JAMES | Journal of Advances in Modeling Earth Systems

COMMENTARY
10.1029/2019MS001720

Special Section:
Historical, Philosophical and Sociological Perspectives on Earth System Modeling

Philosophical Perspectives on Earth System Modeling: Truth, Adequacy, and Understanding

G. Gramelsberger[1], J. Lenhard[2], and W.S. Parker[3]

[1]Theory of Science and Technology, RWTH Aachen University, Aachen, Germany, [2]Department of Philosophy, Bielefeld University, Bielefeld, Germany, [3]Department of Philosophy, Durham University, Durham, UK

Key Points:
- Earth System Modeling can provide "pragmatic" understanding.
- Normative practices help modelers to deal with limits of understanding.
- An adequacy-for-purpose approach to model evaluation has implications for practice.

Correspondence to:
W. S. Parker,
wendy.parker@durham.ac.uk

Citation:
Gramelsberger, G., Lenhard, J., & Parker, W. S. (2020). Philosophical perspectives on Earth system modeling: Truth, adequacy, and understanding. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001720. https://doi.org/10.1029/2019MS001720

Received 18 APR 2019
Accepted 25 NOV 2019
Accepted article online 10 DEC 2019

Abstract We explore three questions about Earth system modeling that are of both scientific and philosophical interest: What kind of understanding can be gained via complex Earth system models? How can the limits of understanding be bypassed or managed? How should the task of evaluating Earth system models be conceptualized?

1. Introduction

Increasingly, philosophers of science are interested in computer simulation in science. There is growing awareness that complex simulation modeling—including Earth system modeling (ESM)—presents a number of challenges to traditional philosophical conceptions of how science works. The contributors to this paper explore several of these challenges, developing ideas that they presented in a special session of the fourth International Conference on ESM (4ICESM), held in Hamburg, Germany, in August 2017.

In section 2, Johannes Lenhard discusses the challenge of understanding via complex Earth System Models; he outlines a "pragmatic" perspective on understanding informed by competing past accounts. Continuing with the topic of understanding in section 3, Gabriele Gramelsberger considers how limits to understanding can be bypassed or managed. She points to the role of normative practices, especially the practice of believing in "true" values when developing models and the practice of sensitive wording when evaluating models and communicating uncertainties. The topic of model evaluation is addressed further in section 4, where Wendy Parker outlines an adequacy-for-purpose approach to evaluation and considers its implications for modeling practice. Section 5 offers some closing remarks.

2. Understanding and Complexity—The Dilemma of Growth (J. Lenhard)

All sciences—natural ones, social ones, and humanities alike—tend to develop a terminology that distances them from ordinary language and thinking. However, there are instances where common and scientific concepts meet, typically not without tension. This section discusses one such instance: the notion of understanding. Scientists might employ procedures only specialists can follow; still, when their goal is to "understand" something, practically everybody knows what they are after. To understand something is part of our common culture. Or so it seems.

Articulating a notion of understanding has been a challenge for philosophy throughout history. The philosophy of science (PhS), in particular, has grappled with the relationship between understanding and scientific explanation. Recently, methods of computer simulation have created a new challenge—they are heavily based on mathematical models that are themselves so complex that some important aspects of their behavior are accessible only via simulation experiments. ESM is a prime example of a field in which this challenge arises. Complex simulation models are indispensable, and the community tries to find strategies to establish and to maintain understanding of the models' behavior and of the Earth System itself. Such strategies figured prominently in many contributions to the 4ICESM conference in Hamburg in 2017. The remainder of this section pursues the idea that dealing with simulation models might influence and even change our concept of understanding. After briefly analyzing understanding as a goal for simulation modeling, a pragmatic concept of understanding is proposed that contributes to the discussion in both ESM and PhS.

2.1. The Dilemma of Growth Makes Understanding an Urgent Topic in ESM

In ESM, two tendencies stand in conflict with each other. On the one hand, simulation modeling is the method of choice (arguably without alternative) for understanding the dynamics of a system as complex

as the Earth System. On the other hand, the growing complexity of the models themselves seems to jeopardize understanding. These two tendencies constitute what can be called a "dilemma of growth." This dilemma has received attention in the ESM community. Manabe (2006), for instance, pointed out that understanding is the central goal of climate modeling and that this means "getting the essence of the mechanism." ESM, Manabe added, is bound to develop ever more complex models, leaving it unclear whether the goal can be reached at all. The principal strategy of mathematical modeling is to reduce complexity by stripping off aspects until only the essence remains. However, atmospheric circulation models do not only include fundamental equations (arguably what "essence" refers to) but also a plethora of additions like parameterizations of clouds and chemistry modules. Furthermore, ESM works with interacting models of atmosphere, ocean, and so forth. The principal strategy hence seems to be hardly feasible for ESM.

