Validation of Climate Models: An Essential Practice

Richard B. Rood

University of Michigan, Ann Arbor

Department of Climate and Space Sciences and Engineering

July 30, 2017

Introduction

This paper addresses the issues of the evaluation and validity of climate models. This is a subject that 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 paper 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 paper is the management, during the 1990s, of the Data Assimilation Office at the National Aeronautics and Space Administration's (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).

[1] Now Global Modeling and Assimilation Office (https://gmao.gsfc.nasa.gov/)

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, Shrader-Frechette, and Belitz (1994) sets the formal argument that in general, numerical models of geophysical phenomena cannot be validated. The argument is two-fold. First is that "the climate" cannot be observed in its entirety. Second is that models are non-unique 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 (Chapter ?); Lemos & 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 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 paper is organized as follows. The next section describes some of the arguments about both abstract and pragmatic attributes of model validation. This is followed by the definition of terms that describe the use of weather and 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.

Abstract and Pragmatic Attributes of Model Validation

Models in General

Though the literature on model validation extends back many more years, because of its importance, the Oreskes et al. (1994) paper serves as a good starting point. There are many words that emerge when both modelers and non-modelers discuss how it is determined whether or not climate models have credibility. Two of the leading words are "verification" and "validation," and Oreskes et al. (1994) state models cannot be either verified or validated. If this conclusion is accepted, then discussion of the how scientists determine the credibility of their models focuses on words such as evaluation, testing, calibration, certification, accreditation, and trustworthiness. These efforts in evaluation are focused on assessing the accuracy of not only the ability of models to represent observations, but also the accuracy of the underlying natural science (physics, chemistry, and biology), and the expression of the natural science on digital computers.

There have been a number of studies since Oreskes et al. (1994) that take on the subject of climate model evaluation and, *de facto*, validation. First noted is the work of Norton and Suppe (2001), who provide a comprehensive paper on the credibility of climate models. At the center of their discussion is that virtually all science relies on models in the collection of observations, formation of conclusions, and the development of knowledge. For example, in the determination of whether or not a satellite instrument is measuring temperature, the satellite data

are compared with a variety of other "conventional" or standardized measurements taken with surface thermometers, balloons, airplanes, and previously validated satellites. This requires collocation of the different sources of observations in both space and time. The evaluation process for the satellite instrument uses models at every level – the extraction of temperature from satellite radiances, the spatial and temporal collocation of information, the evaluation of consistency of information with other geophysical parameters, and the calculations of relating point source measurements techniques to layer averages. The ability of the satellite to measure temperature within a calculated range is determined and generally accepted as validation (e.g., Read et al., 2007). Norton and Suppe's argument is that models are so deeply ingrained into the scientific investigation that to state that models cannot be validated is equivalent to rejecting that scientific investigation is a robust way to generate knowledge.

In Norton and Suppe (2001), they focus on the concept of "truth" being represented by a unique curve and the reduction of model verification and validation to curve fitting. When the underlying uncertainty of observations and the role of uncertainty in the scientific method are considered, the concept of a unique truth curve is a spurious quest. Hence, finding truth is not necessary to establish the usefulness of models. Casual evidence shows that scientific investigation is a successful way to generate and describe knowledge; therefore, a unique representation of a perfectly observed climate is not the robust measurement of validation of our knowledge.

Norton and Suppe (2001) also discuss whether or not numerical experimentation and the use of models and simulation is a consistent analogue to controlled experimentation in a laboratory setting. This remains a discussion topic in philosophical literature. Petersen (2006)

and Frigg and Reiss (2009) provide interesting analysis of this discussion, coming to different conclusions.

Post and Votta (2005), a computational physicist and a computing engineer, pose computational simulation as a new branch of scientific methodology on par with theory and experimentation. Further, they argue that new methods of verification and validation are required in order for computational simulation to take its needed place in both the practice of science and the participation of science in society. In a 2004 Keynote Address to the IEEE International Conference on High Performance Computer Architecture, Post states that "… computational science does not have the predictive reliability of traditional methodologies such as theory, experiment, and engineering design. The results of many major computer applications are often wrong or are misinterpreted, sometimes with disastrous consequences. Computational science must mature as a field if it is to become a reliable methodology for addressing important problems" (Post, 2004). Post emphasizes increasing the roles and rigor of verification and validation as central to the credibility of computational science.

Weather and climate models rely on computational fluid dynamics. There is a large literature on the verification and validation of computational science in the field of computational fluid dynamics. There has been an effort to standardize the language, with, broadly, verification focused on 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 breakdown 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.

In a 1998 paper, Patrick J. Roache discusses verification and validation from philosophical, mathematical, computational, and scientific perspectives (Roache, 1998). Roache (2011) revisits issues from earlier works. With careful attention to language, Roache (2011) 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.

Roache (1998 & 2011) challenge, strongly, the philosophical conclusion of Oreskes et al. (1994) that models of natural systems cannot be validated. In the 2011 paper, Roache, an aeronautical engineer, recognizes and discusses the validation of climate models. Roache states,

> "Perhaps, it is true that we as a technological culture cannot adequately model underground nuclear waste disposal methods or climate changes, but such evaluation should not hinge on a philosophy of science that equally leads to the conclusion that there is no logical reason to prefer a tested theory to an untested one."

As part of Roache's conclusions, it is stated that the philosophical arguments about model validation have no relation to "… computational model validation, which is concerned with simple accuracy rather than some rarefied concept of truth of a theory."

Culture of Climate Model Validation

As Petersen (2006) points out, despite the arguments of philosophers, "(However) in the practice of simulation, the activities of 'validation' and 'verification' are deemed crucial by most simulationists." In climate science, verification and validation are vital aspects of the construction of climate models, and these processes are so ingrained into the cultures of modeling centers, they are often not specifically recognized (Shackley, 2001).

