15
MODEL EVALUATION

Wendy S. Parker

1. Introduction

Assessment of model quality occurs informally throughout the model development process. For instance, when constructing a model, the aim is not to produce just any model of the target system but to produce a good model, and this informs the choices made. Model evaluation, however, is also frequently identified as a distinct step in model development, occurring after a model has been fully constructed. It is this formal evaluative step in model development that will be the focus of the present chapter.

In several scientific and engineering domains, there has been extensive discussion of appropriate terminology and methods to employ in model evaluation, in some cases resulting in official guides for evaluation under the auspices of professional societies (e.g., AIAA 1998). In many other modeling contexts, however, conceptual frameworks and standards of practice for model evaluation are not articulated explicitly, and evaluation activities are only selectively reported. This can make it difficult for individuals not directly involved in the evaluation process to interpret evaluative claims (e.g., that a model is "credible") or to have a sense of the strength of evidence that underlies those claims.

The topic of model evaluation has received relatively little attention from philosophers of science. An influential contribution by Oreskes et al. (1994) called attention to the limits of what can be learned in model evaluation. Teller (2001) emphasized the purpose-relativity of model quality, understood as relevant similarity (see also Cartwright 1983; Giere 1988). More recently, Weisberg (2013) has offered an account of model-target similarity intended to facilitate the evaluation of scientific models, and Parker (2020) has advocated for an adequacy-for-purpose approach to model evaluation. A number of other contributions have emerged as a byproduct of work on the epistemology of computational modeling (e.g., Winsberg 1999; 2010; 2018; Lloyd 2010; Lenhard and Winsberg 2010; Jacquart 2016). Very recently, a massive volume edited by Beisbart and Saam (2019), *Computer Simulation Validation*, brings together both philosophical and scientific perspectives on the evaluation of computational models and constitutes a major addition to the literature.

The present chapter situates existing work within a general philosophical discussion of model evaluation.[1] Section 2 addresses a foundational question: what does it mean for

DOI: 10.4324/9781003205647-19

a model to be a good model? Three common answers are presented: quality as accurate and comprehensive representation, quality as relevant similarity, and quality as fitness-for-purpose. Section 3 considers the task of model evaluation from the perspective of each of these three conceptions of model quality and discusses allied approaches to evaluation that have been advocated by scientists and philosophers. Section 4 outlines several obstacles and challenges that can arise when performing model evaluation, which can prevent evaluators from reaching confident conclusions about model quality. Finally, Section 5 summarizes key points and identifies some directions for future research.[2]

2. Models and model quality

Assessment of the quality of a scientific model depends, at least implicitly, on some conception of model quality, i.e., of what constitutes a good model. This section presents three common conceptions of model quality, which are associated with different views of what scientific models are: quality as accurate and comprehensive representation, associated with a view of models as representations; quality as relevant similarity, associated with a view of models as representational tools; and quality as fitness-for-purpose, associated with a view of models as tools or artifacts, not necessarily representational.

What will here be called the *mirror view* of model quality is only sometimes explicitly espoused, but it seems implicit in much modeling practice (see also Saltelli et al. 2020). In this view, a model is a representation, and it is of higher quality *the more accurately and comprehensively it represents its target system*. The hypothetical limit is a model that mirrors the target system, in the sense that every element (part, property, relationship) of the target system is represented by a corresponding element in the model, and with perfect accuracy.[3] Increasing the comprehensiveness of a model by adding a representation of a target system process that was previously unrepresented, or increasing the fidelity with which some feature of the target system is represented, will count as improving the model on the mirror view, regardless of the purposes for which the model will be used. Conversely, idealizations, distortions, and omissions in representation necessarily detract from model quality on this view, regardless of the purposes for which the model will be used.

On many other views of model quality, however, the intended use of the model *is* relevant to the assessment of model quality. In the philosophy of science, a prominent view is that model quality is a matter of *relevant similarity*: a good model is *similar enough to its target in the relevant respects*, where the relevant respects are determined by the model user's purpose (Giere 1988; 2004; Teller 2001; Weisberg 2013). A closely related view is expressed in terms of representation: a good model represents its target system with sufficient fidelity in the relevant respects, given the modeler's purpose. This way of thinking of model quality is associated with a view of models as representational tools: they are representations, intended to be useful for particular purposes (e.g., predicting X with specified accuracy, explaining Y).