Both PhS and ESM have started to investigate understanding in connection with simulation. Parker (2014) focuses on climate modeling, discerns two sorts of understanding, and discusses how simulation can contribute to them. In a broad sense, she points out that simulation can reveal the implications of physical assumptions so that complex phenomena can be explored by conceptual models. While this is correct, the dilemma of growth arises when models become too complex—in a sense, when too many assumptions are at work. Climate scientists like Sandrine Bony, Isaac Held, and Bjorn Stevens (Bony et al., 2013; Held, 2005) have formulated and analyzed variants of the dilemma. One approach to rescuing understanding invokes modularity and hierarchies. The hope is that small and well-understood models can be knitted together in larger hierarchies so that understanding extends to the whole. However, the prospects of this approach are not yet clear, in part because modularity tends to erode in larger simulation models, leading to a problem of "holism" (see Lenhard, 2018). Below it is argued that practices of ESM—and complex simulation modeling in general—might suggest a concept of understanding that differs from the concept Manabe referred to.

2.2. Understanding Is an Established but Contested Concept

How do philosophers of science explicate the notion of understanding? Since this is not a systematical essay about the history of philosophy, it is helpful to simplify (see de Regt et al., 2009, for a recent anthology on understanding). A majority position in PhS conceives of understanding as an add-on to explanation, where the latter is cast in objective terms like logical relationships. If something explains a phenomenon, it has the potential to also provide understanding. A standard example is the orbit of a planet that is explained by a law of nature (gravitation) plus initial and boundary conditions (like the mass of the planet). This position has been elaborated by a group of philosophers that spans several generations, among them the classic Hempel (in the 1930s), the unconventional Toulmin (1960s), and Bangu (2017) who investigates the role of unification for understanding. This position stresses the objective component of understanding—does the form of an orbit in fact derive from law(s) plus conditions? It might therefore be called the "derivation plus" conception. This position presupposes that the entities in play, together with their relationships, are transparent enough for model users to recognize what derives from what.

A different strand in philosophy (and history) identifies the concept of understanding as an issue where sciences and humanities diverge. This strand juxtaposes understanding with explanation, counting the latter to the sciences and the former to the humanities (see von Wright, 1971, for an influential account). On this sort of view, a chess player understands her opponent, when her expectations of what comes next match with what comes next. Call this the "match" conception of understanding. This conception remains agnostic about the mechanisms that created some phenomenon of interest but stresses the properties of representations as imitations.

Weber (1913) offered an interesting approach that combines both conceptions. He called it "verstehendes Erklären" (understanding-explanation) and thought it was an adequate concept for the field of sociology when it investigates the behavior of individuals and entire societies. Weber proposed to understand societal behavior in terms of rational actors who act in order to advance their purposes ("zweckrational," purpose-rational). To know how a certain purpose relates to certain actions would require "matching" the actor's perspective; concluding that the actions would be rational in this sense—that is, would serve the purpose of interest—would explain them (in a derivation manner). Weber was fully aware that this type of action-purpose explanation is an ideal type, that is, good for orientation but never fully feasible in reality. He warned that this kind of understanding-explanation, due to its subjective components, would require a validation with causal tests.

2.3. A Pragmatic Concept of Understanding

It seems that ESM also searches for something that combines both conceptions of understanding, although in a way different from what Weber envisioned for sociology. The derivation part plays an important role, since general fundamental equations serve as a basis for ESM and they are fundamental (basic, primitive, or "core" are related terms) in the sense that important properties of the model atmosphere's circulation derive from them. This basis is complemented by parts that count as less fundamental, like parameterizations. Additionally, the model is transformed into discrete, algorithmic versions that run on a computer. The entire model assembly then goes through repeated adjustment cycles. Whether these adjustments are located on the level of discretization or on the level of some parameterization, the model assemblies are repeatedly modified via a feedback loop that compares actual (i.e., simulated) with intended model behavior (known data). In this way, model assumptions and parameter value assignments act together and can become interconnected, with choices made in some parts of the model affected by choices already made elsewhere; as a consequence, there may be a loss of some transparency regarding which aspect derives from which assumptions. Even so, modelers often can acquire "a feel" for what kinds of adjustments might be useful for attaining what kind of behavior. A similar kind of "feeling" is known from practices of artisans, technicians, and engineers when they work with instruments and tune machines (cf. Ferguson, 1992).

Such acquaintance with model behavior can be a work-around for building an adequate "inner" representation (like the chess player did) when simplification/idealization strategies are *not* available—as in the case of ESM. However, this work-around does not lead to understanding in the traditional sense. There are no simple models involved that would enable understanding by "capturing the essence of a phenomenon" (Held, 2005, 1609). Nevertheless, simulation provides understanding—if only in the weaker, pragmatic sense of getting acquainted with model behavior.

Simulation methodology matters. Before the use of computer simulation methods, theoretical principles and what derives from them were separated from pragmatic adjustments. The latter could be added but remained mostly unconnected to the principled part. Consider the work of Vilhelm Bjerknes who had formulated the Fundamental Equations in 1904, about half a century before general circulation models could be built. These equations had little use in the practice of weather prediction since there was no way in which one could derive interesting predictions from them. For his practical work, Bjerknes relied on quite independent graphical approaches. Today, with simulation, modeling can knit together principled parts with adjustments. In one and the same process, researchers are getting acquainted with model behavior AND are finding a principled motivation. The first part refers to "match" and the second part to "derivation."

