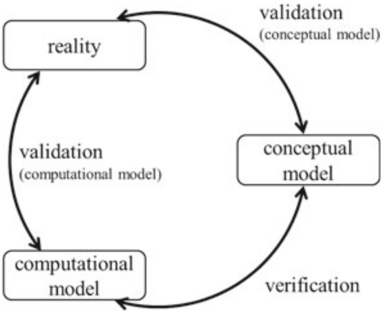


Fig. 42.1 The Sargent circle following Schlesinger (1979, p. 103) adapted for our purposes

adequate representation of the conceptual model, while validation has the objective of showing that the conceptual or the computational model is an adequate representation of the target. As it is often put in terms of equations, verification is about "solving the equations right" (i.e., about really solving the equations from the conceptual model), while validation is about "solving the right equations" (i.e., solving equations that provide adequate representations of the target; e.g., Roache 1997, p. 124, who refers to other authors, however). Accordingly, verification is often said to be a matter of mere mathematics, while validation is supposed to deal with physics or the sciences involved (see e.g., Roy 2005, p. 132, who refers to AIAA 1998). It is arguable that this view is not strictly speaking true (see below for more discussion; see also Chap. 11 by Rider in this volume), but it certainly gives a good sense of how verification and validation differ. Another way of stressing the difference between the methods is to say that verification is *internal* in the sense that it only deals with models and their implementation, while validation is *external* because reference to the target system is made (see e.g., Chap. 4 by Murray-Smith in this volume; see Chap. 15 by Murray-Smith in this volume for the use of experimental data).

The basic difference between verification and validation is often illustrated using a viewgraph called the "Sargent circle", as shown in e.g. Schlesinger et al. (1979, p. 103; see our Fig. 42.1).[6] The circle connects three "nodes", viz., reality (i.e., the target system), the conceptual model and the computational model. Model verification is said to relate the conceptual and the computational model. Model validation, by contrast, makes a link to reality, either from the conceptual model or the computational model. In what follows, it will be useful to have this picture in mind where i. *validation of the conceptual model* is represented by the arrow between the conceptual model and reality, ii. *validation of the computational model* is represented by the arrow between the computational model and reality, and iii. *verification* is represented by the arrow between the computational model and the conceptual model.

Although verification and validation are clearly different at the general level of methods (or types of activities), they are closely related to each other. *First*, both are species of the same genus in that they can both be considered as evaluations of

[6] According to Pace (2004, p. 124), the circle goes back to R. Sargent and was popularized by Schlesinger et al. (1979). See Sargent (1984, p. 116) for a version given by Sargent.

a representation, accuracy being the main standard.[7] As a consequence, the applications of both methods are often very similar. This can be explained using the notion of a benchmark (see Chap. 18 by Saam in this volume): During both validation and verification, the output of computer simulations is compared to benchmarks. If these are analytical solutions to the equations from the conceptual model, then we are engaged in verification. By contrast, if empirical data from the target system are used as benchmarks, we are concerned with validation. The close similarity between verification and validation may explain why validation and verification are sometimes confused.

Second, the methods are related to each other because there is an overlap between the entities considered during verification and validation. Verification and validation of the conceptual model both involve the conceptual model. In an analogous way, verification and validation of the computational model both refer to the computational model. This may explain why validation and verification are taken together in V & V.

Third, some of the three methods under consideration (verification, validation of the conceptual model, validation of the computational model) may be used for the sake of executing another method from the set. This needs a more extensive discussion.

42.3.1 Verification as a Means of Validation of the Computational Model?

Consider first the validation of the computational model. The task here is to show that the results of the computational model agree with the truth to such and such accuracy (see Chap. 2 by Beisbart in this volume for the notions of truth and accuracy). The relationships visualized in the Sargent circle allow that this can in principle be done as follows.

Suppose first that a computational model has been verified with respect to some conceptual model. That is, researchers have substantiated or shown that the solutions to the computational model coincide with those of the conceptual model to such and such accuracy. Assume that the accuracy can be expressed using a distance measure between the output of the simulation for some characteristic C (e.g., precipitation), on the one hand, and the output that would be expected for C had the conceptual model been solved exactly, on the other. We assume that the value of this measure is $\mathrm{A}_{\mathrm{con,com}}$.

Suppose now that the conceptual model is known to represent the target system regarding C with a certain accuracy. We measure the degree of accuracy with which

[7] For a discussion of whether validation should include other standards see Chap. 2 by Beisbart in this volume. The following argument is easily generalized to other standards if the degree to which the latter are fulfilled can be measured using a distance between the output of the models involved and measured data.

the model agrees with the truth using the same distance measure as before, now applied to the predictions of the model for some characteristics C and the true values. Let the degree of accuracy be $A_{\text{con,real}}$. It follows immediately that the accuracy of the computational model with respect to characteristic C, call it $A_{\text{com,real}}$, is no worse than $A_{\text{con,com}} + A_{\text{con,real}}$.[8] This is a consequence of the so-called triangle inequality for distance measures.

Altogether, we have an argument to the effect that the computational model is accurate (with respect to reality) to a certain degree. This amounts to validation of the computational model. Now the argument is partly based upon successful verification. We can thus say that verification of a simulation can make a decisive contribution to validation of the computational model. In the terms suggested in Sect. 42.2.1, we may say that verification can form an important step in validation (this strategy is also mentioned by Parker 2008, pp. 170–171).

This discussion assumes that accuracy can be measured in a quantitative way. This assumption is not necessary for the conclusion. For instance, we can also infer that a simulation is correct in a certain range of qualitative predictions, if the conceptual model is correct in these predictions, and if the predictions of both models coincide in this respect.

However, our conclusion that verification can be considered as a step in validation of the conceptual model does hinge on two crucial assumptions. The first assumption is that the conceptual model has been validated. This validation has to be independent of the validation of the computer simulation; otherwise researchers would be allowed to move in a circle that looks problematic: The computational model is validated, since, among other things, the conceptual model is validated; and the conceptual model is validated because, among other things, the computational model has been validated.[9]

Now it is very often impossible to validate the conceptual model independently from computer simulations. A conceptual model may to some extent be justified because it is built on a theory which is well confirmed with respect to the target. For instance, scientists have no doubts to apply Newtonian mechanics to a macroscopic pendulum, if the gravitational fields are weak and relative velocities small. However, even in such a case, the theory does not imply what the motion of the pendulum is like because additional information about the initial conditions and the shape of the pendulum is needed (in the terms of philosophy of science, the so-called auxiliary hypotheses are required to obtain consequences from the theory). In special cases, we may have accurate measurements that can be combined with the theory, and it may be shown that the accuracy of the measurements together with the theory implies a certain accuracy for the results. However, examples of this type are extremely rare.

[8] Properly speaking, our measures A are distance measures or measures of the error. Accordingly, the accuracy is the higher the smaller the value of A is. What we show here is that the error of the conceptual model is no larger than a certain sum, which is to say that the accuracy is no less than the accuracy that corresponds to the sum.

[9] We will later see that the circle just mentioned can in some sense be rendered acceptable, see Sect. 42.3.2.

In most applications of computer simulations, it is not, and cannot be, known how accurately the assumptions of the conceptual model represent reality.

There is a second assumption that is needed for the conclusion that verification can be a crucial step in the validation of the computational model. This assumption is about the concept of validation and has it that validation need not compare predictions of the computational model with measured data. Some definitions of validation deny this. For instance, in the second characterization of validation quoted above from Oberkampf (Chap. 3 in this volume), he requires that simulation results be compared with data from measurements. Oberkampf thus constrains the means that can be chosen for validation. Due to this restriction, the argument that has been constructed for the accuracy of the computational model above would not qualify as validation, since it does not involve a comparison between simulation output and measured data.

To summarize then, we have shown that verification is a crucial step in a particular strategy to validate a computational model, if validation of such a model does not generally require the comparison between model predictions and measured data. The strategy presumes that the conceptual model has been validated on independent grounds. If this strategy is adopted, then verification is done for the sake of validation of the computational model.

The argument we have suggested for the validation of a computational model can also be turned around as follows: Suppose that some simulation output and corresponding true values are found to disagree with each other with some accuracy on data points about characteristic C (this may be found on the basis of data, where we assume that there is no evidence that the measured data get it substantially wrong). It then follows that it cannot be the case that (i) the computational model is sufficiently verified with respect to C *and* that (ii) the conceptual model is sufficiently accurate regarding C. Thus, if verification has been secured with a sufficient accuracy, then the conceptual model is not sufficiently accurate. This shows that verification can be crucial for *in*validation instead of for validation (see Chap. 6 by Beven in this volume on invalidation). However, sometimes, researchers may not be sure whether or not the simulations have been verified properly. If it then turns out that simulation output and measured data do not agree, then they may not be able to decide what the problem is: Is the conceptual model not sufficiently accurate with respect to reality or is the computational model not sufficiently accurate with respect to the conceptual model (or both)? That is, is validation of the conceptual model or verification doomed to fail? Provided that there is no further background knowledge, researchers face an underdetermination problem because they cannot decide where the problem is. We may say that the researchers cannot distinguish between verification and validation of the conceptual model in some sense because they cannot tell whether verification or validation of this model would fail. In this sense, the distinction is not transparent anymore from the viewpoint of researchers. But this does not show that the distinction does not exist. The underdetermination can be broken if the computer simulation is further investigated, e.g., if it can be shown that there is a problem in the discretization of differential equations.

42.3.2 Verification as Means for Validation of the Conceptual Model?

Let us know focus on the validation of the conceptual model. The relationships shown in the Sargent circle suggest a two-step method for validation of the conceptual model.

For suppose that a computational model has been verified with respect to the conceptual model and that the degree of accuracy that has been determined for some characteristic C is $A_{\text{con,com}}$. Assume further that the computational model has been validated with respect to C and that the accuracy is $A_{\text{com,real}}$. Then we can conclude that the accuracy with which the conceptual model represents the target with respect to characteristics C, $A_{\text{con,real}}$, is no worse than $A_{\text{con,com}} + A_{\text{com,real}}$. So the conceptual model has been validated. In this way, verification can be crucial for validation of the conceptual model.

Unlike our conclusion in Sect. 42.3.1, this conclusion does *not* assume that validation can be done without comparing model output with data. If validation of a model is thought to require that model output be compared to measured data, then we can argue as follows: The strategy just considered for the validation of a conceptual model requires validation of the computational model. Provided the assumption that validation requires an empirical comparison, this validation must involve the comparison between the output of the computational model and measured data. But given that the computational model has been verified with respect to the conceptual model, the comparison between output of the computational model and measured data may be understood as a comparison between output from the *conceptual* model and measured data, where output of the computational model is used as proxy for (typically unavailable) output of the conceptual model. So the validation of the conceptual model too is based upon a comparison of model output with measured data.[10]

However, the research strategy just sketched for obtaining validation of the conceptual model requires that the computational model needs to be validated *independently* from the conceptual model. For, if the computational model was only shown to be accurate because it is an adequate representation of the conceptual model, then the validation of the conceptual model would move in a problematic circle. It may first seem that there is no problem at this point because the comparison between simulation output and measured results is independent of the conceptual model and does not presume the validity of the conceptual model. But at closer analysis, things are more complicated.[11] Empirical validation of a computational model, i.e., the comparison between simulation output and measured data uses a limited sample of data

[10] A word of caution about the comparison between simulation output and measured data. This comparison is only significant for validation, if the measured data reflect the true values. This condition is often not met, there are errors and uncertainties in the data and significant efforts are needed to show that the data reflect the true values to some accuracy. To properly take this into account, we would have to elaborate the Sargent circle. We refrain from doing so and neglect measurement errors because our focus is on errors that are incurred during modeling and simulating a target system.

[11] To simplify the presentation, we will from now on neglect the precise values of the accuracies. This does not bear on the power of the argument.

points. From an agreement between simulation output and measured data on these data points with some accuracy, it is concluded that the agreement holds with this accuracy more generally (i.e., for other results that may be obtained by running the simulation). This is an inductive inference, i.e., an inference that adds content to what is claimed in the premises (which here claim agreement for a limited number sample of data points). More specifically, we are talking about enumerative induction, in which a limited number of instances are used to infer a more general claim.

Now it is well-known that enumerative induction is not always legitimate. For instance, it is not legitimate to conclude from the fact that all children form a certain grade of school are between 10 and 11, that all school children are between 10 and 11. Nor is it legitimate to infer from the fact that the last few people one has seen were female that one will next see females too. As Harman (1965) has suggested, enumerative induction is only justified if the conclusion is part of the best explanation of why the premises hold. And what one can take to be the best explanation depends on one's background knowledge.