If model quality is a matter of relevant similarity (or relevant representational fidelity), then idealizations, distortions, and omissions in modeling do not necessarily detract from model quality; it depends on whether they render the model dissimilar to its target in ways that impede achieving the purposes of interest. Indeed, idealizations, distortions, and omissions can even enhance the quality of a model in many cases, insofar as the resulting model represents the target system in a way that better serves the purpose of interest (see also Bokulich 2013; Potochnik 2018). For example, "artificial viscosity" in fluid simulations is a distortion that allows for a more accurate prediction of the evolution of shock waves (see Winsberg and Mizra 2017 for more examples). Likewise, if the aim is to learn whether a particular causal process plays an important role in producing a phenomenon, it might be advantageous for a computer simulation model to omit that process (while representing other contributing processes with sufficient fidelity) in order to reveal how the phenomenon changes, if at all, when the process is absent.

A third perspective on model quality is closely associated with an understanding of scientific models as tools or artifacts (Caswell 1976; Beck 2002; Knuuttila 2005; 2011; NRC 2007; Boon and Knuuttila 2009; Currie 2017). On this *fitness-for-purpose* view, a model is a good model to the extent that it *has properties that make it a suitable tool for the task at hand*. These properties will often include more than representational properties – properties like manipulability, computational tractability, cognitive accessibility, and so on, can contribute to a model's quality. Moreover, whether a model has such properties can vary with the context of the use, i.e., with the model user, with the methodology employed, and with the background conditions in which the use of the model will occur. For example, a model might be computationally tractable for a user who has access to a supercomputer, but not for a user who has only an ordinary desktop computer. The fitness-for-purpose of a model thus can vary with the context of use (Parker 2020).[4]

As with the relevant similarity view, idealizations, simplifications, and omissions need not detract from the model's quality on a fitness-for-purpose view and are sometimes advantageous. Here, however, they can be advantageous not only for reasons having to do with how the model relates to a target system but also for reasons having to do with how the model relates to model users and other features of the context of use. For example, compared to a complex, hyper-realistic model, a simpler model, which omits many processes at work in the target system and represents others in an idealized way, might better facilitate understanding of a particular phenomenon, given humans' (i.e., users') cognitive limitations (see also Isaac 2013; Potochnik 2018). Indeed, such a view regarding the value of simple models for purposes of understanding is frequently expressed in the study of complex systems.

3. Model evaluation

The aim of model evaluation is to learn about model quality, whether quality is conceptualized as accurate and comprehensive representation, relevant similarity, fitness-for-purpose, or in some other way.[5] Put differently, model evaluation activities are directed at obtaining evidence regarding hypotheses of interest about model quality, such as the hypothesis that the model is similar enough to the target in the relevant respects, given the modeling purpose of interest. This section considers the task of model evaluation from the perspective of each of the views of model quality introduced in Section 2 and discusses allied approaches to model evaluation that have been advocated by scientists and philosophers. Throughout, the analysis attends to two complementary sources of evidence regarding model quality: evidence related to the model's *composition*, i.e., its ingredients and how they are put together, and evidence related to the model's *performance*, i.e., its behavior or output.[6] Although it will not be emphasized below, it is important to recognize that evaluation is typically an iterative process: what is learned when evaluating a model often leads to further adjustments to the model, after which the new version of the model is evaluated, and so on.[7]

Mirror view. From the perspective of the mirror view, model evaluation is an activity that seeks to learn to what extent a model accurately and comprehensively represents a target system. When examining a model's composition, the mirror-view evaluator will be interested in whether any elements of the target system are omitted from (i.e., not represented at all in) the model as well as how closely, from the perspective of theoretical and other background knowledge, the elements of the model come to perfectly representing the corresponding elements of the target system. For example, the evaluator of a mathematical model of an ecosystem might note that some species in the ecosystem have not been represented at all in the model and that interactions among other species have been represented in a quite simplistic way relative to what is known about those species' interactions; this will be judged to detract from the model's quality.

When examining model performance, the mirror-view evaluator will be interested in how closely the behaviors of the model resemble those observed for the target system in corresponding circumstances. For mathematical models and computer simulation models, this typically will involve comparing the values of model variables to observational data. Assessing the fit between model results and observational data is considered a crucial part of the evaluation of such models regardless of the conception of model quality adopted. For the mirror-view evaluator, output for *every* model variable (and combinations/aggregations of such variables) for which observations are available will in principle be of interest, since any such model-data comparison can provide some (indirect) evidence regarding the extent to which the model accurately and comprehensively represents the target system. Performance scores for individual variables might even be averaged or otherwise aggregated to produce an indication of "overall" performance.