2.4. An Uneasy Position for Understanding

Even if ESM can provide a pragmatic kind of understanding, it does not come with the virtues of the derivational type. In particular, because theoretical principles are only part of an ESM, phenomena understood in the pragmatic way via ESM lack the very tight connection to theoretical explanation that is possible in cases of simple analytic derivation from first principles. Most philosophers of science might be disappointed. A notable exception is van Fraassen (1980) who advocates a pragmatic concept of understanding that also does without derivational relationships. The family resemblance between the view articulated in here and that of van Fraassen is signaled by speaking of pragmatic understanding, but an exploration of differences must await a lengthier discussion.

Nevertheless, it is important to think about the limits of understanding, that is, what kind of understanding can be reached (or not) by what kind of modeling. From the perspective of philosophy, it is important to see that the concept of understanding is not historically fixed but rather affected by scientific practices. New instrumentation, importantly the computer, *affects and channels* concepts like understanding. ESM is questioning the limits of the (traditional) concept of understanding and exploring what cannot be understood in the (good) old way. When it comes to conceptions of understanding here, it might be wise to adopt the motto known from Stills (1970): "If you can't be with the one you love, love the one you're with."

To be clear, the claim here is not that pragmatic understanding should or in fact must become the goal of ESM. The case is open in an interesting way. There is a fruitful mutual relationship between ESM and PhS, since both want to find adequate ways of talking about limitations of understanding. There are limitations of established concepts that circumscribe goals (what does it mean to understand something?), and

there are limitations of methods to attain these goals (what aim makes sense given the methodology and instrumentation?). Neither is it justified to stick to all aspects of established concepts—what "understanding" means might be affected by what scientific methods can achieve—nor is it justified to freely adapt concepts, like understanding, to whatever current approaches claim as their success, else understanding would become trivial. When researchers in ESM continue to critically reflect on what should be understood but cannot and what can be understood but in an unorthodox way, they contribute to an ongoing and fruitful philosophical discussion about concepts of understanding.

3. Normative Practices in ESM (G. Gramelsberger)

Limitations of understanding in ESM, as outlined in the previous section, are an intrinsic aspect of a science based on complex models. Such limitations result from uncertainties due to the incompleteness and fallibility of knowledge (epistemic uncertainty), the intrinsically complex character of a natural system (ontic uncertainty), and the overwhelming complexity of ESM (Petersen, 2006, 2011). Nevertheless, limitations of understanding have a positive role too, in that they drive progress in scientific research with the aim of having a better understanding of models—and, through the lens of these models, of natural processes and phenomena. Modeling, as Jules Charney long since proposed, consists of a "'hierarchy of pilot problems', each of which would contain more physical, numerical, and observable aspects of the general forecast than the preceding ones" (Charney quoted in Harper, 2008, 124). Thus, in this context, scientific knowledge is per se knowledge in the making. If this is true, the interesting question is the following: How can current limitations to understanding be overcome in modeling? One possible answer is through practices of normativity.

Norms in science are the set of rules that govern how scientists do their work. The sociologist Merton (1942, p. 278) defined four fundamental norms of science: the universal nature of science and associated values such as objectivity, impersonality, and simplicity; communalism or "open science and data" in today's parlance; the disinterestedness of scientists and the concomitant scrutiny of scientific results and values like reproducibility; and, last but not least, science as the endeavor of organized skepticism—"a methodological and an institutional mandate." These are epistemic norms and values, respectively. Of course, the most basic epistemic norm of all is the search for truth or, at least, evidence. "Truth," from a philosophical perspective, is the horizon to which science tries to progress. As theories and models are semiotic representations of empirical phenomena and interrelations, including abstractions, heuristics, and purposes, they can never be "true" as a whole, but their results can be well confirmed in general and in detail, providing substantial evidence.

Merton's norms are basic norms, and, in particular, the norm of objectivity refers to the ideal that science can progress toward truth without value commitments, purposes, and community bias. However, for complex models, a tension occurs between "truth" as a norm and the intrinsic complexity leading to limitations of understanding. Against this backdrop, the above questions can be rearticulated in the following way: Which normative practices help modelers to deal with limits of understanding and in particular with modeling as knowledge in the making?

3.1. Modeling as Knowledge in the Making

If Charney is right and modeling is a process led by a "hierarchy of pilot problems" including more and more physical, numerical, and observable aspects, then a model is an eternal constructing site. In other word, models are never completed but become increasingly complex; Earth system models mark the current end of this development. However, Earth system models contain submodels and subscale parameters which are at different stages of epistemic development, for instance, oceanic carbon models and parametrizations of the Wegener-Bergeron-Findeisen (WBF) process, respectively.