Several papers have investigated the practice of model evaluation at institutions in both the United States and Europe. Guillemont (2010) acknowledges "modelers tend to distrust the term 'validation' and prefer to use expressions like 'evaluation.'" Guillemont 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. Climate model credentials rely on model performance as evaluated by comparison with observations, with both observations and model performance woven together by consistency with theory.

Complexity of both climate models and Earth's climate is a recurring theme in the efforts to evaluate and establish the validity of climate models. Lehnard 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. Because of these multiple layers of complexity, Lehnard 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. They further conclude that it is unlikely there will be convincing convergence of climate model simulations and the reduction of uncertainty suggested by that convergence.

The weather and climate science community has developed a culture of model intercomparison projects (MIPs). The intercomparisons are not limited to model simulations, but include evaluations of observations and model-observation intercomparison. Perhaps the oldest formalized intercomparisons are associated with the World Meteorological Organization's Working Group on Numerical Experimentation (WGNE)[2]. As the importance of environmental simulation to planning and policy extended to problems beyond weather prediction, motivation and urgency for the intercomparisons grew. Prather and Remsberg (1993) describe an early and rigorous intercomparison for atmospheric ozone models and their use in the assessment of aircraft emissions.

Within the climate community, Gates (1992) describes the Atmospheric Model Intercomparison Project (AMIP). Characteristics of AMIP included simulation design, with the goal of all modeling groups to perform the same simulation. Also, important to the intercomparison was objective evaluation by independent experts, which was often achieved by

[2] http://www.wmo.int/pages/prog/arep/wwrp/rescrosscut/resdeptwgne.html

spanning a community of experts. That is, diagnostics were prescribed that all modeling groups had to provide and the ultimate analysis and synthesis include scrutiny by others than the model developers. The Coupled Model Intercomparison Project (CMIP)[3] 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[4].

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 credibility of future, unobserved, states. The observational-based analysis provides metrics, which emerge 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. That is, as described by Post and Votta (2005), referenced above.

These studies of the evaluation strategies of climate scientists establish that not only are there common practices at many modeling centers, but that the emergence of community intercomparison projects promotes the development of shared standards of evaluation.

---

[3] https://www.wcrp-climate.org/wgcm-cmip
[4] http://www.ipcc.ch/

Furthermore, given the complexity of both the climate models and the Earth's climate, these intercomparison projects are an important strategy to evaluate the climate as a system. Also established is 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. The evaluation process is made more robust by the use of theoretical constraints to integrate together the observations and simulations at the level of quasi-isolated physical processes – a practice that can establish cause and effect.

A notable aspect of simulation-based scientific investigation is that, early on in a discipline, curiosity and practice motivate the development of models as part of understanding observations and their correlated behavior. As simulations become an important part of a field's cultural practice, systematic evaluation becomes more important to credential model results with peers. As models become important to society, whether the design of aircraft's wings, the spread of an emergent disease, or exposing the impacts of atmospheric carbon dioxide emissions, evaluation takes on a far larger scope.

A culture of verification and validation emerges 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.

Definition of Terms

This section defines some key terms, which are summarized in Table 1.

Models are the representation of a system, of a phenomenon, or of an existing or posed object such as a ship or a building. The essence of a model, for example from the American Heritage Dictionary, is a "representation of something, especially a system or phenomenon, that accounts for its properties and is used to study its characteristics."[5] Models are, by definition, not the same as what they represent and are, usually, simpler than the phenomena they represent. Weather and climate models link observations, with mathematical expressions of the physical, chemical, and biological[6] behavior of those observations, with discrete, numerical approximations of those mathematical expressions, with computers that solve those numerical approximations (see Rood, 2010; Gettelman and Rood, 2016).

The previous section documents that modelers of natural science view the arguments over whether or not models can be validated as a semantic argument, based on the etymology of "valid." Dee (1995) says as much and maintains there is no argument about quantifying and describing the predictive skill of models, and defines the process of quantifying model skill as "model validation." Dee (1995) states 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."

---

[5] The American Heritage Dictionary of the English Language, retrieved from https://ahdictionary.com/word/search.html?q=Model

[6] Often it is stated that part of model validation is to evaluate the representation of "the science" or "the physics." In 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 non-local spatial scales. Climate models include not only physical processes, but also chemical and biological processes. Hence, the term "natural science" is used to refer to the theory-based description of observations of the weather and climate that follow from use of the scientific method.

Validation is but one word that enters into the practice of model evaluation. There is a need to define terms to provide a stable foundation for both communications as well as to comply with the principles of scientific investigation and to manage 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 terms based on the background and goals of individuals and expertise groups.

Several of the references in this paper take some effort in defining terms; however, there is no unique glossary. Evaluation will be used as a general term that includes both quantitative measures and qualitative analysis of a model's ability to address its design goals. It is common usage that validation involves comparison of model simulations with observations of nature or experiments to establish the accuracy of the natural science of the model. Accuracy is not definable in an absolute sense. Therefore, accuracy is tempered by measuring the suitability to address a specific application. Verification is often 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 also defined, essentially as a component part of verification or 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. As discussed in Clune and Rood (2011), there are many types of tests that are (or should be), routinely, part of model development. The evaluations related to software testing and verification should be completed before submitting the modeling system for validation by comparison with observations (Clune and Rood, 2011).

Defining the comparisons of simulations with observations of the climate to be validation, the process can be categorized in different ways. Systems validation is defined as

comparison with an established baseline of simulations from an earlier release of the modeling system. For example, 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 problems for which it was designed. For example, the ability to organize the types of storms common in the summer in the U.S. Great Plains. These storms are difficult to represent, require very high-fidelity representation of physical processes, and cross several scales of temporal and spatial behavior.