The consequence for validation of the computational model is as follows: The inductive inference implicit in it is only legitimate if the conclusion that the computational model represents the target in a certain respect with sufficient accuracy is part of the best explanation of why the simulation output and the measured data agree with some accuracy. Now the natural explanation of this agreement is that the simulation is built on a conceptual model that is sufficiently accurate in a certain respect and that the computer simulation is verified with respect to this model. This explanation would imply that all output from the simulation program makes accurate predictions. However, we can barely take this explanation to be best unless we have some reasons to take the conceptual model to form a sufficiently accurate representation of the target. Suppose, for instance, that the model has been constructed by combining some randomly chosen assumptions about the target. Under this assumption, the best explanation of the agreement is more likely that it is due to a fluke. The consequence is that a comparison between output from the computational model and measured data only allows for validation of the computational model if there are some reasons that speak in favor of a sufficient degree of accuracy of the conceptual model.

It may seem that this leads to a problem for the strategy to validate a conceptual model by verifying and validating a computational model. The apparent problem is that the validation of the computational model depends on a sufficient degree of accuracy established for the conceptual model, but the very task of validating the conceptual model is to establish this accuracy.[12] The threat then is that we are moving in a vicious circle if we wish to validate a conceptual model in the way under consideration.

But the threat is not real. There would be vicious circularity if we would have to *know* the accuracy of the conceptual model to conclude that the accuracy of this very model and the successful verification best explain why the simulation output

[12] The accuracy needed in the explanation is, in fact, higher than the accuracy established during validation.

and the measured data agree. But to be able to pick this explanation as the best one, we need only to have *some* independent reasons to take the conceptual model to be accurate and to be properly reflected in the simulations. The idea here is that the identification of the best explanation requires a comparative assessment of various candidate explanations. Typically, none of them is *known* to be true, so the question is which one is best supported by the available evidence or which one fits best with our background knowlege. The explanation that assumes (i) that the conceptual model provides a sufficiently accurate representation of the target and (ii) that the computational model is a suitable representation of the conceptual model, may turn out to be best if there are some reasons in favor of it; for instance more reasons than for an explanation in terms of a fluke. Now such reasons are available if (a) verification has made a strong case for (ii) and (b) the assumptions from the conceptual model are supported by independent evidence. For instance, as indicated before, the model may be built on assumptions from well-confirmed theory. Further, some of the auxiliary assumptions needed to construct the model may be justified in terms of measured data, e.g., about the initial conditions. Thus, if the simulation was verified and if there is some evidence for the conceptual model, the explanation under consideration should turn out best. Thus, if a simulation is verified and if there is some evidence in favor of the accuracy of the conceptual model, then agreement between simulation output and measured data allows researchers to infer that the agreement extends to other possible runs of the simulation program. In this way, then, the computational model inherent in the simulation can be validated.[13]

Altogether then, our discussion leads to the following "reconstruction" of validating a simulation using measured data, where both the conceptual and the computational models are validated. 1. The simulation is verified with respect to the conceptual model regarding characteristic C. 2. Simulation output and measured data about characteristics C are compared to each other. Suppose that they agree with some accuracy (or up to some errors). The researchers conclude from this that simulation output and measured data would agree about C with a certain accuracy for a larger domain of applications. The researchers are justified to conclude this because it follows from the best explanation they have for the agreement between simulation output and measured data, viz., that the conceptual model is sufficiently accurate with respect to C and that the computational model is properly reflected in the simulation output, i.e., that it is verified with respect to the conceptual model in respect C. The researchers are justified in taking this to be the best explanation because they have some evidence in favor of the accuracy of the conceptual model, e.g., because it is built on well-confirmed theory, and because they have verified the simulation or at least taken some steps towards verification such that they can be reasonably confident that the conceptual model is implemented with sufficient accuracy regarding C. In this way, validation of the computational model is secured in terms of an inductive inference. 3. The triangle inequality for distance measures is used to

[13] Note that our argument does not hinge upon Harman's view about enumerative induction. Our argument would also go through if we employed the so-called material theory of induction as defended by Norton (2003).

show that the validation of the computational model and verification together lead to a certain degree of accuracy of the conceptual model, which achieves validation of the conceptual model.[14]

It may be objected that the first step in this reconstruction is not really necessary for the second step and thus for the validation of the computational model (it is clear that it is needed for the third step). The idea might be that some evidence to believe the accuracy of the conceptual model suffices to argue that the agreement between simulation output and measured data is best explained in terms of the explanation mentioned above (i..e, that the conceptual model is sufficiently accurate and that the simulation, i.e., the computational model, has been verified): If one crucial part of the explanation (i.e., validation of the conceptual model) has some evidence, then researchers can take this to be the best explanation and safely infer that the other part of the explanation (i.e., verification) holds too, or so the idea is. While this is in principle true, it will most often not help because the available evidence for the conceptual model is typically not too strong. Further, researchers can certainly be more confident to have picked the correct explanation if there is some evidence that the second part of the explanation holds too. And the more confidence there is to have chosen the correct explanation, the more secure is the inductive inference. Thus, to make a strong case for the accuracy of the computational model using the inductive inference in realistic cases, researchers need at least some evidence that there is verification.

If this is correct, our reconstruction of validation using data (of data-driven validation, for short) shows that verification (at least some effort into verification) is crucial for the validation of both the computational and the conceptual model using data. Since verification is needed for the inductive inference implicit in the second step (validation of the computational model), the reconstruction suggests that verification should be done *before* simulation output and measured data are compared to each other during validation. Note that an essential part of verification, viz. software quality assurance, should be done first anyway, because it is needed to ensure that the program does what scientists intend it to do at a certain coarse level of description. For instance, scientists want to approximate the conceptual model in a particular way, and the software quality assurance is needed to show that they do so.

Depending on how exactly we think about data-driven validation, there are two ways to conceptualize our conclusion. For a first option, we may think that validation is data-driven if, and only if, it involves the comparison between computer output and measured data. We may then say that validation of the computational model encompasses steps 1 and 2 above, and that validation of the conceptual model includes steps 1–3. This would mean that verification is a necessary step in both the data-driven validation of the computational and the conceptual model.

For a second option, we may think that data-driven validation has not only to include a comparison between computer output and measured data. Rather, the idea is too that the data-driven validation is exhausted by the comparison between simu-

[14] This reconstruction can also be understood as showing that the circle identified in Sect. 42.3.1 can be rendered unproblematic.

lation output and measured data and inferences built upon the comparison. Call this narrow view of data-driven validation. Under this view, verification is a necessary precondition of data-driven validation.

However, we decide between both options, we obtain a *two-step view* for data-driven validation: Researchers first need to verify their simulations before they compare simulation output to measured data. As indicated above, this view is common among practitioners. For instance, in the viewgraph in Fig. 2.15 in Oberkampf and Roy (2010, p. 59), most parts of verification (viz. code verification and software quality assurance) are required very early, e.g., before measurements are done. Only the so-called solution verification (see Chap. 11 by Rider in this volume) is done a bit later, but still before the measured data are compared to simulation output.[15]

Before we conclude this section, an additional remark is in order. It may be suggested that the relationships within the Sargent circle allow for another method: We may obtain verification of a simulation by validating the conceptual and the computational models and putting the results together as follows. If the computational and the conceptual model both agree with measured data with some accuracy, then they must agree with each other with some accuracy due to the triangle inequality.

This is true mathematically, but insignificant. First, as argued in Sect. 42.3.1, a conceptual model can often not be fully validated independently from a simulation. Second, as argued in Sect. 42.3.2, validation in terms of a comparison between simulation output and measured data is based upon an inference from a small sample to a larger set of possible applications of the computational model in some domain. This inference is only warranted if the agreement between simulation output and measured data can be explained in terms of an accurate model and verification. This shows that the strategy cannot produce verification from scratch.

42.4 Arguments Against a Clean Separation Between Verification and Validation

As indicated in the introduction, some philosophers have argued that the distinction between verification and validation has been over-exaggerated, to say the least. In particular, Winsberg has reached something like this conclusion in several of his works (Winsberg 2010, pp. 19–25; 2018a, Sect. 4.3 and 2018b, pp. 155–160). Lenhard (e.g., in his Chap. 39 in this volume) too argues that validation and verification cannot be kept apart in practice. In what follows, I will discuss what I take to be the strongest and most interesting points raised by Winsberg and Lenhard.

Winsberg (2010, p. 20; 2018b, pp. 155–156) links the distinction between verification and validation with what might be called a linear picture of computer simulation

[15] The method discussed in Sect. 42.3.1, by contrast, i.e., the validation of a computational model using verification and validation of the conceptual model, does not assume that verification is done before validation of the conceptual model. But under the method, validation of the computational model cannot be achieved unless verification is done.

(2018b, p. 155).[16] This picture assumes that, as a practice, simulation begins with theory (1), on the basis of which a model is built (2). Consequently, the model is "treated", which is to say that parameter values, etc., are fixed (3). The model is implemented in a solver (4), which yields results (5). Winsberg seems to think that this picture accords well with the distinction between verification and validation because validation can be associated with model construction (step 2, maybe also step 3), while verification is connected to the construction of the solver (step 4). Winsberg (2018b, p. 156) claims further that the linear picture leads to two claims, viz. that verification and validation are strictly separable activities and that they belong to different disciplines, viz. mathematics (verification), and empirical science (validation). Winsberg argues that both claims are false (Winsberg 2018b, p. 157). He also rejects the linear model (Winsberg 2018b, p. 158). All this is not meant to deny that the concepts of verification and validation differ and that some activities carried out by simulation scientists can be classified as either belonging to verification or to validation (Winsberg 2018b, p. 157).[17]

To summarize, Winsberg's main claims are as follows:

W1 In practice, verification and validation cannot always be separated in a clean way.

W2 It is false that verification is a mere matter of mathematics and validation a mere matter of empirical science.

We will presently discuss both claims in the next subsections. As a preliminary, we should note that Winsberg, in his discussion, does not distinguish between the conceptual and the computational model; accordingly, he does not discriminate between the validation of the conceptual and the computational model. This is problematic since the difference between both types of model is decisive for the discussion of V & V. Further, in characterizing the linear picture, Winsberg associates validation with model construction, and he seems to think that, under the linear picture, the validation of the model has to be independent of the results of the computer simulation (this is suggested by a remark in Winsberg 2010, p. 20). This is a very strange view of validation indeed, and Winsberg is right in rejecting it. As we have seen above, at least some details of the models often lack independent validity, so the only way is to implement them in simulations and to test the outputs with measured data. This is a central part of validating a conceptual and a computational model.

[16] I prefer the name "linear picture" to Winsberg's "linear model", because we are already concerned with a lot of models.

[17] My account of Winsberg's view is mainly based upon his 2018b, because this seems the most mature expression of it. But there is no evidence for any significant shift in his view anyway. Strictly speaking, the account in Winsberg (2018b) is restricted to climate science, but as the earlier statements of the view show, the latter is meant to apply more broadly. Note though that Winsberg (2018b, p. 162) admits that the clear separation between verification and validation may apply in certain domains of engineering.

42.4.1 The Separation Between Verification and Validation

Turn now to Winsberg's first claim. The argument in favor of it is intimately connected with Winsberg's discussion of the linear picture. Winsberg's main reason to reject this picture is that, in disciplines such as climate science, the model and the computer simulation program are often modified because its results initially do not match measured data. This leads to what Winsberg calls "life cycle" of a simulation (Winsberg 2018b, p. 159). The idea is basically to create a loop that moves back to the model and its construction, once a mismatch between simulation output and measured data has been obtained.

Winsberg is right to stress that models may be changed when they yield results that do not match the data. It's noteworthy, however, that, by moving to the cyclical picture, Winsberg discusses applications of computer simulations that are quite different from those considered by, e.g., Oberkampf and Roy (2010). Winsberg (2010, p. 23) is very clear that he is talking about simulations in which the models lack prior justification; crucial model assumptions have rather the status of educated guesses. In such examples, it seems apt to say that the scientists are in the business of hypothesizing. We cannot expect to obtain predictions with high accuracy from this business. Oberkampf and Roy (2010), by contrast, are interested in predictions with high accuracy. This explains why Winsberg's view differs from those held by Oberkampf and Roy.

The focus of Winsberg's discussion is not entirely unproblematic. For if the model assumptions behind a computer simulation are to some large extent mere hypotheses, then we should not say that the simulations have a real-world system and its behavior as target; rather we should say that they are about hypothetical systems. But then validation is less of an issue. Of course, the model assumptions may be incrementally confirmed, when they lead to increasingly more accurate predictions. In this way, we may say that the final model has been validated to some extent, and this is clearly what Winsberg hopes for (Winsberg 2018b, p. 158). However, as far as validation is concerned, it is unclear whether the first iterations of the cycle involve any proper validation, simply because the model is built on so many hypotheses that it is unclear whether it has a real target—But this is certainly not a knock-down argument, and, in what follows, I will not assume that it is successful.

Suppose then that verification and validation are in principle applicable in the life cycle of simulations. It is not clear what this means for the separation between verification and validation. In Winsberg's description of the life cycle (Winsberg 2018a, b, p. 158), there is no mentioning of activities of verification at all. The question then is whether the separation between verification and validation can be squared with the cyclical picture.