In scientific publications, evaluative discussions of computational models sometimes are strongly suggestive of a mirror view of model quality. Reasons are given for thinking that a model is a "credible" representation of the target system in some general or overall sense. For instance, it might be reported that a model "includes" (i.e., includes some representation of) many target system processes, that the model's core equations are grounded in established theory, and that the model achieves a relatively good fit with available observational data across a range of output variables. In some cases, this approach to model evaluation might reflect a simple commitment to a mirror view of model quality. In other cases, however, it may be intended as a kind of "purpose-neutral" evaluation, motivated by the expectation that the model will be used for a wide range of (perhaps yet-to-be-fully-specified) purposes.[8] Either way, from the fact that a model is "credible" in this general or overall sense, it does not follow that any particular results from the model will be accurate, since a model that represents a target system reasonably well in some overall sense might represent relatively poorly the aspects that matter for a specific question or task.

Relevant similarity view. Unlike the mirror-view evaluator, the evaluator of relevant similarity will be interested in only some aspects of a model's composition and performance, namely, those for which sufficient similarity to the target system is needed in order for the model to serve the purpose of interest. For example, when evaluating an animal model to be used in investigating the toxicity of a chemical, the relevant-similarity evaluator might check whether a particular set of biochemical pathways operative in humans – and expected to mediate any toxic effects of the chemical – are also operative in the animal; the evaluator will not be concerned with aspects of the animal's composition that are expected to make no difference to whether it will be informative about the toxicity of the chemical in humans. Continuing with the example, when focusing on model performance, an evaluator might investigate how the response of the animal to other toxic chemicals compares to the known effects of those chemicals in humans, but many other aspects of the animal's behavior – such as whether it is quieter than humans when eating, whether it wakes up earlier in the morning than humans – are unlikely to be of interest.

The selectiveness of relevant similarity evaluation is formalized in Weisberg's (2013) weighted-feature-matching account of model-target similarity. On his account, models can be assigned a similarity score, depending on the extent to which they have specific features (attributes and/or mechanisms) that match those of the target system, where the features of interest and their relative importance are determined by the modeling purpose (e.g., predicting X, explaining how-possibly Y). What counts as a "match" between features of a model and target system on this account merits further attention, however; in general, relevant features of a model do not need to be identical to those of the target system in order for a model to serve a purpose of interest, yet what it means for features to be "sufficiently similar" is not so clear either (Parker 2015; Khosrowi 2020). A further question is whether Weisberg's account can be usefully applied in practice (Jacquart 2016).

Fitness-for-purpose view. Fitness-for-purpose evaluation seeks to determine whether a model is a suitable tool for the task at hand. In contrast to mirroring and relevant similarity evaluation, fitness-for-purpose evaluation will often need to consider more than just how a model relates to a target; it will need to consider how the model relates to the model user and other aspects of the context of use (Parker 2020). The evidence cited regarding the model's fitness-for-purpose can likewise be broader. Consider, for example, an evaluation of the fitness of a computer model for the purpose of ranking the effectiveness of various possible interventions to curb algae blooms in a given lake. Evidence that the model is fit for purpose could include not only facts about how the model represents certain biological and chemical processes in the lake but also the fact that the model has an interface that allows its users to easily adapt the model to represent the different possible interventions and the fact that the model takes only a short period of time to run on available computers.[9]

A fitness-for-purpose approach to the evaluation of models has been advocated by a number of practitioners in the earth and environmental sciences. An important early contribution comes from Caswell (1976) in the context of ecological modeling. He argues that, since models are artificial systems designed to serve particular purposes, they should be evaluated relative to their intended task environment; for some purposes, such as gaining insight or understanding, whether a model produces output that closely fits observations may be relatively unimportant. Building on this, Beck (2002) notes that environmental models are used not only for "scientific" purposes, such as making predictions or gaining understanding, but also for various "pragmatic" purposes, such as supporting decision-making, formulating public policy, or communicating scientific information to lay audiences, and he raises the question of how to evaluate the fitness of models for such pragmatic purposes. Some progress in this regard is made in a report from the U.S. National Academies of Science, *Models in Environmental Regulatory Decision Making* (2007). It develops an extensive list of considerations relevant to evaluating the fitness-for-purpose of environmental models in regulatory contexts, including considerations like model transparency to stakeholders.

