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.

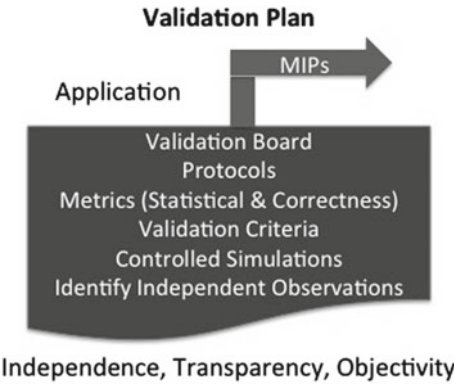

Fig. 30.1 Structure of a validation plan. The selection of the application of the model defines the details of the validation plan. The plan aspires to assure the values of Independence, Transparency, and Objectivity. The plan assumes that the modeling organization will participate in the community-based 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.

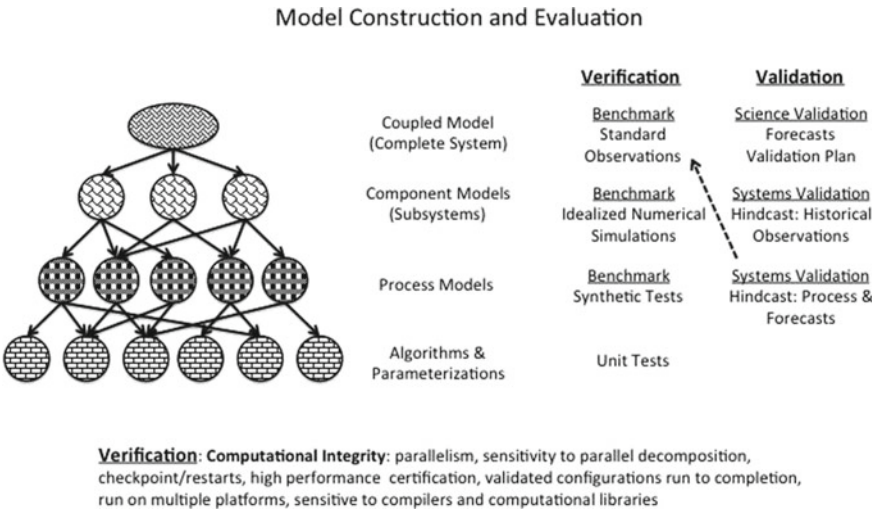

Fig. 30.2 Linking model structure to verification and validation. Following Oberkampf and Turcano (2002). The dashed arrow from Validation to Verification suggests that as Systems Validation with certain datasets evolves to a level of maturity that they no longer represent unique model quality; those 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-469 1STR*, 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.