The answer is obviously yes. Once a conceptual model has been changed, the simulation can be verified with respect to it, and researchers can try again to validate the model—and in fact, both things should be done. Therefore, both activities should be added to the cycle to obtain a more comprehensive picture of simulation (and verification should be put before the comparison between simulation output and measured data). If the modification of the model is very minor, for instance, if no more than the value of one parameter has been changed, then most part of verification may not be necessary as a matter of fact because the simulation program has been verified more broadly for a whole range of parameter values before. Consequently, no new round of verification may be necessary. Likewise, to the extent that validation of the conceptual model was based upon prior knowledge of some model assumptions that have not been changed, validation need not be repeated. What is only an issue in such a case then is the question of whether the new parameter value is consistent with, or even implied by, measured data and prior knowledge. However, if there are more significant modifications, then verification and validation are necessary again.

Now the fact that both methods appear in a cycle does not show that they cannot be separated in practice. If there are no prior reasons to doubt that verification and validation can be separated, then the cyclical picture does not give us reasons to worry about the separation. True, in the cyclical picture, in some sense, verification does not precede the comparison between simulation output and measured data, as is demanded by the two-step view, because verification of model version M2 comes after the (in)validation of model version M1 using data. However, at closer analysis, there is no substantial disagreement with the two-step view, because the latter is about one single model variant. In fact, Oberkampf and Roy (2010, p. 60), who hold the two-step view, stress clearly that a realistic exercise in computer simulation may be iterative. This shows that the cycle as such is a red herring in the discussion. The real issue is whether verification and validation can be separated within one cycle.

Winsberg suggests that this is not so using the following argument (Winsberg 2018b, pp. 159–160; see also Winsberg 2010, p. 24). Sometimes, when a computer simulation lacks the intended level of accuracy because certain approximation schemes do not work, scientists change the model assumptions. A famous example is the so-called Arakawa trick (see Küppers and Lenhard 2005 for a description and a philosophical discussion). Very roughly, scientist Arakawa changed the basic equations to be solved in a simulation about the weather in order to avoid the consequences of numerical instability. In this context, he introduced a certain assumption of energy conservation, which was supposed to be unrealistic in view of the target system. But implementing the trick was considered to be a success because the predictive capacity of the simulation program was enhanced. Winsberg claims that this practice renders the distinction between validation and verification meaningless.

In his Chap. 39 in this volume, Lenhard argues likewise that computer simulation scientists often engage in kludging, which is to say that they apply ad hoc fixes to enhance the accuracy of the output. If kludging is indeed common, as Lenhard thinks is the case, then the point made by Winsberg generalizes. In Lenhard's terms, the problem is that kludges obfuscate a classification of the parameters used in a simulation model. The classification distinguishes between, e.g., parameters that reflect measurable properties, parameters that arise in certain approximation schemes, parameters that are used to characterize uncertainties and so on (see Oberkampf and Roy 2010, p. 623 for the classification to which Lenhard refers). Lenhard's point is that it becomes unclear to which type a parameter belongs if researchers have used kludges to modify the simulation. Consider the example of the Arakawa trick. It is unclear whether a parameter that is introduced in this context has a physical meaning or whether it is simply a parameter in an approximation scheme. Also, if kludges have been applied and if parameters in the program are calibrated in order to improve model accuracy, then the classification of these parameters becomes unclear because their values have been adjusted in a big network of assumptions some of which have been introduced ad hoc.

There is, in fact, a problem with the distinction between verification and validation in examples such as the Arakawa trick, but the consequences to be drawn from this are less significant than one might think. The basic problem with the Arakawa trick and other kludges is as follows: It is unclear what exactly the conceptual and the computational models are after the trick has been applied. Under one description of the new simulation, the conceptual model has been changed because a term has been added to its equations. But this description is problematic because the added term is not motivated on scientific grounds. Under an alternative description, the conceptual model is the same as before, and the new term is supposed to be part of a new approximation scheme used to solve the model equations. So only the computational model has been changed. But this description seems problematic too because it seems very natural to say that the basic equations of the model have been changed (e.g., that a certain assumption about energy conservation has been changed). Now unless we settle for a decision on what exactly the new conceptual model is, we cannot clearly tell what exactly verification and validation of the conceptual model have to show. For instance, if the idea is that the conceptual model was not changed by the trick, then verification should show that the trick improves accuracy of the simulation with respect to the conceptual model. By contrast, if the model was changed, then this is not an issue. In this sense, the distinction between verification and validation becomes difficult. But more precisely, we should say that the distinction is *relative* to an identification of the conceptual and the computational model. This does not imply that the distinction is unclear or that verification and validation cannot be separated.

It is further arguable that scientists should at some point settle on what they regard to be the conceptual and the computational model. Otherwise, one may say, they do not really have an understanding of what they are doing. To settle this question, it seems, they should find out whether or not the implementation of the new equations produces just more accurate solutions to the original equations. If the former is the case, they should not change their conceptual model, but rather think that it is to some extent incrementally confirmed by the fact that the simulation output agrees with the measured data, after the trick has been applied. If the modified simulation program does not yield more accurate solutions to the original conceptual model, scientists should become skeptical about their original conceptual model and think more closely about the assumption of energy conservation, for instance. This suggests that a clear distinction between the conceptual and the computational model is a desideratum of good scientific practice. This would mean that unclarity about the conceptual model is an intermediate stage and thus not significant.

Winsberg's discussion suggests another argument in favor of W1. When a computer simulation yields results that do not match the measured data, this may either be due to a lack of verification or a lack of validation. That is, the problem may either be that the approximation scheme does not work as intended, etc., or that the model is inappropriate (we have already noted this in Sect. 42.3.1). Conversely, if a simulation program does yield the desired accuracy, we may count this as (a small) incremental confirmation of the verification and the validation of the model. This "degeneracy" suggests that validation and verification are entangled with each other (see e.g., Winsberg 2010, p. 24). But the first case (mismatch between simulation output and measured data) does not really show that verification and validation cannot be separated. It only shows that we can sometimes not discern whether verification or validation has failed. The second case (sufficient fit between simulation output and measured data) shows that a certain agreement between simulation output and measured data can make a contribution to both verification and validation. But this contribution is small, because the coincidence may only arise because approximation and modeling errors cancel each other. It is different (cf. Sect. 42.3), if we have some prior reasons to think that the simulation can be verified and validated. Further, that certain coincidences can confirm verification and validation of a model, or contribute to both, is something that we expect on general action theoretic grounds. As argued above, methods and action types are not exclusive categories.

All in all, the best arguments that Winsberg and Lenhard give for W1 show only that the distinction between verification and validation is relative to a distinction between the conceptual and the computational models. Once the two types of models are identified, we can separate between verification and validation. Further, a good understanding of why a simulation program works requires that a conceptual model be clearly identified.

42.4.2 Verification and Mathematics

Turn now to W2, viz., the claim that a clear assignment of verification to mathematics and of validation to the empirical sciences is wrong. Our discussion can be very brief.

Winsberg's main argument in favor of W2 seems to be that the mathematical means to achieve verification are very weak. For instance, he writes: "When models are sufficiently complex and nonlinear, it is rarely possible to offer mathematical arguments that show, with any degree of force, that verification is being achieved" (Winsberg 2010, p. 23). This suggests that we need to some extent rely on validation to make sure that the model equations are appropriately solved in the simulations (Winsberg 2010, pp. 23–24).

Whether this argument is correct turns on what we mean by mathematical methods or means to verify a simulation. As the last quote from Winsberg shows, he concentrates on arguments. He contrasts this with methods that use the output from simulation programs and compare them to benchmarks (Winsberg 2010, p. 22). As far as this comparison with benchmarks is concerned, Winsberg does not clearly distinguish between, e.g., analytical solutions to the model equations and empirical data. But there is a difference, and what is important for verification is a comparison between simulation and, e.g., analytical solutions (see e.g., Oberkampf and Roy

2010, p. 33). The question then is how this method should be classified. Winsberg seems to think that it is empirical, and thus not mathematical. Now the method is in some sense empirical, because it relies on simulation output that has been produced using a device from the empirical world. But it is uncontroversial that verification is empirical in this sense. For instance, Oberkampf and Roy (2010, p. 33) write: "Numerical algorithm verification is fundamentally empirical". Since all parties agree that verification is empirical in this sense, we can put this sense aside. The question then is whether verification in the sense of a comparison between simulation output and benchmark solutions can still be called mathematical. This seems appropriate in view of the fact that the benchmarks are analytical results. As such they are obtained using mathematics. They are non-empirical because they are not based on data from the target system.

In sum then, Winsberg is right about W2 if mathematical arguments are understood in a very narrow sense that excludes the use of any simulation output. Under a broader reading of "mathematical argumentation" that is quite natural, however, the basic techniques of verification are mathematical.[18]

42.5 Conclusions

To summarize our main results: Verification and validation of computer simulations are clearly two different methods because they have different aims. While verification is supposed to show that the computational model implemented in a simulation program represents the conceptual model in some respects to some accuracy, validation is supposed to show that a model represents its target system in some respects to some accuracy. As far as computer simulations are concerned, we need to differentiate between the validation of the conceptual and that of the computational model. This leads to a picture with a circle and three nodes, viz., the conceptual model, the computational model, and reality. There are three ways of relating two of the three nodes to each other, and these correspond to verification (conceptual model vs. computational model), validation of the conceptual model (conceptual model vs. reality) and validation of the computational model (computational model vs. reality).

As this picture indicates, the methods are closely related. First, if a conceptual model is validated independently from a simulation and if the computational model is verified with respect to the conceptual model, then the computational model is validated in comparison to its target system. We can call this validation of the computational model if we do not require that every type of validation involve the direct comparison between simulation output and measured data (note that some authors require this comparison for validation). In practice, however, it is often not possible to argue for the accuracy of the computational model in this way because the conceptual model cannot fully be validated independently from the computer simulation.

[18] As a side remark, we should note that Winsberg does not discuss a very powerful new method of verification, viz., the method of manufactured solutions (see Chap. 12 by Roache in this volume).

In this situation, validation of the computational model and the conceptual model go often hand-in-hand as follows: Verification ensures that both models can be treated as roughly equivalent (in some respect; we will drop this qualification in what follows, although it is meant to apply). Simulation output is compared to measured data from the target system for a limited set of data points. If they agree to sufficient accuracy, then we can conclude that 1. The computational model; 2. The conceptual model represents reality (or the target system, more specifically) with sufficient accuracy, which is to say that both models have been validated. I have argued that this inference from a limited set of data points to the accuracy of the models is only legitimate if there is at least some case that the conceptual model is valid and the computational model is verified. The first condition is typically justified if the conceptual model is built on enough prior knowledge about the target system.

This reconstruction of validation shows that verification is to some extent needed for the data-driven validation of the computational model and utterly necessary for validating a conceptual model using data. This suggests that verification precedes the comparison between simulation output and measured data. Depending on how exactly we define validation in detail, we may either say that the verification is a necessary step within validation or a precondition for validation.

As Winsberg reminds us, verification and validation are often embedded in life cycles of simulation models. It is also correct that the distinction between verification and validation presupposes a clear distinction between the conceptual and the computational model (although Winsberg himself does not make this distinction). In practice, this distinction is sometimes unclear, as the example of the Arakawa trick shows. But a clear identification of the conceptual model seems to be a long-term desideratum. This is at least so if computer simulation is to be more than mere data fitting, but rather the attempt to understand a target system with a model that is based upon prior scientific knowledge about the target system. All in all, Winsberg's argument does not show that the distinction has been over-exaggerated.

Acknowledgements I thank David Murray-Smith and Nicole J. Saam for useful comments.

References

AIAA. (1998). Guide for the verification and validation of computational fluid dynamics simulations, AIAA G-077-1998. American Institute of Aeronautics and Astronautics, Reston, VA, 1998.

Anscombe, G. E. M. (2000). *Intention*. Cambridge, Mass.: Harvard University Press (first edition 1957).

ASME. (2006). Guide for verification and validation in computational solid mechanics. American Society of Mechanical Engineers, ASME V&V 10-2006.

Harman, G. H. (1965). The inference to the best explanation. *Philosophical Review, 74*(1), 88–95.

Küppers, G., & Lenhard, J. (2005). Computersimulationen: Modellierungen zweiter Ordnung. *Journal for General Philosophy of Science, 36*(2), 305–329.

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science, 70,* 647–670.

Oberkampf, W., & Roy, C. (2010). Verication and validation in scientific computing. Cambridge University Press.

Oberkampf, W. L., & Trucano, T. G. (2008). Verification and validation benchmarks. *Nuclear Engineering and Design, 238*(3), 716–743.

Pace, D. K. (2004). Modeling and simulation verification and validation challenges. *Johns Hopkins APL Technical Digest, 25*(2), 163–172.

Parker, W. S. (2008). Franklin, holmes, and the epistemology of computer simulation. *International Studies in the Philosophy of Science, 22*(2), 165–183.