Many other discussions of fitness-for-purpose evaluation, however, largely ignore the context of use of a model, focusing attention instead on how to probe whether a model represents its target system accurately enough in relevant respects to provide sought-after information. Here, the language of fitness-for-purpose (or adequacy-for-purpose) is adopted, but the evaluation is essentially concerned with relevant similarity or relevant representational fidelity. For instance, Baumberger et al. (2017) develop a framework for evaluating the fitness-for-purpose of climate models for projecting long-term changes in climate, but the potential lines of evidence that they identify – coherence with background knowledge, sufficient fit with relevant observational data, and robustness of projections across models – are of interest because they bear on whether models represent sufficiently well the causal processes that will shape the long-term evolution of climate characteristics (see also Knutti 2018 on process understanding and Kawamleh 2022 on process-based evaluation). Another example can be found in the context of hydrological modeling. Beven (2018) argues for the benefits of a falsificationist approach to fitness-for-purpose evaluation, whereby hydrological models – understood as hypotheses about how water catchments function – are tested against relevant observational data and rejected if they fail to meet pre-specified performance criteria identified in light of the modeling purpose.

Pre-specified performance criteria are also an important part of the evaluation of the fitness-for-purpose of computational models in engineering contexts. Here, it is well recognized that the fitness-for-purpose of a model can depend on more than how it represents a target system: computational demands, adaptability, ease of use for model users of a given experience level, etc., can all be relevant (Oberkampf and Roy 2010, 37). Nevertheless, the core of model evaluation is often conceptualized as consisting of two activities: *verification* investigates whether the model's computational algorithm delivers results that approximate closely enough the solutions of the modeling equations that have been selected; *validation* investigates whether those modeling equations represent the target system with sufficient fidelity in relevant respects for the application of interest, primarily by comparing results obtained from the computational model with observational data (see contributions in Beisbart and Saam 2019 for further discussion of these concepts and related practices). Ideally, this comparison is pursued in a systematic way such that individual model components (representing a particular process or part of the target system), and then various combinations of those components, are tested against high-quality observational data obtained from specialized validation experiments, in order to see if pre-specified levels of accuracy are met, where those levels of accuracy are determined by the model application (Oberkampf and Roy 2010). Though verification and validation are often conceptualized as distinct activities, Winsberg (2010; 2019) argues that in practice they are not so neatly separable (see also Lenhard 2018; 2019; and further discussion by Beisbart 2019a).

Evidence synthesis. Regardless of the conception of model quality that is adopted, evaluators may also wish (or be expected) to provide some summary judgment or conclusion about model quality. Doing so in effect involves a kind of evidence synthesis, where the evidence consists of what has been learned about model composition and/or performance. How to perform this synthesis, and when evidence is sufficient to warrant various conclusions about model quality, are complicated matters. Not infrequently, practitioners seem to adopt a kind of informal Bayesian perspective (Schmidt and Sherwood 2015), where particular findings about model composition or performance – such as the finding that the model's results for a given variable closely track observations – are taken to confirm or disconfirm (and thus build or reduce confidence in) a hypothesis about model quality, e.g., the hypothesis that the model is fit for a particular purpose or is a credible representation of the target system (see also Baumberger et al. 2017; Beisbart 2019b; Gelfert 2019).

A quite different sort of approach involves specifying criteria in advance of model evaluation which, if met, will be considered sufficient to warrant a conclusion of interest about model quality. For example, Haasnoot et al. (2014, 112), evaluating a model for screening and ranking different water policy pathways, conclude that their model is fit for purpose after reaching affirmative answers to a series of questions about the model's composition and performance. Similarly, in engineering contexts, evaluators sometimes specify accuracy requirements (with respect to high-quality data from experiments) for a series of model variables, such that meeting those requirements will be sufficient to consider the model (or its results) accurate enough for its intended use. In many modeling contexts, however, it is difficult to confidently specify such a set of sufficient criteria, much less to demonstrate that they are met by a given model, in part for reasons discussed in the next section.

4. Obstacles and challenges in model evaluation

Ideally, the activity of model evaluation will deliver strong evidence regarding model quality, such that confident conclusions – e.g., that a model is fit for purpose P – will be warranted. For a number of reasons, however, confident conclusions can remain out of reach. This section surveys some of these reasons.