Oceanic carbon models, as discussed in a paper by Le Queré (2006) entitled, "The unknown and the uncertain of Earth system modelling," are at early stages of model development, where key observational data are lacking. Although such a case is somewhat unusual, it provides some insights into modeling as knowledge in the making. Le Queré (2006) poses the question: Against what are these models evaluated when observational data are lacking?, And the answer she gives is: "Thus, the first few published [model] results initially take the place of the truth. [...] Although these results may be correct because the physical dynamics has been constrained in part by observations, the biological efficiency remains unexplored" (p. 496). "Truth"

in this sense is a somewhat crude and temporary practice of normativity, until it is replaced by reliable observational data. However, the important contribution of such a modeling phase is that models can inspire new measurement campaigns.

Not surprisingly, the next stage in development is guided by the collection of observational data. These data do not necessarily confirm model results but often contradict them, thus ushering in the "chaos phase"—the most *creative* phase of modeling. Often, model refinement and the progress of knowledge result when evidence from observations contradicts proposed models. Thus, modelers are challenged to explore new possibilities, concepts, and methods. In this phase, Le Queré elaborates, "models tend to go their own way, leading to some incoherence between model results and the state of knowledge" (ibid.). Examples of models at this phase include terrestrial carbon models. Modeling here serves as an epistemic driver to develop a better understanding of the observational findings, which are seen as "true" values—the horizon to which modeling tries to progress.

Finally, a further phase follows, "when the basic concepts have been understood and included in models, and when reliable observations can be used to eliminate outlier model results" (ibid.). Global climate models are examples from this phase. In this phase, uncertainty is reduced by increasing the amount of observational data and by efforts to fine-tune the processes in the models. Now, models can be used for prognosticating, and these prognoses can be evaluated against observational data, which is the standard practice in science. A model at this stage of development is "good" if its results perform well, that is, accord with observational data.

Nevertheless, in the development sketched here, it is not accordance with "truth" but rather incoherence among models and modeling results which becomes fruitful for gaining new knowledge. Engaging with and resolving this incoherence corresponds to Merton's norm of science as the endeavor of organized skepticism.

3.2. What Does "Performing Well" Mean?

How can modelers tell whether their models are "good"? The primary indicator is model performance, a good fit to data. However, even if true values do exist, does "evidence" of a model performing well automatically mean that the model is "true"? Of course, no such transfer of "truth" from data to models occurs. Model results will never match the parameter values describing a system's behavior exactly (Gramelsberger & Feichter, 2011; Lenhard, 2011), but a "good" model is believed to represent the relevant processes of a natural system well. There are thus at least two senses in which models can be "good," as outlined below.

On the one hand, as Charney aptly outlined, developing a "good" model involves improving process understanding through grounding the modeling more soundly in physics. For subscale parametrizations in particular, such as cloud microphysics, improvement in this sense is gained through better representation of physics. However, improvement in process understanding does not necessarily imply improved model performance. A crudely or even incorrectly implemented representation of a process such as the WBF mechanism in cloud physics can perform well within an ESM, and improving WBF representation does not necessarily improve model performance. Such improvement in representation reflects process understanding and, subsequently, can advance understanding of the interplay of model parts and mechanisms. In turn, this understanding can lead to further model development that improves model performance.

On the other hand, leaving the discussion of tuning aside (Mauritsen et al., 2012; Stainforth et al., 2005), "good" refers to empirical accuracy (model results fit observation-based data) and robustness (agreement with other models) (Baumberger et al., 2017). In recent decades, climate science has developed an impressive set of practices for evaluating models. Model experiments contributing to the Assessment Reports of the Intergovernmental Panel on Climate Change, for example, are expected to follow specific practices of evaluation like reporting performance metrics on the system, component, and parameter levels; participating in model intercomparison; and validation of ensemble prognoses.

However, ESMs deal with a wide range of climate variables. These variables are sometimes difficult to measure, requiring complex data modeling, as in the cases of satellite data and reanalysis data (Edwards, 2010). Such data, with their associated uncertainties, may seem to challenge the normative practice of comparing to "true values," but only insofar as data are understood in a naïve way as directly

reflecting true values (Lloyd, 2012); all data are samples and can only be as accurate as the instruments that measure them. A further problem is that models require homogeneously distributed observational data, which do not exist, covering long time periods. However, many practices have been developed to deal with this problem and to standardize data sets, that is, to articulate data norms. Norms for data interpolation, for proxy data, and for reanalysis data are good examples of generating standardized data sets for the entire community (Edwards, 2010). Such "data norms" or "data standards" are necessary conditions for model intercomparison and model evaluation.

3.3. The Unique Practice of Sensitive Wording in Climate Science

Finally, there is a unique way in which climate science manages the epistemic challenges that are inherent in the use of complex models to study a complex system: through the normative practice of sensitive wording. The development of this normative practice is rooted in the sociopolitical expectations placed on climate science to produce information that can be trusted in decision making. Hardly any other discipline has developed such a serious commitment to sensitive wording, motivated by ethical considerations. This commitment is most apparent in the Intergovernmental Panel on Climate Change Assessment Reports. Among other ways, it manifests in the use of the term "projection," rather than "prediction" or "forecast" (Bray & von Storch, 2009) and in the introduction of a "likelihood language" to describe scientific uncertainties, prescribing standardized terms such as "likely" or "very likely" according to definitions based on a probabilistic scale (Landström, 2017; Moss & Schneider, 2000).