Roache, P. (1997). Quantification of uncertainty in computational fluid dynamics. *Annual Review of Fluid Mechanics, 29,* 123–160.

Roy, C. J. (2005). Review of code and solution verification procedures for computational simulation. *Journal of Computational Physics, 205*(1), 131–156.

Sargent, R. G. (1984). A tutorial on verification and validation of simulation models. In: S. Sheppard, U. Pooch, & D. Pedgen (Eds.), *Proceedings of the 16th conference on Winter simulation* (pp. 115–121). IEEE Press.

Schlesinger, S., et al. (1979) Terminology for model credibility. *Simulation, 32*, 103–104.

Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science, 71*, 767–779

Weisberg, M. (2007). Who is a Modeler? *British Journal for Philosophy of Science, 58,* 207–233.

Winsberg, E. (2010). *Science in the age of computer simulation*. Chicago: University of Chicago Press.

Winsberg, E. (2018a). Computer simulations in science. In E. N. Zalta, (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition). https://plato.stanford.edu/archives/sum2018/entries/simulations-science/.

Winsberg, E. (2018b). *Philosophy of climate science*. Cambridge: Cambridge University Press.

Chapter 43
The Multidimensional Epistemology of Computer Simulations: Novel Issues and the Need to Avoid the Drunkard's Search Fallacy

Cyrille Imbert

Abstract Computers have transformed science and help to extend the boundaries of human knowledge. However, does the validation and diffusion of results of computational inquiries and computer simulations call for a novel epistemological analysis? I discuss how the notion of novelty should be cashed out to investigate this issue meaningfully and argue that a consequentialist framework similar to the one used by Goldman to develop social epistemology can be helpful at this point. I highlight computational, mathematical, representational, and social stages on which the validity of simulation-based belief-generating processes hinges, and emphasize that their epistemic impact depends on the scientific practices that scientists adopt at these different stages. I further argue that epistemologists cannot ignore these partially novel issues and conclude that the epistemology of computational inquiries needs to go beyond that of models and scientific representations and has cognitive, social, and in the present case computational, dimensions.

Keywords Computer simulations · Novelty · Epistemology · Validation · Goldman · Consequentialism · Social epistemology · Random numbers · Computational values · Monte Carlo · Random number generator · Modeling norms · Modularity · Opacity · Verification · Reproducibility · Models · Invisibleness of failure · Naturalism

The drunkard's search or streetlight fallacy corresponds to a type of situation where people search at the easiest site, even if what they are searching for is unlikely to be there. Typically, the drunkard searches for her keys under a streetlight even if they were lost somewhere else.

C. Imbert (✉)
Archives Poincaré, CNRS, Université de Lorraine, 91 avenue de la Libération - BP 454., F-54001
NANCY Cedex, France
e-mail: Cyrille.Imbert@univ-lorraine.fr

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_43

43.1 Introduction: Computer Simulations, a Revolutionary Epistemology?

The search for innovation and novelty are major goals across social fields. Unsurprisingly, when new technologies, artifacts, techniques, methods, practices, perspectives, or issues are developed, bold statements about their potential impacts are made. So-called "revolutions" or "turns" are regularly announced across science and it is legitimate to be methodologically cautious about such claims.

It is hardly controversial that computers have largely transformed science and help to extend the boundaries of human knowledge. Yet, it might be the case that computers merely bring about more inferential power but that concerning how scientific results are justified and come to be trusted, science is left unchanged by the computational revolution.

Philosophers of science have had a long-standing tradition of analyzing experiments, theories, and scientific reasoning. However, specific epistemological analyses of simulations did not develop until the 1990s with work by philosophers like Paul Humphreys, Eric Winsberg, or Manfred Stöckler, historians of science like Peter Galison, scientists interested in philosophical issues like Fritz Rohrlich, or scholars at the crossroads of several fields like Evelyn Fox Keller. These different authors mostly agreed that computational methods not only provided a new powerful way to practice science, but also did not match existing categories and called for specific and novel analyses, above and beyond those concerning experiments, theories, or models.

In a thought-provoking and conservative article, Roman Frigg and Julian Reiss stood against this move and argued that claims about the novelty of computational science were overblown and ill-grounded and that there was no more to the epistemology of simulations than the epistemology of modeling (Frigg and Reiss 2009). Making final and flawless contributions is difficult for those who pioneer in a field and various aspects of Frigg and Reiss's jubilant refutation were convincing. The need to guard against the lure of apparent novelties was later confirmed, for example, by the criticism by Barberousse and Imbert (2013) of revolutionary claims about cellular automata based simulations (a case that was recurrently used in favor of novelty claims) or by the sober and deflationary analysis by Beisbart of the deeply argumentative nature of simulations despite their genuine similarities with experiments (Beisbart 2018, see also Barberousse et al. 2009). However, Frigg and Reiss were not content to refute claims about the novelty of specific aspects of simulations. They extrapolated that simulations "raise few if any new philosophical problems" (593) and suggested considering the literature on simulations "as contributing to existing debates about, among others, scientific modeling, idealization or external validity, rather than as exploring completely new and uncharted territory" (595).

Paul Humphreys quickly responded that this general non-novelty claim was simply false. In (Humphreys 2009), he highlighted that issues such as the epistemic opacity of computational processes, the importance of syntax, complexity questions, or the specific role of time in simulations all make the epistemological and semantic analysis of computational science novel, beyond genuine overlaps with existing philosophical analyses of science.

The present chapter focuses specifically on the epistemology of simulations, how their results are validated, and whether the problems that arise in this context are novel. Frigg and Reiss' debunking paper was a sanitizing contribution. Nevertheless, I will also argue that their main conclusion is false because simulations raise new epistemological questions or raise traditional questions that require novel or specific answers for simulations.

I devote Sect. 43.2 to philosophical preliminaries: I first discuss how the notion of novelty should be cashed out here and argue that using a conceptual framework similar to the one used by Goldman to develop social epistemology is appropriate for the investigation of the present question. In Sect. 43.3, I list computational, mathematical, representational, and social loci on which the validity of simulation-based belief-generating processes hinges. Additionally, I emphasize that their epistemic impact depends on the practices that scientists adopt to face these problems. I further argue in Sect. 43.4 that epistemologists cannot ignore these issues and conclude in Sect. 43.5 that this analysis agrees with those which emphasize that the epistemology of science needs to go beyond that of scientific representations and has cognitive, social, and here computational, dimensions.

43.2 Methodological and Conceptual Preliminaries

First, the notion of novelty should be clarified if it is to frame the discussion. What is scientifically novel is contingent upon which claims, theories, or perspectives have been defended within a field. Thus, the real issue is whether an object of inquiry should be analyzed along the same lines as other objects. The difference can be illustrated as follows. In the context of computer simulations, it is blatantly obvious that complexity and computational resources must be taken into account to analyze the constraints that frame computational inquiries. Accordingly, focusing on what is possible in practice (Simon 1957; Humphreys 2004; Wimsatt 2007) and emphasizing the importance of the scarcity of resources for agents is appropriate. However, resource-boundedness is a general constraint that frames *both* computational and noncomputational inquiries. Humphreys suggests the general epistemological principles that "it is the invention and deployment of tractable mathematics that drives much progress in the physical sciences" and that "most scientific models are specifically tailored to fit, and hence are constrained by the available mathematics" (see Humphreys 2004, 55–56 and Barberousse and Imbert 2014 for a detailed discussion). Still, it so happens that in existing discussions about models, complexity issues merely arise as a peripheral point to justify the need to make approximations. In brief, whereas the development of computational science somewhat relaxes computational constraints and resource-boundedness constitutes a much more restrictive straightjacket for traditional noncomputational inquiries (see again Barberousse and

Imbert 2014), quite paradoxically, the need to adopt a bounded-resource perspective is blatant and apparently novel in discussions about computational inquiries.

Also, the notion of the novelty of questions, discussions, or of "uncharted territory", partly places novelty in the wrong location. While some aspects of computer simulations trigger new questions (e.g., concerning code or the epistemic opacity of computational processes), others raise questions of types that are already analyzed by epistemologists but need no less epistemologically revolutionary answers. For example, the role of human faculties in the architecture of human knowledge is a central issue in mainstream epistemology. This role sometimes changes. Over the centuries, the development of measurement instruments transformed empirical science and made it less dependent on our senses. However, till the advent of computers, methods, languages but also objects of inquiries were adapted to human reasoning and inferential abilities, the reliability of which was crucial to that of scientific results. The development of computer-assisted science keeps transforming this situation (Humphreys 2009, 616). Computers carry out increasing parts of inferential processes and scientific inquiries are less adapted to our inferential capacities in their objects and methods. However, humans remain the architects of these inquiries, the devisers and warrantors of methods and instruments, and the recipients of scientific results. In brief, science is no longer *human-tailored* but remains *human-centered* (Imbert 2017, 771), and we are faced with the "*anthropocentric predicament*, of how we, as humans, can understand and evaluate computationally based scientific methods that transcend our abilities" (Humphreys 2009, 617). Overall, the development of computational science requires reexamining the place of human faculties in knowledge and analyzing the evolving distribution of roles between human capacities and the epistemic instruments that we use as surrogates.

Second, a conceptual framework is needed to investigate the scope of this epistemological inquiry about the epistemological novelty of simulations and their validation. General epistemology analyzes issues such as the nature, sources, or structure of knowledge and justified belief. Specific, applied epistemological inquiries can be pursued about the specific practices or processes through which beliefs are acquired within fields for which the promotion of epistemic objectives is important, such as adjudication, library science, journalism, or science. For example, the epistemology of science investigates how novel scientific contents are unraveled by processes such as mental reasoning, calculus, thought-experiments, experiments, or computer simulations (El Skaf and Imbert 2013).

A consequentialist framework like the one used by Goldman for social epistemology (Goldman 1999, 87) provides a useful tool to analyze the various aspects of such belief-generating processes. The inquiry is pursued relative to some epistemic states, such as knowledge, error, ignorance, or consensus, which are considered to have primary or fundamental value. Then, practices can be described as having instrumental value depending on how much they promote or impede the development of such epistemic states. Further, it is useful to adopt a fine-grained description of the resulting epistemic states. Goldman proposes the notion of "mental infosphere" at a time *t*, which consists of the beliefs of all the people inhabiting the globe at *t* (Goldman 1999, 161). Then, this conceptual framework is used "to widen epistemology's vista" (ibidem, preface) and to show how particular communication systems, adjudication rules, media funding systems, testimonial rules, etc., have a positive or negative impact on the dissemination of false or true beliefs in the mental infosphere. Similarly, one can analyze how much the practices involved in belief-generating processes that comprise computer simulations favor the development of true, error-free, consensual, reliable, etc., beliefs in the mental infosphere, or in its scientific part.

Belief-generating processes involving computer simulations require the expertise of specific scientists, organized in specific ways, dedicated to specific practices, and using specific tools, languages, or types of resource. Thus, trivially, their epistemological appraisal corresponds to a task in its own right since it requires analyzing and understanding new types of objects and processes. Further, because inquiries involving computational methods represent a large part of scientific activities, this task is an important one in scientific epistemology.

However, the importance and specificity of this *task* do not imply that, once epistemologists tackle it, they are always faced with novel *problems*. Nevertheless, could it really be the case that the computational, logical, mathematical, cognitive, and social specificities of computer simulations do not make a single epistemological difference and that all the epistemological problems that they raise boil down to problems that epistemologists have already solved in different contexts? If this is so, applied epistemologists should celebrate this cosmic coincidence and rejoice that their past works have such unintended scope.

In any case, even if simulations reveal epistemological problems that are similar or identical to those raised by other scientific activities, they can still be epistemologically different. Indeed, answering a traditional question about a novel object is not usually trivial. When one tries to solve equations of a new type, one tackles a novel problem. Saying that a mathematician who has successfully done so has achieved nothing new, because this is (once again!) the same old stew or problem of solving an equation would be mathematically naive. Naturally, it may be that answers to epistemological questions about simulations are sometimes identical to answers to similar questions about other activities, but it cannot be *assumed* that this will be systematically so. Finally, showing that a problem about some type of object is actually the same as another problem about another type of object is usually not simple; and showing that a problem reduces to another, or that the solution of the latter can be adapted to solve the former, is usually an achievement.