Limited observations of the target system. First, scientific models are often employed when, for practical or ethical reasons, target systems are inaccessible to observation and experiment under conditions of interest. As a consequence, there are limited relevant observations of the target system, which can significantly hinder model assessment. For example, assessment of the fitness of today's climate models (for the purpose of projecting future temperature change in response to rising greenhouse gas concentrations) is hindered by the fact that, during the past periods for which reliable observations of the climate system are available, greenhouse gas concentrations were lower than in the scenarios for which projections are being made. In such situations – when available data were collected under conditions quite different from those that are ultimately of interest – it can be difficult to tell what a model's performance on the data indicates about its fitness-for-purpose (Parker 2009). This is especially so when models could have been constructed in awareness of, or even partially tuned to reproduce, the available data (Baumberger et al. 2017).

Model opacity. Another obstacle is model opacity, i.e., the inscrutability or incomprehensibility of aspects of a model, including its behavior, to an evaluator (see also Humphreys 2004 on epistemic opacity). Especially when models are complex and nonlinear, they are somewhat opaque even to individuals intimately involved in their development. A relevant-similarity evaluator, for instance, may find it difficult to understand – just by observing the behavior of a complex computational model – why it behaves in a particular way and may thus be unsure what that behavior indicates about the fidelity with which the model represents relevant target system processes (Baumberger et al. 2017; see also Lenhard and Winsberg 2010 on analytic impenetrability). Opacity can be just that much greater for evaluators who were not involved in the development of a model, especially when that development involved ad hoc elements (e.g., kludging) and when the model is poorly documented, i.e., when little accompanying information is provided and/or the model code is undocumented. Such an evaluator may have a difficult time deciding where to focus their evaluation efforts and determining whether the results of model tests provide strong evidence regarding model quality. They may also be left unaware of how non-epistemic (social, political, ethical) values shaped choices in the model's development, which in some cases might be relevant to their evaluation (see, e.g., Parker and Winsberg 2018; Hirsch Hadorn and Baumberger 2019; Lusk and Elliott 2022).

Holism in assessment. Holism is a challenge that arises primarily when assessing relevant similarity or fitness-for-purpose: in many cases, what is learned about the composition or performance of a model component in isolation cannot on its own serve as evidence regarding model quality (Parker 2020; see also Lenhard and Winsberg 2010; Lenhard 2018; 2019). Suppose that the purpose of a modeling study is to predict whether applications to a university will increase or decrease in number over the next several years. Finding that a model grossly underestimates a factor that is an important determinant of application numbers might or might not be evidence that the model is not fit for purpose, depending on whether that error is sufficiently compensated for by errors elsewhere in the model or by the broader methodology in which the model is embedded (e.g., a bias correction step). Likewise, whether the degree of similarity between a component of a model and a part of a target system counts as evidence for or against a relevant similarity hypothesis (for the model overall) can depend on how similar other components are and in what ways. The fact that components of models sometimes cannot be assessed in isolation makes evaluation a more complicated task, both practically and cognitively, especially when models are complex.[10]

Quantifying quality. Further challenges arise when evaluators seek to quantify model quality, i.e., to assign each of several models a quantitative score indicative of its quality. Such scores might be used, for instance, to differentially weight results from different models or to select from a set of models the ones that are best for a given purpose. A fundamental challenge here is quantifying the contribution of model composition to model quality (see also Baumberger et al. 2017). Weisberg's (2013) weighted feature-matching approach, mentioned in Section 3, might be one way forward for relevant-similarity evaluators, insofar as its scoring procedure takes account of both mechanisms (pertaining to composition) and attributes (covering performance aspects). Yet relevant-similarity evaluators will still need to determine how to assign weights indicating the relative importance of various mechanisms, how to avoid double-counting when both mechanisms and attributes they help to bring about are among the relevant features, and more.[11]

A different approach that is sometimes employed in practice is for evaluators to limit their attention to models that, based on expert judgment, seem of at least roughly equal quality from the perspective of composition, and then assign quality scores based on performance metrics.[12] Challenges here include determining which performance metrics should be employed and how they should be combined to produce an overall quality score. Mirror-view evaluators will need to choose from a host of measures of model-data fit (root mean square error, max absolute error, etc.) for each model variable for which observational data are available and will need some method for aggregating findings across variables into an overall score. Relevant-similarity and fitness-for-purpose evaluators will, in addition to choosing among measures of model-data fit, need to identify which model variables to focus on and how to weight performance on these variables to produce an overall quality score (Knutti 2018). Typically, there will be many reasonable ways to proceed for all three types of evaluators, with different choices resulting in somewhat different assessments of the relative quality of different models. In other words, there will be uncertainty about the models' relative (and absolute) quality.