Sensitive wording is also important when considering concepts like "truth," "true values," "verification," and even "validation." In a landmark paper, Oreskes et al. (1994) argued for the normative practice of sensitive wording about model evaluation, avoiding the term "verification." Numerical models in Earth science are too complex to be verified, they argued. The same holds for "validation," often used synonymously with verification; its use is similarly misleading, as "the term valid might be useful for assertions about a generic computer code but is clearly misleading if used to refer to actual model results in any particular realization" (Oreskes et al. 1994, p. 642). Their analysis recommends the terminology of "confirmation" in model evaluation instead—though, as argued by Parker below, sensitive wording may be required even when speaking of confirmation. Another term could be "reliability." Petersen (2006) differentiated between two notions of reliability: the reliability of a model result for a specific domain and the reliability of a model related to its methodological quality. (These notions are distinct, of course, from reliability in the sense of calibrated probabilistic forecasts.) However, both "confirmation" and "reliability" are relational attributes, which do not connote that the truth of a model has been established, as "verification" and "validation" may seem to do. "Confirmation" and "reliability" allow for appreciating that modeling is knowledge in the making and that science, first and foremost, is led by intellectual curiosity—a value Merton forgot to mention in his list.

4. Evaluating Models: An Adequacy-for-Purpose View (W. Parker)

Alongside efforts to develop, improve, and understand ESMs, we need to evaluate them. How should the task of model evaluation be conceptualized? Here is one intuitively appealing starting point, already touched upon in the last section: A model is a "good" model of a real system just to the extent that it *accurately represents* that system. Model evaluation is then understood as an activity that seeks to learn how accurately a model represents a real system. A standard approach would be to compare the model's assumptions and predictions to observations of the system, documenting the degree of model-data fit in various respects. Model improvement occurs, on this way of thinking, when there is an increase, in some *overall* sense, in the representational accuracy of the model. In the ideal limit, one imagines arriving at a model that is a perfectly accurate and complete representation of the real system. (see, e.g., Teller, 2001, discussing the "Perfect Model Model" and Knuuttila, 2011, p.267, on "the idea that scientific representation should aim for as accurate a representation as possible.")

An alternative perspective sees models as *tools* that scientists wish to use for particular predictive, explanatory, and other purposes (Currie, 2018; Giere, 2004; Knuuttila, 2011; Morrison & Morgan, 1999). Model evaluation is conceptualized as an activity that seeks to learn whether a model is *adequate* for one or more *purposes of interest* (for related views, see Caswell, 1976; Rykiel, 1996; NRC, 2007; Parker, 2009; Baumberger et al., 2017). The evaluator focuses on whether the model represents the real system *sufficiently accurately* in those respects that are *relevant* for the achieving the purpose(s). On this "adequacy-for-

purpose" view, model quality itself is purpose relative. Similarly, model improvement occurs relative to some range of applications, with priority placed on developing models in ways that serve those applications.

These two perspectives differ in numerous respects. Yet this does not mean that one must be chosen as the "correct" way to think about model evaluation; each view—and indeed others as well—might be advantageous in some circumstances. (Some additional views are the verification and validation framework of Oberkampf & Roy, 2010; the informal Bayesian approach suggested by Schmidt & Sherwood, 2015; and the possibilist approach advocated by Katzav, 2014.) Increasingly, however, ESMs and other climate models are being called upon to provide specific information in support of decision making, suggesting that an adequacy-for-purpose perspective will often be an attractive option in this context; what is of interest in such cases is not how accurately an ESM represents the world in some overall sense, but whether it can provide the specific information that is sought. In what follows, an adequacy-for-purpose view is explored in more detail, and some implications of the view are highlighted.

4.1. Further Fundamentals of an Adequacy-for-Purpose View

In order to employ an adequacy-for-purpose view, one needs to have some idea of what it means for a model (or any other tool) to be adequate-for-purpose. Upon reflection, it becomes clear that many different notions of adequacy are possible. A model might be adequate for a purpose P *in principle*, in the sense that if it were used in just the right way and in favorable circumstances, then P could at least sometimes be achieved. More often, scientists will be interested in whether a model is adequate for a purpose *in practice*, that is, given the way the model actually has been or will be used. Even then, it is useful to distinguish further varieties of adequacy. Sometimes, what is of interest is whether a model is adequate for purpose P in a given *instance*, that is, whether the use of the model in a particular instance did or will result in P. A model that is used to correctly predict the occurrence and duration of a particular heat wave is adequate (in this sense) for that predictive purpose, even if in many other cases it fails to correctly predict that a heat wave is coming. Perhaps more often, scientists are interested not in a particular instance of use, but in whether, in a given *type* of use of the model, P will *very often* be achieved. They want to know whether, given the way they actually will be using their model (e.g., as an element in a particular forecasting system), they will very often succeed in predicting heat waves. (This can also be expressed in terms of forecast *skill*.) The point is that in order to evaluate a model's adequacy-for-purpose—and to avoid misunderstandings—one needs to be clear about which variety of adequacy is of interest.