Overall, it is difficult to tell in advance how much the epistemology of computer simulation shares with that of other activities. Computer simulations have the same target as other scientific activities, rely partly on the same theoretical material, use common parts of applied mathematics, and are partly carried out by agents that are subject to similar cognitive, scientific or social constraints. Like experiments, they can be part of "big science", often involve using material and nonmaterial instruments, massive budgets, various collaborators, and may yield big data. Because all such features are not epistemologically neutral, the epistemology of computer simulations cannot be radically new, nor should it be carried out separately from epistemological inquiries about instruments, mathematics, computer science, statistics, experiments, and, naturally, scientific representations and models (see also Frigg and Reiss 2009, 611, Humphreys 2009, 615). In brief, there is no doubt that the epistemology of simulations has a lot in common with that of other scientific activities. Identifying genuine overlaps, disentangling and explaining shallow similarities from deep ones, and determining what is epistemologically specific to computer simulations strictly speaking and what is a general feature of computational science corresponds to a research program for applied epistemologists (for a critical overview of the case of simulations and experiments, see e.g., Imbert 2017, 34.5) In any case, claiming from the start that nothing novel is to be found in the epistemology of simulations remains puzzling. It is as if Columbus, after only one month in America, had claimed that there was nothing specifically new or interesting on this continent because local indwellers also had two legs and no road panel pointing at hot discoveries was in sight for newcomers.

For the following discussion, I adopt the following characterization of computer simulations (see Imbert 2017, 34.2.1 for more details):

> A computer simulation corresponds to the actual use of a computer to unfold the behavior of a physical system S, by generating a description of a potential trajectory of S in the state space of a computational model of this system by applying repeatedly an algorithm that computes the description of the next state of the trajectory from the description of the previous states.

Analyzing how the conclusions that are reached with the help of computer simulations can be validated requires discussing more than computer simulations per se. Computer simulations are embedded within larger scientific inquiries aimed to answer specific questions about certain target systems (El Skaf and Imbert 2013, 3454, Frigg and Reiss 2009, 596). At the end of the day, the key issue is not whether computer simulations faithfully represent some target systems but whether the data that they yield can be used to provide target questions with answers that are likely to be correct. Below, I shall consider that validation describes the process of making sure that this is the case. Validation in this sense is directed at inquiries and investigates the soundness of the production and use of computational results. As for any form of reasoning, the value of the final results hinges on both the content of the material that feeds them (typically premises, theories, or models) and whether the inferential process (which here includes the running of the simulation) is properly carried out. As we shall see, the process of validating simulation-based inquiries goes beyond the adoption of good scientific methods and has a social or communal dimension.

43.3 Dimensions of Computational Inquiries, or Where Things Can Go Wrong Epistemically

Belief-generating processes relying on simulation-based inquiries are extremely complex, from the elaboration and running of computer simulations to the reception of results in scientific communities and beyond. An important task for epistemologists is to pin down within these processes the various problems that scientists must solve for the final results to be valid, and where detrimental effects can be triggered and spoil the process.

Philosophers of science do not have a strong tradition of investigating scientific or epistemic failure. Inquiries about errors and practices are mostly carried out by sociologists of science (e.g., see the symmetry thesis about the explanation of false and true results in Bloor 1976), philosophers with specific orientations, such as naturalists, pragmatists, or advocates of a practice turn (see e.g., Kitcher 1993, Wimsatt 2007, Woods 2013), or philosophers investigating issues for which the question of errors can hardly be discarded, such as investigations about ampliative reasoning or statistics. Still, it is informational to locate within belief-generating processes the key factors (or "process variables") that influence the validity of results or on which the epistemic impact of the results hinges, if only to better control these epistemic processes. Another task is to analyze specifically the epistemic impact of the adoption at these hinge points of particular choices, behaviors, practices, or policies. When the corresponding epistemic effects are specific to the context of computer-based inquiries, or when such key factors are specific to such inquiries, the validation of simulation-based inquiries is a novel problem, which calls for specific epistemological analyses. I illustrate in the following paragraphs such cases.

43.3.1 The Production of Computational Results: Can We Control the Beast?

Because computer simulations involve carrying out a wide variety of tasks, they can fail in many, often specific, ways. Errors may come from the hardware, e.g., if single-bit alterations are caused by physical interferences. Failures may be rooted in the types of miscomputations or malfunctions that can affect digital computers and communication systems (Fresco and Primiero 2013), in particular in the written code or the software, which do not always do what we believe they do. Problems may come from the type of algorithms that we use to compute functions, to approximate real functions, or to solve equations, but also from their implementation; from the discretization of mathematical objects to make them amenable to computable descriptions; from a mismatch between the algorithms (or the models) and the type of computational architecture that is used (supercomputers now use parallel architectures, which require various adaptations), etc.

I do not presume to present here an exhaustive list; quite the contrary. Inventorying and analyzing specific ways of failing, from the hardware to the inquiry level, and assessing the factors that favor or neutralize them is a substantial epistemological task. Potentially, it requires a wealth of distinct expertise concerning various parts of the process, which clearly makes the situation epistemically uncomfortable for epistemologists. Still, this is no reason to discard or ignore this task, which is no less important than the analysis of potential sources of errors in other belief-generating processes, such as cognitive biases and fallacies of reasoning, logical errors in formal inferences, typical ways of failing in thought-experiments, or bandwagon effects and cascades in belief exchanges within communities. I highlight below aspects of the production of computational inquiries that make their validation specific.

43.3.1.1 Computational Practices: An Evolving Field with Specific Epistemic Values

The success of computational science depends first on the ability of scientists to develop specific technical, mathematical, and human strategies to solve hardware and software problems effectively (for questions pertaining to verification, see Sect. 43.3.2 below and Chap. 10 of this volume by Rider).

These questions cannot be discarded or identified by an armchair inquiry, since the issues of where reliability bottlenecks lie, how they are faced, and which questions are hot concerning computer simulations and their validation keep evolving with technological and scientific progress. For example, for the first computers, “the overwhelming problem was to get and keep the machine in working order,” as is reflected in the names of societies such as the Association for Computing Machinery (Dijkstra 1972, 860). Also, the existence of single-bit alterations may no longer be a worry for ordinary simulations, even if it remains an issue for sensitive simulations, for which error-correcting code memory (ECC memory) devoted to scientific computing needs to be used. Similarly, in the late 1960s, the development of computer power had triggered needs that could not be answered by programmers’ abilities. “Programming ha<d> become an equally gigantic problem” (ibidem, 860), and “the software crisis” had arisen, with the development of low quality, inefficient, or difficult to maintain software. How scientists have managed this crisis ever since is a question worth exploring.

Anyhow, beyond discussions about the validity of particular inquiries, the *average* validity and *global* impact of computational inquiries depend on various properties of hardware and software. How much hardware and software is globally *efficient, easily usable, standardized, maintainable, adaptable* for follow-up inquiries, *transferable* to other scientific problems, etc., influences how much sound results are produced. These properties correspond to epistemic values that are specific to computational inquiries, so investigations about their impact are clearly novel. For example, a science in which all codes are radically different and all practitioners develop their specific solutions is unlikely to be efficient and globally reliable. In contrast, the existence of shared codes of good practices and commonly developed software tools, or the adoption of common standards (e.g., in terms of hardware, programming languages, or mathematical tools) is bound to have positive effects. This shows that the epistemology of computational inquiries, like that of instruments, goes beyond that of individual practices and overlaps with social epistemology. How much the above properties need to be traded against one another and where actual computational practices within empirical science lie on this multidimensional map are other questions worth investigating.

43.3.1.2 Applied and Computational Mathematics for Limited Social Agents: The Case of Random Numbers

To carry out computer simulations, scientists need to find ways of solving various mathematical problems. Understanding how they do so requires going beyond traditional questions in the epistemology of mathematics such as how we interact with mathematical entities, make reference to them, or access mathematical truths, since the epistemological issue of how we *develop* mathematical knowledge remains largely untouched by answers to these foundational questions. By contrast, the epistemology of applied mathematics and computational science deals directly with the issue of how logical and mathematical content is *unfolded*, knowledge *extracted*, and problems *solved* given our limited wherewithal, the complexity of the task, and the features of the formal tools that we use (see e.g., Wimsatt 2007, El Skaf and Imbert 2013, Fillion and Corless 2014, Lenhard and Carrier 2017). Accordingly, it involves analyzing how heuristics for mathematical problems work and what we can expect from them; how to describe the quality of approximate solutions, how to develop mathematical strategies to analyze and control computational errors and, more generally, which features influence how applied mathematicians crawl their way through complex problems. While applied mathematics is not limited to its use in computational science, it is central to this field, and much of it is developed for the needs of computational inquiries.

For illustrative purposes, I now present the case of the production of random numbers by simulation practitioners and highlight factors on which the reliability of this task depends. The production of random numbers is a central problem of modern science. Randomness is a key concept across various theories, and fields and scientific arguments involving statements about random properties are frequent. A specificity of computer simulations is that they often rely on the use of token numbers that *instantiate* the property of randomness (*versus* involve statements that attribute it). Accordingly, the validity of statements *about* random properties merely relies on the semantic relation between the content of these statements and what they denote. By contrast, that of computer simulations and computational inquiries using random numbers also relies on our ability to produce such random numbers. Various epistemological questions arise in this context. How easy is it to produce the random numbers that our computer simulations need? Which factors have an impact on this production? Can we expect the random numbers that are usually used in computer simulations to be good enough for the preservation of the validity of inquiries? I provide evidence that these questions cannot be ignored and have nontrivial answers that require going into the details of socio-computational practices, and perhaps the psychology of practitioners.

Whereas almost all sequences of binary digits are random, producing random numbers is extremely difficult. The need for scientists to produce many such numbers increases the difficulty of the task. Random number generators (hereafter RNG) must satisfy various requirements such as producing uniformly distributed, reproducible, random numbers, which have periods that are much larger than the samples used. Jointly fulfilling all these requirements is in general difficult, but even more so in the context of parallel computers (Hellekalek 1998). Parallel architectures involving thousands of processors were developed in the 1990s and supercomputers are now massively parallel. Then, to the extent that parallelization is possible, parts of the computational task can be computed synchronically, which speeds up computation. Typically, replicas for Monte Carlo simulations can be produced independently. Thus, good random generators should be parallelizable. For parallel RNG (hereafter PRNG), other requirements are the absence of correlations and, for reasons of efficiency, the need to generate numbers independently (ibidem, 85). After reviewing existing methods to fulfill these requirements, Hellekalek concluded that it was “not at all trivial to find high-quality RNGS for parallel machines” (ibidem, 82) and that some aspects of the problem (e.g., correlations between disjoint substreams of consecutive numbers) were “dangerous territory” (ibidem, 85). Indeed, his analysis showed that the application of parallelization techniques to standard RNG could “perform terribly” (ibidem, 86). Thus, his paper was named “Don’t Trust Parallel Monte Carlo”—arguably a big stone in scientists’ shoes, given the importance of parallel computers and Monte Carlo methods for computer simulations.

Naturally, things have improved since Hellekalek’s paper, but full optimism may still be inadequate. Scientists’ needs have also increased massively and access to supercomputers is difficult. Simulations in nuclear medicine can require as many as 10^{20} random numbers for computations carried out on thousands of processors. A fundamental problem is that scientists do not have theorems or techniques to prove the independence of two parallel random streams (Hill 2015, 68). Further, strong autocorrelations within pseudorandom numbers can appear far apart and spoil the application of parallelization techniques (De Matteis and Pagnutti 1988). In practice, the testing of PRNG is based on a battery of statistical tests, such as BigCrush TestU01, which represent the current state of knowledge about random numbers. and few PRNG satisfactorily pass the test. The epistemological moral is that it can require pointed expertise to determine whether a (P)RNG is sufficient for a scientific inquiry.

The next question for epistemologists is to assess whether the (P)RNG that are *actually* used in science are in general satisfactory. After all, if scientists always use the currently best RNG, troubles are unlikely, and epistemologists should not bother. Evidence can be found that epistemological optimism may be misplaced again. For a scientist without expertise about RNG, the easiest option is to use the “rand” functions from standard libraries that are used by her community. The problem is that “almost all of these generators are badly flawed” (Jones 2010), and the somewhat inconvenient advice here is to “always use <one’s > own random number generator” (ibidem). The use of dedicated libraries is no guarantee either. In 2004, Joel Heinrich, a researcher in high-energy physics, pointed out serious defects in pseudorandom generators provided by standard libraries such as Linux C and C++ as well as a major bug in CLHEP (*A Class Library for High Energy Physics*) class library for random generators, that is, tools frequently used by physicists (Heinrich 2004). Similarly, 40 out of 58 generators in the GNU scientific library were shown to have defects due to inadequate initialization schemes (Matsumoto et al. 2007). Indeed, an inadequate use of a good RNG can also spoil the broth. For example, a simple way to seed an RNG is to use the time function existing in most libraries. However, this is not acceptable if one launches too many jobs because a large number will start at the same time on different nodes. This leads to a repetition of the same calculations many times and surreptitiously spoils the statistics (Jones 2010). Unfortunately, this type of problem can be hard for practitioners and for the external community to detect. Overall, for random number production, the use of communal scientific resources does not protect against failure. Thus, because many scientists lack the relevant, evolving expertise and do not always resort to experts for such local choices, computational results may often be sullied.