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. Such comparisons are primarily statistical, characterizing the spatial and temporal behavior. An important task of such validation is to determine the bias of the model with respect to observations. Variability, especially the representation of extremes, is an important metric because of the impacts of extremes on human and natural systems.

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 validation often relies on the collection of extraordinary datasets that, for example, trace the evaporation of water from the Earth's surface, its incorporation into atmospheric flow, its conversion from vapor to liquid and ice, and its return to the surface as precipitation. In this case, there might be focus on a specific cloud system and tracing the conservation of water and energy

through the lifecycle of a storm. The challenge is for the model to represent storms consistent with the transport and phase changes revealed by the observations. This process-based approach evaluates 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 to a model simulation that has been developed as a benchmark through, perhaps, a calculation at extraordinary resolution with a numerical method that is too expensive to be run routinely. 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 iterative, deliberative process, as opposed to a chain of hierarchical activities streaming up or down a decision tree.

The multilayered evaluation process uses several different types of models (Rood, 2010). The primary and implicit focus of this paper is the physical model. These 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. As noted above, in 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 non-local spatial scales. The fluid dynamical processes and local-scale processes represent the conservation laws and both are elements of the physical model.

If we have a system of partial differential equations that describe the behavior of the climate, then we can analyze those equations to reveal some powerful constraints. Using the atmosphere as an example, in middle latitudes (e.g. 45 degrees north or south), far enough above the surface of the Earth to ignore frictional drag, the flow is well described as a balance between forces due to differences in pressure (pressure gradient force) and the Earth's rotation (Coriolis force). This is the geostrophic balance. That a discrete, numerical model reduces to the geostrophic balance under appropriate simplifications is an important check on the model formulation. Geostrophic balance can be viewed as a conceptual or heuristic model, and describes correlated behavior based on fundamental theoretical considerations. Heuristic models provide 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" 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 structural decomposition. 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.

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 one year to the next (e.g., Johnson, Battisti, and Sarachik, 2000). Statistical models rely on having adequately observed behavioral relationships and for that behavior to remain the same (stationary) with time.

Two roles of statistical models in the evaluation of weather and climate models are introduced. A simple statistical model of weather is persistence; namely, tomorrow's weather is the same as today's weather. Persistence is "correct" more than half the time and defines a baseline for evaluation. Therefore, a measure of model "skill" for a physical model is to be better than persistence.

The second role of the statistical model in evaluation is to use the products from a physical model to derive statistical relationships analogous to those from observations. Though this comparison is not a robust measure of correctness, it is often a productive way to explore statistical models, physical models, and the robustness and completeness of observations. There is also the potential to investigate whether or not the statistics remain stationary; that is, changing with time and a measure of a changing climate.

Focusing on physical models, it is useful to identify two categories. Comprehensive models seek to model all of the relevant couplings or 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 in some way. As climate models have evolved, complexity has increased in incremental ways with coupling of atmospheric models with land, ocean, and ice models depending on the application. A weather model of the early 1990s could be viewed as a

climate model with specified boundary conditions. Today, a climate model and the most advanced weather models are made of coupled component models.

Climate Models and Observations in Practice

In the ideal practice of science, we have observed phenomena being 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 simple 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.

Figure 1 is a block diagram showing four elements of the practice of computational science to investigate and predict the Earth's climate. These four: Observations, Infrastructure, Models, and Assimilation, 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. Hence, these four elements are not connected with lines and arrows; rather, an iterative, deliberative process is suggested. Such an iterative process requires diligence and peer-based scrutiny to assure the integrity of science-based investigation.