Note that, on the conceptions of adequacy-for-purpose just articulated, a model (and indeed any other tool) is adequate for a purpose *in a particular instance or type of use*. A model can be adequate for a purpose P in one instance or type of use but not in another, due to differences in user, methodology, or circumstances. Continuing with the heatwave example, the model might fail to be adequate-for-purpose (in the instance or type sense) if it is used in conjunction with a data assimilation scheme—an aspect of methodology—with particular limitations; with a different data assimilation scheme, the same model might be adequate. (As this suggests, it might be preferable to take the forecast system as a whole to be the tool whose adequacy is evaluated. Likewise, sometimes scientists will be interested in the adequacy-for-purpose of a set of models, for example, the adequacy of an ensemble for revealing the full range of outcomes that are plausible.) The relevance of the user is easiest to see for pedagogical and explanatory purposes: Whether a model is adequate (in the instance or type sense) for developing a correct explanation of a particular phenomenon can depend not only on whether the model represents sufficiently well the processes that give rise to the phenomenon and on what kind of experiments are performed on the model (e.g., turning processes off to see the effects) but also on the user's background knowledge and reasoning abilities. Evaluation of a model's adequacy-for-purpose in practice will require attention to these broader considerations—the methodology, the user, and the circumstances of use—not just how the model represents the world.

It is also important to understand how concepts like truth, confirmation, and confidence can play a role under an adequacy-for-purpose view. On this view, model evaluation seeks to learn whether it is *true* that a model is adequate for one or more purposes of interest. *Confidence* that a model is adequate (or not) for a purpose of interest can stem from what is known about how the model was constructed, as well as from how the model is found to perform in various tests, for example, against relevant observational data or other models (Baumberger et al., 2017). Such tests can sometimes be said to *confirm* (i.e., provide some support for) or *disconfirm* (i.e., provide some evidence against) hypotheses about a model's adequacy for particular

purposes. To return to the heatwave example, if the forecast model shows little skill in predicting the occurrence of heat waves when tested on past data (e.g., in hindcast mode), this disconfirms the hypothesis that the model is adequate (in the type sense) for very skillfully forecasting future heatwaves. One might also seek to test and confirm/disconfirm other sorts of hypotheses that bear on conclusions about a model's adequacy, for example, the hypothesis that the model has at least a moderately accurate representation of a particular physical process that is crucial to the model's being used successfully for the purpose at hand. What one should not claim to confirm is the model as a whole; this would make little sense, as it is usually known from the outset that the model incorporates some assumptions that are false (and perhaps not even approximately true). Here, it becomes apparent how sensitive wording is required even when employing the concept of confirmation (see also section 3.3).

Unfortunately, while it is relatively clear in the heatwave example whether the model's performance on past data is evidence for or against its adequacy-for-purpose, in other cases, it can be unclear. Complicating factors include—among other things—model tuning, limited information about observational uncertainties, and purposes that require simulating the behavior of the real system under boundary conditions different from those for which data are available (as in climate change projection; see also Parker, 2018; Schmidt & Sherwood, 2015, p.156). When it is unclear whether there is good evidence that a model is adequate for a purpose of interest, however, it is worth considering whether there is some related purpose for which the evidence of model adequacy (or inadequacy) is clearer. For example, it might be unclear whether today's climate models are adequate for projecting future climate with a specified level of accuracy, but there might be good reason to think that they are adequate for providing *plausible* projections (Parker, 2009). It is also possible, of course, that a model will be adequate for purposes that go beyond those for which it was originally expected or hoped to be adequate. This might be discovered in the ordinary course of examining model behavior or when surprising modeling results turn out to be largely correct, giving rise to new lines of inquiry.

4.2. Some Implications for Practice

Adopting an adequacy-for-purpose view has numerous implications for practice. It was already noted above, for example, that evaluating adequacy-for-purpose in practice requires considering not just how a model represents a real-world system but also the methodology in which the model will be embedded (e.g., a broader forecast system), the background circumstances in which it will be used, and sometimes even properties of the model user(s). Three additional implications will be discussed here.

A first implication concerns the construction and selection of *performance metrics* in model evaluation. On an adequacy-for-purpose view, evaluators should aim to tailor metrics of performance to the purpose of interest. That is, they should try to identify metrics that will be most informative about the model's adequacy for that purpose, giving greater attention in to performance on those that are thought to be most relevant. The selection of relevant performance metrics will often rely on "process understanding," especially understanding of which processes in the system strongly shape the behavior or phenomenon of interest (Baumberger et al., 2017; Eyring et al., 2019; Herger et al., 2018). It is important to keep in mind, however, that for many purposes, the evaluation of a model's adequacy-for-purpose should consider more than just performance: the model's resolution, which simplifications and idealizations it includes, how and to which data it has been tuned, and so on, can all be relevant considerations. Put differently: Even a tailored metric of *performance* is not necessarily a metric of *quality-for-a-purpose*.