Even then, reliability may be preserved by the transmission of good computing practices within communities (see e.g., Wilson et al. 2014). Thus, the average validity of computer simulations using random numbers depends on whether communities are organized in a way that favors the adoption of sound practices and indirectly promotes reliability. Social epistemologists of science should then investigate whether the right information is easily accessible and actors are incented to do the right thing. Good practice guides like that of Jones (Jones 2010) can help practitioners adopt appropriate practices or understand that they need assistance. Explicit publication standards in good journals can also point out sensitive aspects, if, for example, authors are systematically requested to provide details about the nature, properties, and implementation of the RNG that they have used. Overall, on this and other issues, the reliability of computational science is contingent on the adoption of appropriate practices at the community level and on how individual scientists tend to behave in this unsafe environment.

43.3.1.3 Changes in Modeling Practices, Justification Strategies, and Typical Usages

Based on their familiarity with particular types of simulations, some philosophers have tried to extrapolate and to single out specific features of the epistemology of simulations. For example, in early writings, cellular automata were seen as typical illustrations of the epistemological novelties brought about by computer simulations. More recently, Winsberg has suggested that the knowledge produced by simulations results from inferences that are *downward* (from theories to phenomena), *motley* (the justification process is a combination of disparate elements), and *autonomous* (see, e.g., Winsberg 2010, passim).

Unsurprisingly, such claims are easy to falsify. Simulations are versatile, mostly neutral inferential tools. Like other general tools, they can be used in various (epistemic) contexts and depending on the cases, the appropriateness of their use can be justified in different ways. In brief, simulations are epistemologically heterogeneous and the project of finding some general, novel features about how their results are justified seems doomed to fail. Because models are also versatile tools, their epistemological uses are also heterogeneous. At the end of the day, it is no surprise that epistemological features instantiated by simulation-based inquiries can also be instantiated by (pen-and-pencil) model-based inquiries. The conclusion should not be that the general epistemology of simulations boils down to that of models but rather that neither (the class of) model-based inquiries nor (the class of) simulation-based inquiries correspond to epistemological kinds that provide appropriate units of analysis for general investigations about validation strategies.

To discuss validation strategies and assess the different types of roles that simulations can play within these strategies, inquiries should be described at a fine-grained level by specifying their goals, what is known about the target systems, what tools and resources are available, etc. Then, it may be the case that simulations sometimes open up a space for novel validation strategies or, more frequently, for new versions of existing strategies. Generic modeling practices such as approximating, idealizing or abstracting are a way to use models which, though tractable and simple, produce results that suit the particular goals of inquiries. While these procedures are already analyzed in the literature on scientific models, it remains worth investigating whether these generic practices have specific versions and require particular epistemological scrutiny in the context of simulation-based inquiries. Similarly, ensemble forecasting can be seen as nothing novel since it amounts to combining different incompatible epistemic sources (here simulations) to make (predictive) judgments. However, it is usually agreed that this procedure calls for specific analyses in the context of simulations and climate analysis.

Importantly, the epistemology of science should analyze what validation strategies are used in suitably described contexts, but also how frequently these strategies are used, and why. Epistemological features like those highlighted by Winsberg, though not specific to simulations, may correspond to strategies that develop with computational science. Arguably, because computer simulations are a powerful tool, they are likely to be used in more complex and uncertain cases, which would not be investigated otherwise and which constrains the selected strategies. Then, *because of this type of use*, computational inquiries may seem, *on average*, to have specific features and an epistemology of their own and they may modify how science is usually practiced. For example, simulations may more often involve approximations and departures from the truth, epistemologically mixed or impure methods, trade-offs between epistemic goals, etc., even if, when philosophers of science analyze their aspects individually, the practices that they find are not radically different from those identified for other types of model-based inquiries (see Imbert 2017 for more details and similar analyses about unexplanatoriness and simulations). Overall, the failure to note the difference between analyses in terms of properties of token inquiries and analyses in terms of frequent features of inquiries may be another cause of dissent concerning the novelty of the epistemology of simulations.

In any case, computational science may require a specific analysis with respect to typical modeling or justification practices within communities. For example, the availability of computational power may change which modeling strategies are most often used. As noted by Frigg and Hartmann (2017, Sect. 3.1), computational power may encourage scientists "to swiftly come up with increasingly complex models." This may lead in turn to an improvement of the empirical adequacy of predictions, but not necessarily to a better understanding of underlying mechanisms. In the end, such changes may modify which goals are valued and which modeling norms are dominant within the cultures of communities using simulations. Differences of these types can hardly be analyzed by scrutinizing exclusively the content of particular representations.

43.3.1.4 Division of Scientific Labor, Computational Inquiries, and the Preservation of Validity

Various types of tasks requiring different types of expert knowledge need to be adequately carried out for the production of valid computational results. Because no single individual can possess all the relevant knowledge, scientists need to divide the global task into subtasks, and delegate their completion to specific humans or machines. Then, how much computational science can felicitously "<push> back the boundaries of what can be known" (Humphreys 2004, 154), depends on how much safe practices of dividing labor, which does not compromise the validity of the global inquiry, can be applied. The possibility to divide inquiries into standardized nontrivial units or modules that can be carried out independently and recombined together to yield sound results is beneficial, in particular for the validation of simulations. For example, different actors with pointed expertise can be in charge of each module and produce more reliable collaborative inquiries; some failures can be more easily localized by means of local tests; other failures may have local impacts, etc. The advantages of modularity are not specific to computer simulations. However, how much modularity is possible and beneficial for simulations and their validation requires a specific investigation. The following argument can be used to clarify the situation and explain why validating simulations can be difficult.

P1 If, in a perfectly modular structure, each individual module works, so does the whole structure.

P2 Simulation-based inquiries are perfectly modular.

P3 It is straightforward to check whether the modules of simulation-based inquiries work.

∴ It is straightforward to check whether simulation-based inquiries work

Evidence seems to suggest that conditions P2 and P3 are often false for simulations. For example, while well-designed modularity is desirable, it is often conspicuous by its absence from programs: "patches, ad hoc constructions, bandaids and tourniquets, bells and whistles, glue, spit and polish, signature code, blood-sweat-and-tears, and, of course, the kitchen sink—the colorful jargon of the practicing programmer seems to be saying something about the nature of the structures he works with" (Millo et al. 1979, 277). In various cases, there is uncertainty about how much modules actually work and whether potential departures from exactness are a worry. Typically, mathematical functions are often approximately computed by the versions that libraries provide. If users are not strongly aware of the limits of the specific functions within libraries, the results can be corrupted (see the case of random numbers described above). Lenhard also supplies the example of the practice of "kludging", i.e., using quick-and-dirty and hard to maintain solutions to make software work (Lenhard, forthcoming, see also Chap. 38 by Lenhard in this volume). This implies that knowledge concerning the validity domain of parts of the software can become lost. Further, in many cases, the fact that black-boxes are used makes it impossible to check deeper into the modules. Overall, this means that the global validity of computational inquiries can become corrupted and uncertainty often remains as to whether this is the case. In practice, modularity may be a solution, but not always a blissful one.

The various reasons for this corruption of modularity, whether it is inevitable, and the strategies developed by practitioners to preserve validity when modularity is eroded, are other questions worthy of study. Answers may differ depending on the aspects or fields considered. For software architecture, there are clearly "reasons for degeneration: ongoing evolutionary pressure, piecemeal growth. Even systems with well-defined architectures are prone to structural erosion" (Foote and Yoder 1999, Chap. 29, quoted by Lenhard). At the same time, "a sustained commitment to refactoring can keep a system from subsiding into a big ball of mud" (ibidem). How much safe modularity is preserved depends on which types of tools, practices, and norms are actually adopted within a scientific community, from the hardware to the modeling level. This is again a contingent issue, which epistemologists cannot analyze by armchair analyses. Importantly, different factors pull in different directions. Modularity brings about epistemic advantages, such as the facilitation of piece-wise validation and understanding of inquiries and their results. However, preserving modularity can be extremely costly. Similarly, reusing and adapting modules beyond their initial domain of validity to produce more results quickly is a legitimate concern, even if this tends to make errors more likely. Describing more precisely the trade-offs between these different epistemic goals can help to understand the constrained epistemic choices that resource-limited practitioners and communities are faced with, why some practices are considered as good, acceptable or sloppy, or why some types of errors or problems can be expected within computational inquiries. Depending on the orientations that are taken by communities concerning these matters, different types of computational science are possible.

43.3.1.5 Computer Simulations for All: What Epistemological Effects?

Computational science increasingly benefits from the development of various tools at the hardware, software, or modeling levels. Individual scientists would not be able to complete many inquiries without all these computational, mathematical, and modeling facilities. This situation keeps lowering the epistemic cost of the *running* of computer simulations (in terms of what one needs to know). Even scholars within communities with no strong training in computer science and mathematics can develop potentially valid simulations. But is this really safe: can individual scientists *really* afford epistemic ignorance and still produce sound computational results, or is this a lure? Actually, the epistemic price to *validate* results properly may remain high.[1] Indeed, the question of whether partly reliable tools and facilities work well usually calls for a context-specific answer, and determining this answer requires expert knowledge concerning both the tools and the subject matter. Thus, new tensions arise from these modern facilities, which offer opportunities to produce a wealth of results across scientific fields but come with new risks of failure. While this tension exists for other complex activities, it is extremely acute here. How scientists eventually behave, i.e., how much they cope on their own or ask other experts or collaborators for help hinges on many factors. These include the cost of human and computational resources, how much failure is risky and acceptable, whether errors are often detected by peers and tarnish scientific reputation, etc. The productivity and reliability of computational science can vary significantly depending on what is the case concerning such factors.

43.3.2 The Reception and Post Hoc Assessment of Computational Results

A bad result that is used has a detrimental impact. A sound result that is ignored has no beneficial effect. In both cases, we are epistemically worse off. Accordingly, epistemology must also scrutinize how results are publicly validated, accessed, trusted, and used once they are produced. I highlight below a couple of issues that make this problem specific for computational inquiries.

43.3.2.1 Epistemic Access

For mind-produced results, inferential processes and their conclusions, *qua linguistic entities*, are accessible to the authors, who personally carry out these activities. Publication extends this access to the public. Things are different for computer simulations. These are carried out by external processes. Thus, the authors no longer have a privileged epistemic position. Furthermore, even if the content of computer simulations can be described logically (putting aside issues concerning physical implementation) and can be made accessible provided that scientists preserve bit-reproducibility (Demmel and Nguyen 2013), in practice, practitioners usually cannot access the details of computational processes. In other words, simulations remain *globally epistemically opaque* (Humphreys 2009), even when they are *locally transparent* (Imbert 2017, 726): a human mind can inspect any part of the process though it cannot inspect all the parts. Things are worse for the more distant scientific audience. In most cases, the public can access a tiny fraction of the results through tables or graphs. In some cases, the whole data set and the code are available for inspection, while in rarer cases, this is true for the whole state-by-state simulation. However, it

[1] Similarly, knowing the main effects of drugs may give lay people the illusion that they can safely decide whether they should take them when they are sick.

is virtually never so for the bitwise computational process. Overall, how much of the process can be accessed, directly (by human minds) or indirectly (with software facilities), depends on computational questions, publication policies, issues related to openness and proprietary use, or the development and maintenance of storage facilities and exploitation software. (Note that the overlap with similar issues for experimental science is merely partial). Since public validation and good use are contingent on the possibility of access, the epistemic impact of simulations clearly depends on how this problem is socially and technologically treated within scientific communities.

43.3.2.2 Verification of Program Correctness

Accessing results is one thing, trusting them and using them is another. It does not matter that some type of process often produces adulterated results if its users can identify and use sound cases. In brief, whether it is possible to certify the reliability of simulations is crucial to their felicitous use. Here again, the epistemology of simulations overlaps with that of other activities but it has its specificities.

At a low level, program verification is a matter of verifying whether token computational runs have the appropriate causal behavior, which is a specific version of the problem of inductive inference (Fetzer 1988, passim). At a higher level, it can be seen as that of verifying whether algorithms and their coded counterparts do what they should. De Millo et al. (1979) argued that program verification does not work like proof verification. Mathematical proofs are usually sketches of proofs (*versus* formal proofs in the logician sense), and their logical validity is publicly discussed by mathematical communities. Program verification is different, because proofs of program correctness for real-life systems are long, tedious, and repetitive, and are not usually published nor publicly discussed (De Millo et al. 1979, 276). Dijkstra (a defender of program verification) counterargues that proofs of program correctness can also be the object of lively exchanges between scientists. Further, trivial mathematical theories also have simple statements “whose finite proofs are impossibly long” (Dijkstra 1978). Thus, for both proofs and programs, mathematicians need to find concise and elegant proofs.

Fortunately, for the purposes of this chapter, there is no need to endorse a position about the nature and ideals of program verification. Epistemology deals with what we can do *in practice* and what we actually do, given our epistemic, computational, and cognitive wherewithal and the incentives within our epistemic communities. Similarly, Wikipedia may change nothing of the nature of knowledge, justification, and science, but it changes how agents with limited resources and cognitive biases access knowledge. Thus, its existence modifies our epistemological situation and changes which beliefs propagate throughout human societies. Here, even if proofs of program correctness and mathematical proofs share the same nature and ideals, in practice, strong epistemological differences between them remain. If proofs of program correctness are usually not published, are less valued as scientific achievements, less scrutinized, much longer and extremely repetitive, then, from an epistemological point of view, program verification works differently.