5. Concluding remarks

Model evaluation is an important part of the model development process, occurring informally even during the building of models, and more formally once they are fully constructed.

The aim of model evaluation is to learn about the quality of one or more models, whether the quality is conceptualized as accurate and comprehensive representation, relevant similarity, fitness-for-purpose, or in some other way. The conception of model quality that is adopted carries implications for the practice of model evaluation, including whether evaluation must attend to the purposes for which models are being used and whether factors other than how the model relates to a target system, such as aspects of the context of use, are relevant.

Whatever the operative conception of model quality, evidence regarding model quality can come via two complementary routes: by examining the model's composition, i.e., its ingredients and how they are put together, and by examining the model's performance, especially its performance against observations of the target system. A number of obstacles and challenges can arise in the course of gathering such evidence and attempting to reach conclusions about model quality, including limited observations of the target system, model opacity, holism in assessment, and uncertainty about how to quantify model quality. Because of obstacles and challenges like these, it is sometimes difficult to reach confident conclusions about model quality.

Many questions about model evaluation merit further attention from philosophers of science. To name just a few: How do practices of model evaluation vary across different types of scientific models and in different scientific fields? How should evidence regarding model quality be synthesized to reach conclusions about model quality? How and to what extent should non-epistemic values figure in the evaluation of scientific models? A topic that can be expected to attract attention in the near future is the evaluation of "models" produced via machine learning methods; they present an especially interesting case for philosophical analysis, given their opacity, their questionable representational status, and their increasing use in high-stakes practical applications.

Notes

[1] The discussion of existing work will – of necessity – be far from comprehensive, especially when it comes to scientific work on model evaluation. The author apologizes for omissions of important works.

[2] This chapter is concerned with the evaluation of scientific models whose targets are real systems or phenomena, such as earth's atmosphere or the spread of flu virus through a population. The evaluation of models that have only imagined/imaginary target systems will not be addressed, e.g., a model of the population dynamics of a hypothetical species with four sexes and particular mating strategies. Likewise, the evaluation of statistical/data models, which are intended to capture relationships among variables in datasets, may merit separate treatment.

[3] All elements of the target system might be represented in a model if the target system is specified such that it encompasses only a finite set of elements, e.g., particular relationships in a set of chemical reactions.

[4] A "fitness-for-purpose" view of model quality is often adopted in scientific practice today, though exactly what practitioners mean by "fitness-for-purpose", and whether they understand it to be relative to a context of use, is sometimes unclear.

[5] This is not to suggest that practitioners always have a clear and explicit conception of model quality; in some cases, for instance, evaluation proceeds in a way that simply follows what is usually done in a particular lab, community, or field.

[6] Jacquart (2016) understands relevant similarity to be a matter of a model's composition and adequacy-for-purpose to be a matter of a model's performance. This differs from the present discussion, which allows that a model's performance might make it relevantly similar to a target, a model's composition might be essential to its fitness-for-purpose, etc.

[7] Likewise, even after a model is fully constructed and put to use, it may subsequently undergo further development and evaluation. This is common, for instance, in weather and climate modeling,

and is reflected in the labels given to successive versions of a model, e.g., CESM1.0, CESM1.1, CESM1.2.

[8] Thanks to Donal Khosrowi for prompting me to consider this possible motivation and for supplying the language of "purpose-neutral" evaluation.

[9] For a similar, real example of fitness-for-purpose evaluation, see Haasnoot et al. (2014).

[10] Rice (2019) argues that many highly idealized models are "holistically distorted representations" that are "greater than the sum of their accurate and inaccurate parts." If so, then even when a mirror view of model quality is adopted, it might be misguided in some cases to assess models by examining the representational fidelity of each component in isolation and aggregating the findings. (Note, however, that Rice's analysis is not concerned with the assessment of model quality; it is intended to challenge the view that, when models are used successfully for explanation and understanding, it is because their idealized/inaccurate parts do not "get in the way" of the accurately representing parts that do the real work.) Taking an artifactual perspective, Carrillo and Knuuttila (2022) offer a view of "holistic idealizations" that downplays the idea that they are distortions and emphasizes that they "result from more systematic research programs that integrate different concepts, analogies, measuring apparatus and mathematical approaches" (50).