A second implication concerns the interpretation of *model uncertainty*, including structural and parameter uncertainty. Rather than thinking of model uncertainty as uncertainty about what would constitute a perfect model, one thinks of it as uncertainty about which model structure(s) and parameter values would make for an adequate model in a given instance or context of use. In some situations, for example, when conducting perturbed physics ensemble studies, it will be helpful to think of parameter uncertainty as uncertainty about the parameter values that will give the *best* results for output quantities of interest, given the chosen model structure. Sampling this uncertainty and then propagating it via the ensemble will give an estimate of uncertainty about the best results that the model can give for those output quantities; to reach conclusions about the behavior of the real system, one will need to consider how far from the truth even the best results from the model might be (see, e.g., Sexton et al., 2012). The value of a given parameter that leads to the best results for a given purpose, of course, might not be the true value (if such a true value exists), given errors elsewhere in the model. Moreover, the value of a parameter that leads to the best results could differ from one purpose

to the next. This might happen, for instance, if the different values help to compensate for different errors elsewhere in the model, where some of those errors matter more for some purposes and others matter more for other purposes.

A third implication concerns the interpretation of past successes of a model. It is tempting to think that a model accrues credit or merits increased confidence in a general way as it accumulates successes in use. In a limited sense, this can be right: the broader the range of successes of a model, the more confident one can be that the model's equations are at least approximately capturing the physical processes that drive the system's behavior, at least on some spatiotemporal scales. An adequacy-for-purpose view, however, urges evaluators to consider explicitly what the *particular* successes (and failures) achieved thus far indicate—if anything—about the adequacy of the model for the purpose at hand. It may be that the model's past successes, even if impressive, actually give one little reason to be confident that it is adequate for the purpose at hand, for example, because the latter requires simulating with sufficient accuracy a somewhat different set of physical processes, or simulating smaller-scale details, than were required for those past successes.

5. Concluding Remarks

Earth system models and other computer simulation models play a range of important roles in climate science—and in many other fields—but also raise challenges that are of both scientific and philosophical interest. The focus of this brief discussion has been on a set of interconnected challenges related to understanding and model evaluation. Here we have only scratched the surface of these issues, which merit further discussion by philosophers and scientists alike. We are grateful for the opportunity to have started a conversation on these issues at the 4ICESM conference and look forward to future exchanges.

Acknowledgments
J. L. gratefully acknowledges funding by DFG SPP 1689.

References

Bangu, S. (2017, Scientific explanation and understanding: Unificationism reconsidered. *European Journal for Philosophy of Science*, *7*(1), 103–126. https://doi.org/10.1007/s13194-016-0148-y

Baumberger, C., Knutti, R., & Hirsch Hadorn, G. (2017). Building confidence in climate model projections: An analysis of inferences from fit. *WIREs Climate Change*, *8*(3), e454. https://doi.org/10.1002/wcc.454

Bony, S., Stevens, B., Held, I., Mitchell, J. F., Dufresne, J.-L., Emanuel, K. A., et al. (2013). Carbon dioxide and climate: Perspectives on a scientific assessment. In G. R. Asrar, & J. W. Hurrell (Eds.), *Climate science for serving society: Research, modeling and prediction priorities*, (pp. 391–413). Dordrecht: Springer.

Bray, D., & von Storch, H. (2009). 'Prediction' or 'Projection'? The nomenclature of climate science. *Science Communication*, *30*, 534–543. https://doi.org/10.1177/1075547009333698

Caswell, H. (1976). The validation problem. In B. C. Pattern (Ed.), *Systems analysis and simulation in ecology*, (Vol. IV, pp. 313–325). Academic Press.

Currie, A. (2018). From models-as-fictions to models-as-tools. *Ergonomics*, *4*(27), 759–781. https://doi.org/10.3998/ergo.12405314.0004.027

De Regt, H., Leonelli, S., & Eigner, K. (2009). *Scientific understanding: Philosophical perspectives*. Pittsburgh: Pittsburgh University Press.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge: MIT Press.

Eyring, V., Cox, P. M., Flato, G., Gleckler, P. J., Abramowitz, G., Caldwell, P., et al. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*(2), 102–110. https://doi.org/10.1038/s41558-018-0355-y

Ferguson, E. (1992). *Engineering and the mind's eye*. Cambridge: MIT Press.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, *71*(5), 742–752. https://doi.org/10.1086/425063

Gramelsberger, G., & Feichter, H. (Eds) (2011). *Climate change and policy. The calculability of climate change and the challenge of uncertainty*. Berlin, Heidelberg: Springer.

Harper, K. C. (2008). *Weather by the numbers. The genesis of modern meteorology*. Cambridge: MIT Press.

Held, I. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, *86*, 1609–1614. https://doi.org/10.1175/bams-86-11-1609

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., & Sanderson, B. M. (2018). Selecting a climate model subset to optimize key ensemble properties. *Earth System Dynamics*, *9*, 135–151. https://doi.org/10.5194/esd-9-135-2018

Katzav, J. (2014). The epistemology of climate models and some of its implications for climate science and the philosophy of science. *Studies in History and Philosophy of Modern Physics*, *46*, 228–238. https://doi.org/10.1016/j.shpsb.2014.03.001

Knuuttila, T. (2011). Modeling and representing: An artifactual approach. *Studies in History and Philosophy of Science A*, *42*(2), 262–271. https://doi.org/10.1016/j.shpsa.2010.11.034

Landström, C. (2017). Tracing uncertainty management through four IPCC Assessment Reports and beyond. In M. Heymann, G. Gramelsberger, & M. Mahony (Eds.), *Cultures of prediction in atmospheric and climate science*, (pp. 214–230). London: Routledge.