Further, verification of programs is de facto a partly specific problem for scientific activities outside mathematics and computer science. Millions of lines of code are regularly written for the purpose of scientific activities. The more software facilities develop, the more scientists with no specific background in computer science write code, which is neither verified by formal methods nor undergo the interested scrutiny of computer scientists or mathematicians. This raises the question of how these codes are actually tested and how reliable related procedures are. In the absence of a grand theory of testing, "programmers are probably better off using the tools and insights they have in great abundance. Instead of guessing at deeply rooted sources of error, they should use their specialized knowledge about the most likely sources of error" (De Millo 1978, 41) and rely on their "intuition and problem-dependent knowledge in a disciplined manner to test for a variety of specified error types" (Shapiro 1997, 31). Further, program correctness merely guarantees that the implementation matches the specifications; but these can themselves be flawed (Shapiro 1997, 32), and unexpected physical, mathematical or computational conditions or situations can bring about failure. In brief, testing often relies on a messy combination of formal and nonformal, subject-specific, and partly dirty strategies. Thus, while the epistemology of computer simulations and software engineering is at the crossroads of other disciplines and overlaps with them, it does not reduce to them and requires a specific scrutiny from philosophers of the empirical science, even if they still lack the corresponding culture.

43.3.2.3 Verification of Mathematical Correctness

Even if computer simulations work properly at the hardware and software level, they may be unsatisfactory because they compute solutions that are not close enough to the unknown solutions of the target models or equations. In the frequent absence of mathematical theorems to guarantee that this is so, assessing whether this is the case is difficult. I will not develop this point here, as it has already been discussed in the literature (see e.g., Winsberg 2010, Chap. 2, Frigg and Reiss 2009, 603).

43.3.2.4 Reproducibility

Scientists can be willing to replicate or reproduce simulations and their results. Replication is costly and not all scientific results or simulations are replicated. Nevertheless, the possibility of replication is a cornerstone of science and the validation of scientific results. In principle, it is possible for entities that can be defined or presented unambiguously by linguistic means. By contrast, replicating experiments or thought-experiments can be controversial, which may feed epistemological problems like that of the experimenter's regress (Collins 1985).

Over the past decade, there has been a growing awareness that present scientific practices and publication rules often do not match replicability standards (Baker

2016). This is referred to as the replication, replicability, or reproducibility crisis in science. Until recently, it was seen as touching almost exclusively experimental science. As it turns out, computational activities are also concerned. Failure to reproduce computational results or to replicate a computation can stem from various sources: the authors may not share their code; the representation of real numbers may vary; the order of associative operations such as addition and multiplication may make a difference in floating point representations; programming languages, compilers, operating systems, and finally computational architectures may make a difference (Hill 2015), etc. How serious is the problem for computer science and computer simulations in particular? Some researchers like Claerbout have struggled over the years to create a reproducible research environment and have reported how difficult this has been (Fomel and Claerbout 2009). More recently, Collberg and Proebsting tried to replicate computer science research presented in 601 papers from the respectable Association for Machinery conferences and journals (Collberg and Proebsting 2016). They defined different degrees of repeatability based on how difficult they found it to repeat the research. In spite of their efforts, 47% of the 601 target papers turned out to present non-repeatable research. It is unlikely that computer scientists are more careless concerning reproducibility than researchers who simply use computer simulations for their research. Accordingly, epistemologists should not indulge in wishful thinking concerning the replicability of computer simulations, and, arguably, computer simulations also raise a specific reproducibility problem.

43.3.2.5 Trust

Overall, the impact of computational results depends on how much and when the results of computer simulations are trusted, and whether this trust is misplaced or not. If one takes a general, abstract, bird's eye view of this problem, it looks familiar and seems to boil down to the issue of how and when scientists accept being epistemically dependent on their peers and using their results (Hardwig 1985). The answer can be described in terms of networks of trust or trust indicators such as the scientific reputation of journals, scientists, or institutions. However, at a more fine-grained level, how much trust toward computational results is distributed and how these trust indicators are fed depends on the details of practices across fields. Here again, computer simulations may require specific scrutiny.

43.3.2.6 Publication Procedures and the Setting of Appropriate Standards

Publication procedures contribute to the production, assessment, and diffusion of good results. Tuning them appropriately for computational inquiries can be specifically beneficial with respect to some of the problems described above. I shall give brief examples here.

Because access to relevant information is crucial for replication or validation by peers, but also for novel inquiries that use existing data (e.g., those generated by big simulations such as the Millenium Run), editorial rules, or requests concerning what information authors must provide, as well as openness and proprietary issues, can influence the epistemic impact of computer simulations.

Editorial rules can also be used to keep "educating" members of scientific communities about what can spoil simulations (e.g., if authors are requested to provide detailed information about the (P)RNG they use). This is particularly true since computational science evolves at a brisk pace and communal practices need to keep adapting to guide individuals.

Beneficial results can also be achieved through appropriate authorship practices. Collaborative science is now widespread, which may undermine epistemic accountability and feed a decrease of reliability (Andersen 2014, Imbert 2014). In this context, major journals, like *Nature* or *JAMA*, have started adopting policies to make authors list their respective contributions, and what they endorse responsibility for (Rennie 2001), as well as who the guarantors are (Rennie 1997). In the context of computer simulations, adapted versions of these policies may be adapted to indicate the crucial scientific roles that must be endorsed to carry out and validate simulations properly. This may put virtuous pressure on practitioners, e.g., concerning the interpretation of agent-based models by computer scientists with no object-specific expertise, or the internal validity of simulations carried out by researchers with little expertise in computational methods.

43.4 Should Epistemologists of Science Bother, After All?

As seen above, belief-generating processes involving computer simulations can fail in various places, spoiling their epistemic impact. Can epistemologists ignore these issues? Epistemologists of science have a strong tradition of focusing on scientific representations. So far, the issue of the validation of simulations has mostly been tackled through the lens of the epistemology of models and question such as whether partial misrepresentations (e.g., due to idealizations, approximations or abstractions) threaten the conclusions that can be drawn from models (Frigg and Hartmann 2017). I now provide general arguments to the effect that an adequate epistemological analysis of computer simulations should extend beyond these questions.

43.4.1 Target Models, Actually Investigated Models, and Failure

When a simulation is carried out, a computational model is always exactly explored, even if it differs from the model targeted for investigation. So the validity of a simulation always boils down to that of this computed model. Therefore, why extending investigations beyond those of models?

Unfortunately, such a position begs the question. First, the model that is actually computed can be unknown, e.g., because unnoted errors spoil the investigation, so the above position makes scientific failure more difficult to analyze. Second, the exact description of these actually computed models should include hosts of gory mathematical and computational details as well as information about the software and hardware, i.e., much more than philosophers analyzing models usually discuss. Thus, at the very least, one must distinguish between the target models that one would like to investigate and those that are actually investigated, knowingly or not. Third, focusing on the content of target models also remains unsatisfactory. From an epistemological point of view, what matters is less the potential of target models than what is actually extracted from them by practitioners. Typically, if some Monte Carlo practitioners use low-quality random numbers, their results may be incorrect, whether or not the target average quantity in the model represents the target system property correctly. Overall, analyses about models and mathematical–computational practices are complementary. Just as investigations about the death toll on roads cannot be reduced to analyses of the driving code and road maps, discussions about the epistemology of models and scientific representations cannot save us the effort of epistemological investigations about mathematical–computational practices.

43.4.2 The Valuable Redundancy Argument

Let us assume for the sake of the argument that the epistemological analysis of computational models could exhaust that of computer simulations. Even so, much independent epistemological work would be needed to describe how other aspects of simulations, such as coding, mathematical practices, verification procedures, etc., favor the production of reliable results. This can be understood with an analogy to classical mechanics. Even if one knows exactly the trajectory of a deterministic system, there remain hosts of regularities to be discovered between other variables describing the system. These regularities are somewhat redundant, since anything about the system's behavior can be derived from the knowledge of its trajectory. Nevertheless, discovering such regularities remains epistemically valuable. Similarly, investigations into the reliability of computational practices and their epistemic impact are valuable, even if the validity of computer simulations is determined by the very content of the models that are investigated.

43.4.3 The Procrustean Objection

Finally, one might argue that many of the questions described above are novel but belong to formal or empirical science. As such, they might be discarded from epistemology, leaving nothing substantially novel in the epistemology of simulations. For example, Frigg and Reiss emphasize that questions, e.g., about the relationships between numerical and actual solutions or the impact of truncation errors are "purely mathematical problems" (Frigg and Reiss 2009, 592, 602).

This type of answer is perplexing. Once one has provided clear-cut and consensual notions of what counts as logical, mathematical, epistemological, etc., inquiries, one can analyze which questions fall within the scope of these inquiries. This is what I have done above with the consequentialist epistemological framework used by Goldman for social epistemology. In this perspective, nothing precludes that some problems or sub-problems belong to several disciplines—or one should explain why, whereas disciplines are historical and partly conventional constructs, there cannot be a partial overlap between them. The sub-problems that need to be tackled to pursue epistemological inquiries can also be considered to be epistemological ones, although perhaps derivatively. It would be implausible to claim that (the solutions of) philosophical or epistemological questions cannot involve (those of) mathematical or scientific questions.

Let us take an example. Suppose that one pursues epistemological investigations about journalistic practices and how much they promote the diffusion of true beliefs (see e.g., Goldman 1999, Chap. 6). Then, the solutions of various cognitive, technological, sociological, or economical questions about journalism and communication systems are relevant to these investigations and overlap with them. Further, these investigations coincide with those pursued by "theoretical journalists", who search for demonstrably reliable journalistic practices. However, epistemologists of journalism do not assess either the reliability of particular pieces of information or whether journalistic rules are applied correctly.

The case of science is analogous. Scientists try to develop safe practices to extend scientific knowledge. Thus, proving results about the reliability of particular methods, applying sound practices, and assessing the validity of particular inquiries is directly their task, even if it may provide indirectly relevant information for epistemological inquiries. Epistemologists analyze science and the reliability of its practices in order to present a faithful picture of science, given the epistemic, technological, sociological, etc., conditions in which it is practiced. Then, the assessment of the general reliability of scientific practices or possibility or impossibility results about these practices is common concerns for both inquiries. Emphasizing this overlap does not amount to confusing the goals and tasks of scientists with those of epistemologists.

Overall, if it deals with the epistemic analysis of natural belief-generating processes, epistemology inevitably intersects the fields that investigate specifically these processes, what they are and what they can be like. Accordingly, impossibility and complexity results in mathematics, logic, and computer science, results in cognitive and social psychology about reasoning and biases, results about the aggregation of individual judgments and preferences, or sociological analyses of how scientific communities work intersect epistemological inquiries. In the present case, mathematical questions about the complexity of verifying programs or psychological and sociological questions about how computational communities are organized epistemically and how scientists behave within them overlap with epistemological inquiries about the validation of computer simulations. Discarding the epistemological dimension of these intersectional questions for the purpose of a non-novelty argument amounts to elaborating an inadequate Procrustean version of epistemology.

43.4.4 The Absence of Data Argument and the Ostrich Strategy

Pointing at problems that can spoil the validity of computational inquiries is one thing, nevertheless, some may be already solved or may have a minor impact. Thus, one would need to know which of these problems frequently generate errors that threaten the validity of simulations, and which can be idealized away.

Unfortunately, it is extremely difficult to provide data about how often and why computer simulations fail since correct results are usually unknown and cannot be used as an external standard. However, the absence of accessible evidence about something in no way disproves its existence. There are many reasons why failures of computer simulations are not likely to be detected or publicized when they occur. First, simulations are not self-certifying activities in the sense that simulating a system does not produce direct evidence by itself that the simulation is successful. By contrast, juggling is self-certifying: when one juggles correctly, one immediately knows about it. Second, when computational inquiries unknowingly fail, usually some data are still produced. Once criteria of syntactic correctness are met, computer simulations always yield numbers, and practitioners need to deploy specific vigilance to track potential troubles. Third, not all robustness tests (e.g., by using different computational architectures, codes, libraries, etc.) can be carried out. Fourth, external detection of failure is often difficult because the details of computational activities are usually not public, replication is difficult, and incentives for replications are low. Fifth, because problems can be potentially ascribed to various tasks in the process, localizing failures means facing a specific version of the Duhem–Quine problem (Winsberg 2010, 24, Frigg and Reiss 2009, 604). This undermines scientific accountability and may encourage sloppier practices. Finally, even when failures are detected or suspected, nothing may happen, unless something major is at stake. Scientific life is short, resources are scarce, publicly localizing others' errors is time-consuming, and pay-offs for doing so are usually low. Accordingly, scientists may simply do nothing and let the results that seem fishy feed the gray zone of science. Overall, it is difficult to assess computational failure satisfactorily. Direct methods, e.g., by counting public detections of errors or retractions, are likely to grossly underestimate it and computational science runs the risk of the *invisibleness of its failures*.