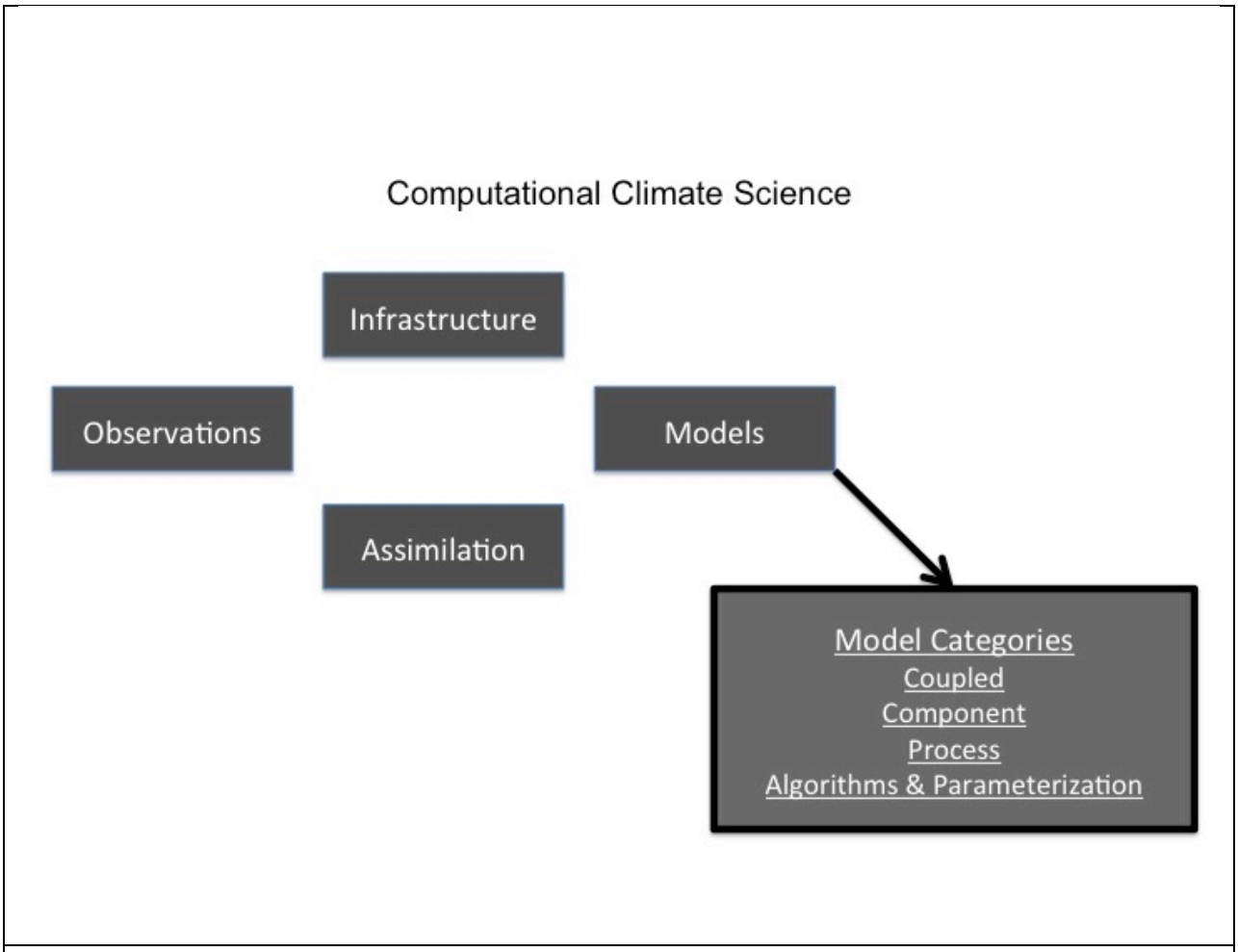

Figure 1: Foundational structures of computational climate science. The model structure is broken down into major categories as described in text.

Of the four elements, observations are at the foundation. Science relies on measured phenomena, observations. Models rely on observations. The observations of climate and climate change are complex. The incomplete definition of climate as "average weather" suggests the importance of wind, temperature, and water. Considering just water, we are immediately led to needing to know how water moves in the soil and is transferred between soil and air by plants. Knowing temperature leads us to needing to know about ozone, an atmospheric gas relevant to climate, but also subject to intense investigation for other reasons. The essence of climate and climate change requires a focus on oceans and ice and the measurements that characterize their

behavior. 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, pure body of knowledge; models and their applications steer observational needs.

Climate science is evolving and emerging from many different fields of natural science – meteorology, oceanography, hydrology, glaciology, etc. As a consequence, climate models are made by coupling component models: atmosphere, ocean, land model, sea ice, glacier and ice sheet, chemistry, biological – the list goes on (Edwards, 2010). 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 specific types of infrastructure are mentioned. 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. 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, relying on the definitions above, infrastructure requires the verification of its computational integrity, and enters into the portfolio of climate model components that require validation.

The second type of infrastructure is that which facilitates model analysis and model intercomparison. Examples of this type of infrastructure include the Earth System Grid Federation[7] (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 – a foundation of the scientific method. 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 (see below).

Focusing on the models, their validation, and their relation to observations, 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.

As described above, a climate model can be viewed as a coupled composite of component models. Each of the component models can be view 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 standard way to build models (e.g.,

[7] http://esgf.llnl.gov/

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. These model parts are decomposed in the breakout in Figure 1.

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 build climate models and, then, different observations to evaluate the model 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 last box of Figure 1, Assimilation, is considered.

Assimilation is the melding of model predictions with observations (Rood, 2010). Originally developed to provide the initial condition for weather forecasts, assimilation has become a core practice of weather and climate science. Weather forecasts are accurate enough in time ranges of hours to a day that they are used to generate estimates of observations of sufficient accuracy to provide quality control by monitoring observing systems (e.g., Stajner,

Winslow, Rood, and Pawson, 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 filling in processes that are not observed. Of course, these data-influenced estimates of processes are reliant upon the model parameterizations, which were, originally, defined with the help of other observations. The broad use of assimilated datasets know 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.

Validation of Climate Models in Practice

The verification and validation of weather and climate models requires the evaluation of many decisions, steps, and processes. 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
- the ability of the coupled model to represent conservation of energy, mass, and momentum

As described above, 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.

Independence, Transparency, and Objectivity: Basic principles of verification and validation: 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 system-wide testing, verification, and validation procedures. Standard tools and methods as well as a commitment to communication are needed. 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.

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 parts of validation. Such independence serves to evaluate robustness of logic and correctness of 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.

In order to support independent review, transparency of the test plan and validation plan is required. 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.

For climate science, transparency is especially important. The results from 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 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).

Transparency and the documentation of models and their validation support another attribute of scientific investigation, reproducibility. Ultimately, validation requires that scientists at other institutions recreate results with independent means. 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 that results are correct as based on evidence of quantitative measures. The end result, in this case, is 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.

Identification of Independent Observational Data: It is useful, first, to consider 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 (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.