[11] In the context of statistical model selection, scoring criteria like the Akaike information criterion (AIC) take account of model composition by penalizing models for having more adjustable parameters; models receive a higher quality score to the extent that they can fit some set of data well with a smaller number of adjustable parameters. When it comes to models of real-world phenomena, the quality of a model's composition is usually understood to be a matter of much more than the number of adjustable parameters it contains.

[12] Note that, for fitness-for-purpose evaluators, composition will need to be evaluated taking account of the model user, methodology, and background circumstances, not just the model's target system.

References

AIAA. 1998. "Guide for the Verification and Validation of Computational Fluid Dynamics Simulations." *American Institute of Aeronautics and Astronautics*, AIAA-G-077-1998. Reston, VA.

Baumberger, Christoph, Reto Knutti, and Gertrude Hirsh Hadorn. 2017. "Building Confidence in Climate Model Projections: An Analysis of Inferences from Fit." *WIREs Climate Change* 8(3): e454.

Beck, Bruce. 2002. "Model Evaluation and Performance." In *Encyclopedia of Environmetrics,* Volume 3, edited by Abdel H. El-Shaarawi and Walter W. Piegorsch, 1275–1279. Chichester: Wiley and Sons.

Beisbart, Claus. 2019a. "Should Validation and Verification Be Separated Strictly?" In *Computer Simulation Validation*, edited by Claus Beisbart and Nicole J. Saam, 1005–1028. Switzerland: Springer.

———. 2019b. "Simulation Validation from a Bayesian Perspective." In *Computer Simulation Validation*, edited by Claus Beisbart and Nicole J. Saam, 173–202. Switzerland: Springer.

Beisbart, Claus, and Nicole J. Saam, eds. 2019. *Computer Simulation Validation*. Switzerland: Springer.

Beven, Keith J. 2018. "On Hypothesis Testing in Hydrology: Why Falsification of Models Is Still a Really Good Idea." *WIRES Water* 5(3): e1278.

Bokulich, Alisa. 2013. "Explanatory Models versus Predictive Models: Reduced Complexity Modeling in Geomorphology." In *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, edited by Vassilios Karakostas and Dennis Dieks, 115–128. Switzerland: Springer.

Boon, Mieke, and Tarja Knuuttila. 2009. "Models as Epistemic Tools in Engineering Sciences: A Pragmatic Approach." In *Handbook of the Philosophy of Science, vol. 9, Philosophy of Technology and Engineering Sciences*, edited by Antoni Meijers, 687–720. Amsterdam: Elsevier.

Carrillo, Natalia, and Tarja Knuuttila. 2022. "Holistic Idealization: An Artifactual Standpoint." *Studies in History and Philosophy of Science* 91: 49–59.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Caswell, Hal. 1976. "The Validation Problem." In *Systems Analysis and Simulation in Ecology*, vol. 4, edited by Bernard C. Patten, 313–325. Cambridge, MA: Academic Press.

Currie, Adrian. 2017 "From Models-as-Fictions to Models-as-Tools." *Ergo* 4(27): 759–781.

Gelfert, Axel. 2019. "Assessing the Credibility of Conceptual Models." In *Computer Simulation Validation*, edited by Claus Beisbart and Nicole J. Saam, 249–270. Switzerland: Springer.

Giere, Ronald N. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

———. 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71(5): 742–752.

Haasnoot, Marjolin, W. P. A. van Deursen, Joseph H. A. Guillaume, Jan H. Kwakkel, Ermond van Beek, and Hans Middelkoop. 2014. "Fit for Purpose? Building and Evaluating a Fast, Integrated Model for Exploring Water Policy Pathways." *Environmental Modelling and Software* 60: 99–120.

Hirsch Hadorn, Gertrude, and Christoph Baumberger. 2019. "What Types of Values Enter Simulation Validation and What Are Their Roles?" In *Computer Simulation Validation*, edited by Claus Beisbart and Nicole J. Saam, 961–980. Switzerland: Springer.

Humphreys, Paul W. 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press.

Isaac, Alastair M. 2013. "Modeling without Representation." *Synthese* 190: 3611–3623.

Jacquart, Melissa. 2016. *Similarity, Adequacy, and Purpose: Understanding the Success of Scientific Models*. PhD diss, University of Western Ontario. https://ir.lib.uwo.ca/etd/4129.