It is often sound policy to leave aside issues that one cannot treat correctly. Nevertheless, the difficulty of directly observing some phenomena and the unavailability of objective standards for evaluation purposes are frequent in science, and they do not discourage scientists. The epistemological analysis of adjudication systems raises similar problems, because one never knows what the right verdict should be (Goldman 1999). Yet indirect ways out of the deadlock can be found out. For example, when a lay jury and a jury involving judges, or a real jury and a mock one give different verdicts, the two cannot be correct. This was used to investigate the effects of jury size and decision rules (Kalven and Zeisel 1966, Hastie et al. 1983). Here, the inability to replicate computer simulation results may, for example, be used as a general indicator of their invalidity. If some authors do not manage to replicate some computational results, this could mean that these results are sensitive to the method used, and their supposed scope is usurped. Alternatively, the method could have been badly implemented, or there may be some initial vagueness concerning the target model, which often surfaces when codes need to be effectively written. Overall, dropping the case of the epistemological assessment of computational practices on the ground of armchair arguments or, because it is difficult, would be tantamount to behaving like ostriches, which according to rumor bury their heads in the sand in the face of danger. Further, given the evidence that computational practices can fail in various specific ways, the burden of proof lies on the shoulders of the epistemologists of science who claim that sources of failure for simulations can be idealized away or ignored, except when it comes to their pet research topic (typically misrepresentation for philosophers of scientific models).

At the end of the day, I have no optimistic or pessimistic general conclusion to make about the present validity of computational inquiries. The point is rather that, given the *type* of activity that they are, and all the factors that can spoil them, it is not difficult to figure out states or domains of science in which simulations are sloppy or unreliable methods. Thus, it is worth investigating what is the case, why, and whether things can be improved epistemically.

43.5 Conclusion and Moral

Computer simulations have changed science. Over the past decades, it has been claimed that they also needed a novel, if not revolutionary, epistemology. Some such claims were over-stretched and the criticisms they raised were legitimate. However, one should be careful not to throw away the baby with the bathtub water. I have tried to present a sober version of the thesis of the epistemological novelty of simulations. I have adopted for clarification purposes a conceptual framework borrowed from Alvin Goldman and used it to emphasize that the computational, mathematical, representational, social, and potentially psychological dimensions of computational inquiries and their reception within scientific communities require specific epistemological investigations if one is to understand their validity and their epistemic impact. These investigations often raise novel questions, especially with respect to objects of novel types, like hardware and software, or call for novel and context-specific answers to traditional questions. I have not discussed the cognitive dimension of inquiries based on computer simulations, even if it is potentially an important one. For example, how we cognitively handle code or complex computational models, control computational activities, interact with computers, analyze and grasp computational data, which specific skills are required for these activities, and which type of biases are more frequent in this context are questions worth investigating. I have not discussed either how much the epistemology of activities like theorizing, predicting, evaluating, corroborating, explaining, or understanding, is altered when it is carried out by means of computer simulations (see Imbert 2017 for the last two questions).

Are these conclusions about the epistemological specificity of simulations so surprising? Over the past three decades, epistemologists and philosophers of science have provided analyses that Kitcher characterizes as belonging to the return of naturalists (Kitcher 1992). Against epistemological investigations that are almost exclusively centered on the content of representations, such approaches emphasize the epistemological importance of studying the various aspects of belief-generating processes, in particular, their psychological and social dimension (see Kitcher 1993, Solomon 1994, Goldman 1999, and Kitcher 2002 for an insightful overview). The above analyses fit within this naturalistic perspective and show the need to include a computational dimension, broadly construed, to epistemological analyses when computers are part and parcel of scientific belief-generating processes, which is an increasing majority of cases.

Overall, such a naturalistic epistemology is bound to be demanding for students of science. Philosophers of empirical science usually have a cognitively costly education, both in philosophy and empirical science. This makes the study of scientific representations a natural level of inquiry and an ecological niche for them, after decades of logic-oriented analyses of science. However, if the epistemology of science and computer simulations, in particular, requires delving deep into psychological, social, or computational aspects of scientific processes, an unfortunate combination of different types of expertise is needed to develop it. Furthermore, one cannot expect these epistemological questions about the uses of computers in the empirical science to be disciplinary central for philosophers of mathematics and computer science, sociologists, or psychologists. Naturally, analytically minded epistemologists should hail results showing that aspects or dimensions of computer-based belief-generating processes can be ignored or treated independently. Also, searching where it is easier can be methodologically sound and rational up to a certain point. Nevertheless, epistemologists should guard against the streetlight effect and unjustified simplifications for fear of producing an incomplete and distorted picture of the epistemology of computer simulations.

Attempts to refute extreme or early versions of claims do not provide solid evidence for considering that their moderate versions are totally false. Using such refutations to discard incipient and burgeoning analyses about a novel issue looks like falling prey to confirmation bias. Frigg and Reiss, after rejecting the idea that aspects of the epistemology of simulations are novel, defend a conservative normative stance about which scientific orientations should be adopted. They recommend considering analyses about simulations as merely feeding existing debates, in particular, those about scientific models. Although synergies are needed and overlaps are worth investigating, such a perspective is unduly narrow. Its blind adoption as a communal view may have a chilling effect and distract from important questions that deserve attention, at least if one wants to understand how scientific knowledge is developing at the current time.

Acknowledgements Past interactions with Roman Frigg, Stephan Hartmann, and Paul Humphreys about various issues discussed in this chapter were extremely stimulating, and I probably owe them more than I am aware of. I am also very grateful to the editors, whose comments contributed significantly to improving this chapter.

References

Andersen, H. (2014). *Epistemic dependence in contemporary science: Practices and malpractices*. In L. Soler, S. Zwart, M. Lynch, & V. Israel-Jost (Eds.), *Commentary on epistemic dependence in contemporary science: Practices and malpractices by Hanne Andersen* (pp. 161–173). Routledge Studies in the Philosophy of Science, London: Routledge.

Baker, M. (2016). 1,500 Scientists lift the lid on reproducibility. *Nature News, 533*(7604), 452.

Barberousse, A., Franceschelli, S., & Imbert C. (2009). Computer simulations as experiments. *Synthese, 169*(3), 557–574.

Barberousse, A., & Imbert, C. (2013). New mathematics for old physics: The case of lattice fluids. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 44*(3), 231–241.

Barberousse, A., & Imbert, C. (2014). Recurring models and sensitivity to computational constraints. Sherwood J. B. Sugden (Ed.), *Monist, 97*(3), 259–279.

Beisbart, C. (2018). Are computer simulations experiments? And if not, how are they related to each other? *European Journal for Philosophy of Science*, 1–34.

Bloor, D. (1976). *Knowledge and social imagery* (Routledge Direct Editions). London, Boston: Routledge & K. Paul.

Collberg, C., & Proebsting, T. A. (2016). Repeatability in computer systems research. *Communications of the ACM, 59*(3), 62–69.

Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice*. London, Beverly Hills: Sage Publications.

De Matteis, A., Pagnutti, S. (1988). Parallelization of random number generators and long-range correlations. *Numerische Mathematik, 53*(5), 595–608.

DeMillo, R. A., Lipton, R. J., & Sayward, F. G. (1978). Hints on test data selection: Help for the practicing programmer. *Computer, 11*(4), 34–41.

DeMillo, R. A., Lipton, R. J., & Perlis, A. J. (1979). Social processes and proofs of theorems and programs. *Communications of the ACM, 22*(5), 271–280.

Demmel, J., & Nguyen, H. D. (2013). Numerical reproducibility and accuracy at exascale. In *2013 IEEE 21st Symposium on Computer Arithmetic* (pp. 235–237).

Dijkstra, E. W. (1978). On a political pamphlet from the middle ages. *ACM SIGSOFT software engineering notes, 3*(2), 14–16.

Dijkstra, E. W. (1972). The humble programmer. *Communications of the ACM, 15*(10), 859–866.

El Skaf, R., & Imbert, C. (2013). Unfolding in the empirical sciences: Experiments, thought experiments and computer simulations. *Synthese, 190*(16), 3451–3474.

Fetzer, J. H. (1988). Program verification: The very idea. *Communications of the ACM, 31*(9), 1048–1063.

Fillion, N., & Corless, R. M. (2014). On the epistemological analysis of modeling and computational error in the mathematical sciences. *Synthese, 191*(7), 1451–1467.

Fomel, S., & Claerbout, J. F. (2009). Guest editors' introduction: Reproducible research. *Computing in Science Engineering, 11*(1), 5–7.

Fresco, N., & Primiero, G. (2013). Miscomputation. *Philosophy & Technology, 26*(3), 253–272.

Frigg, R., & Reiss, J. (2009). The Philosophy of simulation: Hot new issues or same old stew? *Synthese, 169*(3), 593–613.

Frigg, R., & Hartmann, S. (2017). Models in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2017. Metaphysics Research Lab, Stanford University, https://plato.stanford.edu/archives/spr2017/entries/models-science/.

Goldman, A. I. (1999). *Knowledge in a social world*. Oxford, New York: Clarendon Press, Oxford University Press.

Hardwig, J. (1985). Epistemic dependence. *Journal of Philosophy, 82*(7), 335–349.

Hastie, R., Penrod, S., & Pennington, N. (1983). *Inside the jury*. Cambridge, Massachusetts, United States: Harvard University Press.

Heinrich, J. (2004). Detecting a bad random number generator. CDF/MEMO/STATISTICS/PUBLIC/6850. University of Pennsylvania. https://www-cdf.fnal.gov/physics/statistics/notes/cdf6850_badrand.pdf.

Hellekalek, P. (1998). Don't trust parallel Monte Carlo. In *Proceedings Parallel and Distributed Simulation Conference* (pp. 82–89), Alberta, Canada.

Hill, D. R. C. (2015). Parallel random numbers, simulation, and reproducible research. *Computing in Science Engineering, 17*(4), 66–71.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, *169*(3), 615–626.

Imbert, C. (2014). The identification and prevention of bad practices and malpractices in science. In L. Soler, S. Zwart, M. Lynch, & V. Israel-Jost (Eds.), *Science after the practice turn in the philosophy, history, and social studies of science* (pp. 174–187). Routledge Studies in the Philosophy of Science, London: Routledge

Imbert, C. (2017). Computer simulations and computational models in science. In *Springer handbook of model-based science* (pp. 735–781). Springer Handbooks, Cham: Springer.

Jones, D. (2010). *Good practice in (pseudo) random number generation for bioinformatics applications*. Technical report, UCL Bioinformatics Group.

Kalven Jr, H., & Zeisel, H. (1966). *The American jury*. London: The University of Chicago press.

Kitcher, P. (1992). The Naturalists Return. *The Philosophical Review, 101*(1), 53–114.

Kitcher, P. (1993). *The Advancement of science: Science without legend, objectivity without illusions*. New York: Oxford University Press, 1993.

Kitcher, P. (2002). The third way: Reflections on helen longino's the fate of knowledge. *Philosophy of science, 69*(4), 549–559.

Lenhard, J., forthcoming. Holism, or the erosion of modularity-a methodological challenge for validation. *Philosophy of Science.*

Lenhard, J., & Carrier, M. (2017). *Mathematics as a tool-tracing new roles of mathematics in the sciences*.

Matsumoto, M., Wada, I., Kuramoto, A., & Ashihara, H. (2007). Common defects in initialization of pseudorandom number generators. *ACM Transactions on Modeling and Computer Simulation, 17*(4).

Rennie, D., Yank, V., & Emanuel, L. (1997, August 20). When authorship fails. A proposal to make contributors accountable. *JAMA, 278*(7), 579–585.

Rennie, D., Flanagin, A., & Yank, V. (2001). The contributions of authors. *JAMA, 284*(1), 89–91.

Shapiro, S. (1997). Splitting the difference: The historical necessity of synthesis in software engineering. *IEEE Annals of the History of Computing, 19*(1), 20–54.

Simon, H. A. (1957). *Models of man: Social and rational mathematical essays on rational human behavior in a social setting*. New York: Wiley.

Solomon, M. (1994). Social Empiricism. *Noûs*, *28*(3), 325–343.

Foote, B., & Yoder, J. (1999). *Pattern languages of program design 4 (= Software Patterns. 4)*. Addison Wesley.

Wilson, G., Aruliah D. A., Brown C. T., Hong N. P. C., Davis, M, et al. (2014). Best practices for scientific computing. *PLOS Biology, 12*(1).

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, Mass: Harvard University Press.

Winsberg, E. B. (2010). *Science in the age of computer simulation*. Chicago: Etats-Unis.

Woods, J. (2013). *Errors of reasoning: Naturalizing the logic of inference*. College Publications.

Index

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2