Le Queré, C. (2006). The unknown and the uncertain of earth system modelling. *Eos*, *87*, 499–450. https://doi.org/10.1029/2006eo450007

Lenhard, J. (2011). Artificial, false, and performing well. In G. Gramelsberger (Ed.), *From science to computational sciences*, (pp. 165–176). Diaphanes: Zurich, Berlin.

Lenhard, J. (2018). Holism or the erosion of modularity—A methodological challenge for validation. *Philosophy of Science*, *85*(5), 832–844. https://doi.org/10.1086/699675

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 Part A*, *43*(2), 390–401. https://doi.org/10.1016/j.shpsa.2012.02.001

Manabe, S. (2006). *Early development of climate modeling and the prospect of the future. Talk given at workshop "Dealing with Uncertainty. Simulation, Evaluation and Public Communication"*. Berlin-Brandenburg Academy of Science. organized by G. Gramelsberger

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., et al. (2012). Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, *4*(3), M00A01. https://doi.org/10.1029/2012ms000154

Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legal and Political Sociology*, *1*, 115–126. (Republished as "The Normative Structure of Science" in Merton, R. K. (1973), *The sociology of science. Theoretical and empirical investigations*, Chicago: University of Chicago Press, 267-278.

Morrison, M., & Morgan, M. S. (1999). Models as mediating instruments. In M. S. Morgan, & M. Morrison (Eds.), *Models as mediators*, (pp. 10–37). Cambridge University Press.

Moss, R. H., & Schneider, S. H. (2000). Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment and reporting. In R. Pachauri, T. Taniguchi, & K. Tanaka (Eds.), *Guidance papers on the cross cutting issues of the Third Assessment Report of the IPCC* (pp. 33–51). Geneva: World Meteorological Organization.

National Research Council (NRC) (2007). *Models in environmental regulatory decision making*. Washington, DC: The National Academies Press.

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

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the Earth sciences. *Science*, *263*(5147), 641–646. https://doi.org/10.1126/science.263.5147.641

Parker, W. S. (2009). II—Confirmation and adequacy-for-purpose in climate modelling. *Aristotelian Society Supplementary Volume*, *83*(1), 233–249. https://doi.org/10.1111/j.1467-8349.2009.00180.x

Parker, W. S. (2014). Simulation and understanding in the study of weather and climate. *Perspectives on Science*, *22*(3), 336–356. https://doi.org/10.1162/posc_a_00137

Parker, W. S. (2018). Climate science. In *Stanford Encyclopedia of Philosophy*, (Summer 2018 ed.. Available at:). https://plato.stanford.edu/entries/climate-science/

Petersen, A. C. (2006). *Simulating nature: A philosophical study of computer-simulation uncertainties and their role in climate science and policy advice*. Apeldoorn, Antwerpen: Het Spinhuis Publishers.

Petersen, A. C. (2011). Climate simulation, uncertainty, and policy advice—The case of the IPCC. In G. Gramelsberger, & H. Feichter (Eds.), *Climate change and policy*, (pp. 91–112). Berlin, Heidelberg: Springer.

Rykiel, E. J. (1996). Testing ecological models: The meaning of validation. *Ecological Modelling*, *90*, 229–244. https://doi.org/10.1016/0304-3800(95)00152-2

Schmidt, G. A., & Sherwood, S. (2015). A practical philosophy of complex climate modelling. *European Journal for Philosophy of Science*, *5*(2), 149–169. https://doi.org/10.1007/s13194-014-0102-9

Sexton, D. M. H., Murphy, J. M., Collins, M., & Webb, M. J. (2012). Multivariate probabilistic projections using imperfect climate models part I: Outline of methodology. *Climate Dynamics*, *38*, 2513. https://doi.org/10.1007/s00382-011-1208-9

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., et al. (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, *433*(7024), 403–406. https://doi.org/10.1038/nature03301

Stills, S. (1970): *"Love the one you're with" is part of the album "Steven Stills"*. Stills credits the musician Billy Preston for the quoted line.

Teller, P. (2001). Twilight of the perfect model model. *Erkenntnis*, *55*, 393–415. https://doi.org/10.1023/a:1013349314515

Van Fraassen, B. (1980). *The scientific image*. Oxford: Oxford University Press.

Von Wright, G. H. (1971). *Explanation and understanding*. London: Routledge and Kegan Paul.

Weber, M. (1913). Über einige Kategorien der verstehenden Soziologie. In J. Winckelmann (Ed.), *Gesammelte Aufsätze zur Wissenschaftslehre,* 1988. Mohr Siebeck: Tübingen.