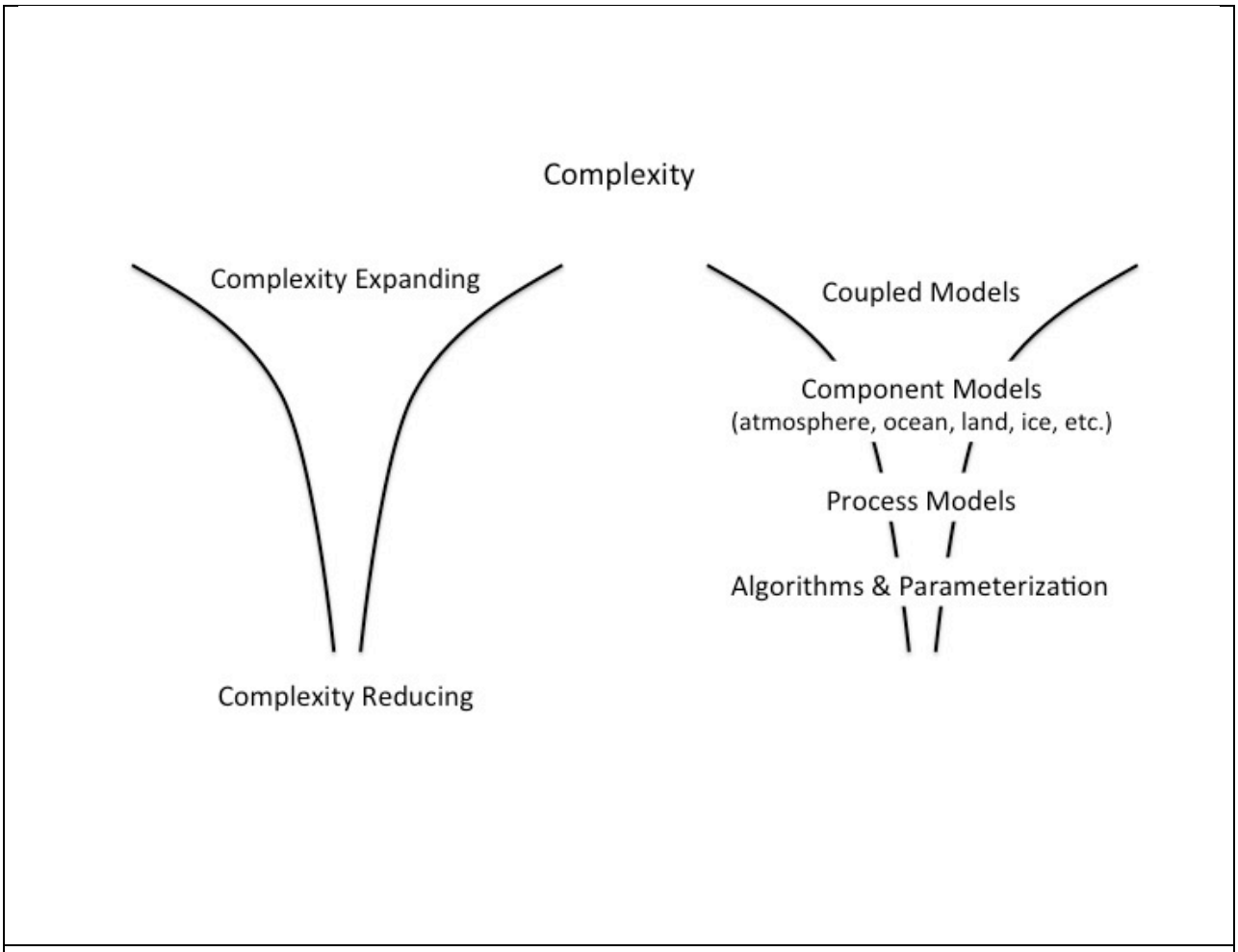

Figure 2: Left panel is a schematic of complexity, represented as a funnel. Validation of observations looks at reducing complexity. Climate models are complex and their development increases complexity. Therefore, validation requires managing increasing complexity. Right panel overlays the decomposition of a coupled climate model into the hierarchy of complexity.

Validation approaches for satellite observations have a fundamental difference with climate model validation. Figure 2 is a schematic designed to represent complexity. The narrow base of the funnel represents less complexity; the wide mouth represents more. 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 downward in the funnel, towards less complexity. A climate model, however, looks towards the

mouth of the funnel. As component models are combined into climate models more and 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 differences between the observational data sets are often related to details in treatment of, for example, day-night (sun-dark) influences on instruments, satellite orbital changes, or instrument degradation (e.g. Mears and Wentz, 2017). The discrepancies between models and observations 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 dance between simulations and observations and the intertwined roles of models challenges both the researcher and the communication of the research for societal use.

A sustained, iterative interaction between simulation and observations is part of the practice of climate science. For convincing model validation, it is necessary to identify when observations and models are not independent. Some 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. 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.

For the purpose of this paper, it is assumed that due diligence has been exercised to assure simulation-observation independence.

A Schema for Climate Model Validation: The goals of testing, verification, and validation are to assess correctness at all stages of development and implementation. Validation, the comparison of model simulation with observations relevant to a particular application, establishes the accuracy of the model relative to observations, analyzes whether cause and effect are meaningfully represented, and provides a description of uncertainty. The validation process communicates the trustworthiness of a model to its users.

The first step in validation is the development of a validation plan. The plan addresses the principles of independence, transparency, and objectivity. The elements of a validation plan are highlighted in Figure 3 and described below.

An organization must write and agree upon the evaluation criteria and metrics at the beginning of a model development phase. This 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.

If the validation criteria are known at the beginning, then definition is added to the validation process. During development, it is always the case that scientists and developers can identify further improvements. The plan, therefore, defines an end point, based on how well the model performs at a particular snapshot and assures that the model addresses particular user needs.

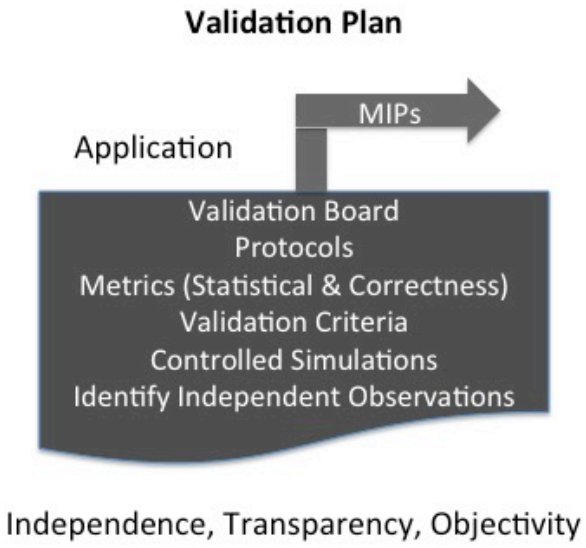

Figure 3: Structure of a validation plan. The selection of the Application of the model defines the details of the validation plan. The plan aspires to assure the values of Independence, Transparency, and Objectivity. The plan assumes that the modeling organization will participate in community-based model intercomparison projects (MIPs).

Adherence to a validation plan improves the ability of an organization to allocate human and computational resources. An organization is better able to meet goals within budget and on schedule. The end result is that model and model products will be based on evidence from a scientific process.

Within the plan, independence is assured by 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 falls 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. 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, less scientifically correct schemes can often be modified to meet specified statistical measures. However, in the end, correctness of the equations and their representation as numerical algorithms improves 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. 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. To maintain an old, verifiably less scientifically correct component is not justified from a scientist's perspective. The degradation of statistical skill is, in fact, pointing towards more scientific investigation to improve the model. The use of a more rigorous component is a better foundation for future development. The validation plan, therefore, needs to consider what potential negative scores can be tolerated to accommodate more robust future development.

The design of the validation plan should assure that the model is susceptible to quantitative validation. This requires an approach analogous to controlled experimentation. 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. A full validation might take many months to support the execution of the simulations, the analysis, and the deliberations. This cost motivates fine grain testing (Clune and Rood, 2011) and bundling a set of testable changes in a validation exercise.

Given the complexity of the component models that are coupled to make a climate model, validation strategies that are built on validation of component models are often used. Hence, climate model validation has the challenge of multiple changes being tested in a coupled environment. Component level validation is necessary, but not sufficient. For the model to be convincingly validated, the number of model changes needs to be limited. Analysis of the expected outcomes of the individual changes need to be posited and included in the evaluation criteria.

Final validation requires that the coupled model be validated. This is expensive and validation strategies continue to evolve. A priority for future validation practice is to develop strategies for fine-grain testing of coupled models to provide a more robust foundation for coupled model validation.

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[8]. Model intercomparisons have a strong influence on model validation and greatly enhance the

[8] https://www.wcrp-climate.org/wgcm-cmip

confidence in climate model results. The intercomparisons are an important foundation for describing the uncertainty. Therefore, experimental design and validation criteria for simulations extend outside of institutions to allow community-based experiment and validation protocols.

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).

Returning to Figure 2, the model categories from Figure 1 are placed in relation to 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 are composites of algorithms and parameterizations, which represent a quasi-isolated, highly observed geophysical feature such as a type of Arctic cloud (e.g., Roesler, Posselt, and Rood, 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 design of controlled simulation experiments and the design of validation exercises that have realizable metrics is self-evident. The requirement for an application or purpose of a particular model to limit the complexity is, likewise, self-evident.

Figure 4 pictorially describes the validation schema. 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 (computational accuracy) and validation (representation of the Earth).

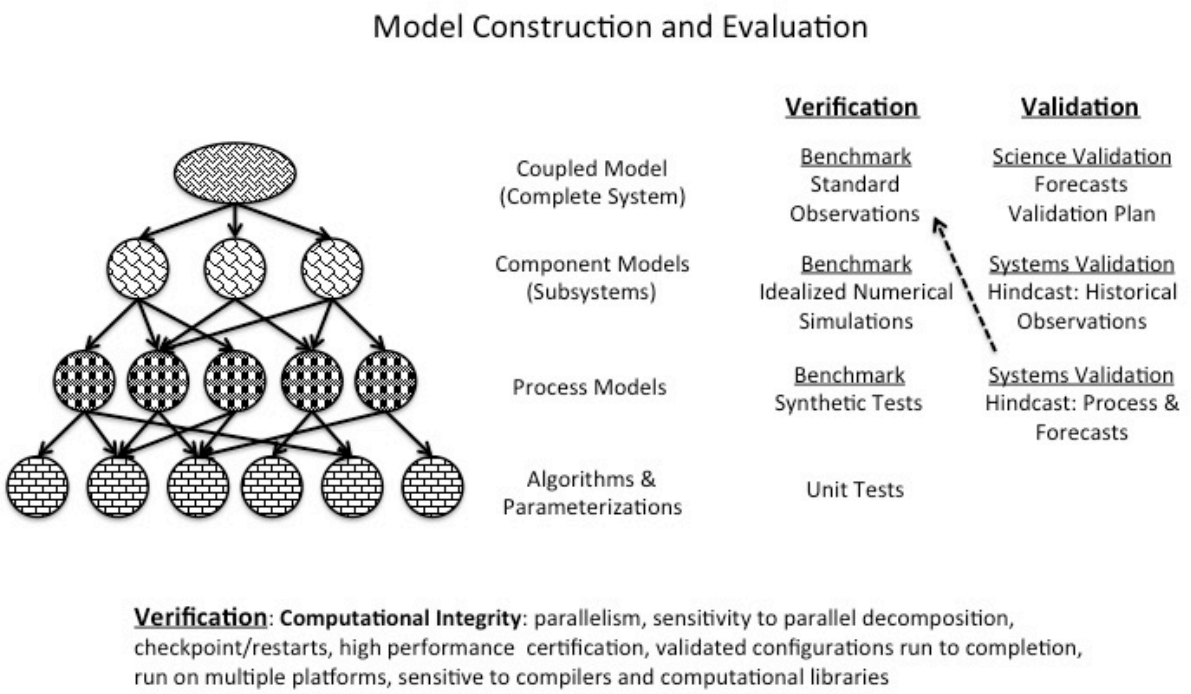

Figure 5: Linking model structure to verification and validation. Following Oberkampf and Turcano (2002). The dashed arrow from Validation to Verification suggests that as Systems Validation with certain datasets evolves to a level of maturity that they no longer represent unique model quality, those tests move to verification. That is they become benchmark standards that all models are expected to achieve.

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 associated with numerical representation of the science-based equations set.

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, for example, 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. For example, does a re-mapping 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 (perhaps, function or subroutine), and errors revealed at the unit test level support effective 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, the ability to configure 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 quality. Such tests might be viewed as minimal standards or benchmarks as all models are expected to do the benchmarks well. Confidence is added to the testing procedure if a more complex, coupled model can be configured for testing, rather than the test models serving as stand-alone models.

Proceeding to the highest levels of complexity, component models and coupled models, there is 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.

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.

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 are likely to include a set of 10-day weather forecasts from standard specified initial conditions that include all seasons.

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. Such 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 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. A simple example would be the ability of the model to represent the annual cycle of ozone variability. More sophisticated standard measures, which are designed to represent the combined effects of transport and chemistry, such as those described in Douglass et al. (1999) become automated from standardized output and provide quick and profound measures of model performance. The diagnostics of Douglass et al. (1999) rely on strong theoretical constraints; that is, heuristic models.

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-levels systems level evaluation can be automated. An excellent

example of publically available automated validation information for the Community Earth System Model can be found online[9].

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[10].

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 simulation of the 20th century, or more generally the post-industrial to 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 20th century allows examination of important modes of air-land-sea interactions, response to volcanoes, and some aspects of solar variability. Longer time-scale variability associated with oceans and ice are not fully represented in the 20th century record.

A possible disadvantage of the 20th 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 the ozone hole. These changes mean that we do not

[9] http://www.cgd.ucar.edu/amp/amwg/diagnostics/plotType.html
[10] http://iri.columbia.edu/our-expertise/climate/forecasts/enso/2017-June-quick-look/

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 relative short lived (< 5 years) and involve heating and cooling, water vapor responses, 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 model for broader use. However, further scientific validation better substantiates its credibility.

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. 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; 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-time scale of the needed forecasts allows the validation process to be concluded in weeks to months. For longer times scales 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 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 evolve as communities of users exercise the model over, on the order, of 18 – 24 months. This is time scale appropriate for deliberative research and peer review as well as development and validation of a model.

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 problems. The validation plan focuses on not just statistical measures of defining parameters, but on physically based, correlated behavior. That is, processes related to cause and effect. Improvement in the representation of processes stands along with statistical measures of performance.

In the validation exercise, it is near certain that some metrics will improve and some will degrade. It is likely that sensitivity experiments will reveal direct connections between improvements and degradations. 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.

Synthesis and Summary

Daniel Farber, a Professor at the University of California Berkeley, explored 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:

> "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."

This chapter has deconstructed and organized the practice of weather and 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, weather and 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 usable information for planning and practice. Climate scientists and

computational scientists, in general, have developed a culture of verification and validation that establish model's credibility and legitimacy.

Predictions and projections will always be uncertain; that 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 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 for thinking 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.

Table 1: Definitions

	Term	Definition
Models		
	Model	"representation of something, especially a system or phenomenon, that accounts for its properties and is used to study its characteristics" American Heritage Dictionary (https://ahdictionary.com/word/search.html?q=Model)
	Physically-based (physical) Model	use first-principle laws of conservation energy, momentum, and mass to represent and predict weather and climate
	Component Model	physical model of atmosphere, ocean, land, ice, chemistry, biology, etc. A discipline based model of a major sub-discipline of climate science
	Coupled Model	a model built from connected component models
	Comprehensive Model	seeks to represent all of the relevant couplings or interactions in a system
	Mechanistic Model	some variables or boundary conditions are prescribed and the system evolves relative to the prescribed parameters
	Process Model	physical model focused on some quasi-isolated processes that can be measured with high fidelity
	Statistical Model	predict future behavior based on past, observed behavior
	Heuristic Model	describes correlated behavior based on fundamental theoretical considerations
	Candidate Model	a model under development, submitted to validation, intended to improve upon previously validated models
	Diagnostic Model	model used to determine and to test the processes representing a set of observations
	Prognostic Model	model used to make a prediction
	Local-scale physics parameterization	fine-scale structure of model based on statistical relationships between observed variables and of an evolving dynamical system

	Algorithm	fine-scale structure of model based on numerical formulation of physical processes and functions that are directly derived from the underlying equation set
Evaluating Models	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 design goals
	Testing	checks the performance, quality, reliability – generically, in a way that is narrowly defined compared to the model as a whole
	Unit Tests	fine-grained, low level tests to assure that instructions or algorithms are correctly implemented
	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
	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 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

References

Cash D. W., Clark W. C., Alcock, F., Dickson N. M., Eckley, N., Guston D. H., Jäger, J., and Mitchell R. B. (2003), Knowledge systems for sustainable development, *Proc. Natl. Acad. Sci. USA, 100,* 8086–8091.

Clune, T. L., and Rood, R. B. (2011). Software testing and verification in climate model development. *IEEE Software, 28*, 49-55, doi: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, 1–14.

Douglass, A. R., Prather, M. J., Hall, T. M., Strahan, S. E., Pasch, P. J., Sparling, L. C., Coy, L., and Rodriguez, J. M. (1999). Choosing meteorological input for the global modeling initiative assessment of high-speed aircraft. *J. Geophys. Res.,* 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.

Frigg. R., and Reiss, J. (2009). The philosophy of simulation: Hot new issues or the same old stew. *Synthese, 169*, 593-613, DOI: 10.1007/s11229-008-9438-z.

Gates, W. L., (1992). AMIP: The Atmospheric Model Intercomparison Project. *Bull. Amer. Meteor. Soc., 73*, 1962–1970.

Gettelman, A., and Rood, R. B. (2016). Demystifying Climate Models: A Users Guide to Earth Systems Models. Berlin, Heidelberg: Springer Berlin Heidelberg. http://dx.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.

Johnson, S.D., D.S. Battisti and E.S. Sarachik, 2000. Empirically derived Markov models and prediction of tropical Pacific sea surface temperature anomalies. *J. Climate, 13*, 3-17.

Lenhard, J. and 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. and Rood, R. B. (2010). Climate projections and their impact on policy and practice. *Wiley Interdisciplinary Reviews: Climate Change*, *1*, 670-682, DOI: 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. and Wentz, F. J. (2017). A satellite-derived lower tropospheric atmospheric temperature dataset using an optimized adjustment for diurnal effects, *J. 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., and Suppe, F. (2001). Why atmospheric modeling is good science. In C.A. Miller and P.N. Edwards (Eds.), *Changing the Atmosphere: Expert Knowledge and Environmental Governance* (pp. 67-105). Cambridge, MA, USA: The MIT Press.

Oberkampf, W. L., and Trucano, T. G. (2002). Verification and Validation in Computational Fluid Dynamics, *SAND2002 – 0529,* Albuquerque, NM, USA: Sandia National Laboratories.

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

Petersen, A. C. (2006). *Simulating Nature: A Philosophical Study of Computer-Simulation Uncertainties and Their Role in Climate Science and Policy Advice*. Amsterdam: Het Spinhuis.

Post, D. (2004). The Coming Crisis in Computational Science. *LA-UR-04-0388*, Los Alamos, NM, USA: Los Alamos National Laboratory.

Post, D. E., and Votta, L. G. (2005). Computational science demands a new paradigm. *Physics Today, 58*, 35-41.

Prather, M. J., and Remsberg, E. E. (Eds.) (1993). The atmospheric effects of stratospheric aircraft: Report of the 1992 models and measurements workshop. *NASA Ref. Publ. 1292*.

Read, W.G., A. Lambert, J. Bacmeister, R.E. Cofield, L.E. Christensen, D.T. Cuddy, W.H. Daffer, B.J. Drouin, E. Fetzer, L. Froidevaux, R. Fuller, R. Herman, R.F. Jarnot, J.H. Jiang, Y.B. Jiang, K. Kelly, B.W. Knosp, L.J. Kovalenko, N.J. Livesey, N. H.C.Liu, G.L. Manney, H.M. Pickett, H.C. Pumphrey, K.H. Rosenlof, X. Sabounchi, M.L. Santee, M.J. Schwartz, W.V. Snyder, P.C. Stek, H. Su, L.L. Takacs, R.P. Thurstans, H. Vomel, P.A. Wagner, J.W. Waters, C.R. Webster, E.M. Weinstock, and D.L. Wu, (2007). Aura Microwave Limb

Sounder upper tropospheric and lower stratospheric H2O and relative humidity with respect to ice validation. *J. Geophys. Res. 112,* D24S35, doi:10.1029/2007JD008752.

Robock, A. (1983). El Chichón provides test of volcanoes' influence on climate. *Nat. Wea. Dig., 8,* 40-45.

Roesler, E. L., Posselt, D. J., and 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. *J. Geophys. Res., 122,* 4378-4400.

Rood, R. B. (2010). The role of the model in the data assimilation system. In W. Lahoz, B. Khattatov, and R. Menard (Eds.), *Data Assimilation: Making Sense of Observations* (pp. 351-379). Berlin, Heidelberg: Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-74703-1_14

Roy, C. J., and Oberkampf, W. L. (2011). A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Comput. Methods Appl. Mech. Engrg., 200*, 2131-2144.

Santer, B. D., Solomon, S., Pallotta, G., Mears, C., Po-Chedley, S., Fu, Q., Wentz, F., Zou, C-Z, Painter, J., Cvijanovic, I., and Bonfils, C. (2017). Comparing tropospheric warming in climate models and satellite data. *J. Climate, 30*, 373-392.

Shackley, S. (2001). Epistemic lifestyles in climate change modeling. In C.A. Miller and P.N. Edwards (Ed), *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 J. Numer. Anal., 5*, 506-517.

Stajner, I., Winslow, N., Rood R. B., and Pawson, S. (2004). Monitoring of observation errors in the assimilation of satellite ozone data. *J. Geophys. Res., 109,* D06309, doi: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. *J. Geophys. Res., 106,* 7183-7192.

Theurich, G., DeLuca, C., Campbell, T., Liu, F., Saint, K., Vertenstein, M,. Chen, J., Oehmke, R., Doyle, J., Whitcomb, T., Wallcraft, A., Iredell, M., Black, T, da Silva, A. M., Clune, T., Ferraro, R., Li, P., Kelley, M., Aleinov, I., Balaji, V., Zadeh, N., Jacob, R., Kirtman, B., Giraldo, F., McCarren, D., Sandgathe, S., Pechham, and Dunlap IV, R. (2016). The Earth

System Prediction Suite: Toward a coordinated U.S. modeling capability. *Bull. Amer. Meteor. Soc., 98*, 1229-1247

Williams, D. N., Balaji, V., Cinquini, L., Denvil, S., Duffy, D., Evans, B., Ferraro, R., Hansen, R., Lautenschlager, M., and Trenham, C. (2016). A global repository for planet-sized experiments and observations. *Bull. Amer. Meteor. Soc., 98*, 803-816 http://dx.doi.org/10.1175/BAMS-D-15-00132.1

Roach references
IPCC validation reference
Jablonowski reference
Uncertainty and Usability reference
Define accuracy, correctness, consistency, evaluation