Kawamleh, Suzanne. 2022. "Confirming (Climate) Change: A Dynamical Account of Model Evaluation." *Synthese* 200: 122.

Knutti, Reto. 2018. "Climate Model Confirmation: From Philosophy to Predicting Climate in the Real World." In *Climate Modelling: Philosophical and Conceptual Issues*, edited by. Elisabeth A. Lloyd and Eric Winsberg, 325–359. Palgrave MacMillan.

Khosrowi, Donal. 2020. Getting Serious about Shared Features. *British Journal for the Philosophy of Science* 71(2): 523–546.

Knuuttila, Tarja. 2005. "Models as Epistemic Artefacts: Toward a Non-Representationalist Account of Scientific Representation." *Philosophical Studies from the University of Helsinki* 8: 12–78.

———. 2011. "Modeling and Representing: An Artifactual Approach." *Studies in History and Philosophy of Science A* 42(2): 262–271.

Lenhard, Johannes. 2018. "Holism, or the Erosion of Modularity: A Methodological Challenge for Validation." *Philosophy of Science* 85(5): 832–844.

———. 2019. "How Does Holism Challenge the Validation of Computer Simulation?" In *Computer Simulation Validation*, edited by Claus Beisbart and Nicole J. Saam, 943–960. Switzerland: Springer.

Lenhard, Johannes, and Eric Winsberg. 2010. "Holism, Entrenchment, and the Future of Climate Model Pluralism." *Studies in History and Philosophy of Science A* 41(3): 253–262.

Lloyd, Elisabeth A. 2010. "Confirmation and Robustness of Climate Models." *Philosophy of Science* 77(4): 971–984.

Lusk, Gregory, and Kevin C. Elliott. 2022. "Non-Epistemic Values and Scientific Assessment: An Adequacy-for-Purpose View." *European Journal for Philosophy of Science* 12: 35.

NRC (National Research Council). 2007. *Models in Environmental Regulatory Decision Making*. Washington, DC: National Academies.

Oberkampf, William L., and Christopher J. Roy. 2010. *Verification and Validation in Scientific Computing*. Cambridge: Cambridge University Press.

Oreskes, Naomi, Kristin Shrader-Frechette, and Kenneth Belitz. 1994. "Verification, Validation and Confirmation of Numerical Models in the Earth Sciences." *Science* 263(5147): 641–646.

Parker, Wendy S. 2009. "Confirmation and Adequacy-for-Purpose in Climate Modeling." *Aristotelian Society Supplementary Volume* 83: 233–249.

———. 2015. "Getting (Even More) Serious about Similarity." *Biology and Philosophy* 30(2): 267–276.

———. 2020. "Model Evaluation: An Adequacy-for-Purpose View." *Philosophy of Science* 87(3): 457–477.

Parker, Wendy S., and Eric Winsberg. 2018. "Values and Evidence: How Models Make a Difference." *European Journal for Philosophy of Science* 8(1): 125–142.

Potochnik, Angela. 2018. *Idealization and the Aims of Science*. Chicago: Chicago University Press.

Rice, Colin. 2019. "Models Don't Decompose That Way: A Holistic View of Idealized Models." *British Journal for the Philosophy of Science* 70(1): 179–208.

Saltelli, Andrea, Gabriele Bammer, Isabelle Bruno, Erica Charters, Monica Di Fiore, Emmanuel Didier, Wendy Nelson Espeland, John Kay, Samuele Lo Piano, Deborah Mayo, et al. 2020. “Five Ways to Ensure that Models Serve Society: A Manifesto.” *Nature* 582: 482–484.

Schmidt, Gavin A., and Steven Sherwood. 2015. “A Practical Philosophy of Complex Climate Modelling.” *European Journal for the Philosophy of Science* 5(2): 149–169.

Teller, Paul. 2001. “Twilight of the Perfect Model Model.” *Erkenntnis* 55: 393–415.

Weisberg, Michael. 2013. *Simulation and Similarity*. New York: Oxford University Press.

Winsberg, Eric. 1999. “Sanctioning Models: The Epistemology of Simulation.” *Science in Context* 12(2): 275–292.

———. 2010. *Science in the Age of Computer Simulation*. Chicago: Chicago University Press.

———. 2018. *Philosophy and Climate Science*. Cambridge: Cambridge University Press

———. 2019. “Computer Simulations in Science.” In *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/win2022/entries/simulations-science/

Winsberg, Eric, and Ali Mizra. 2017. “Success and Scientific Realism: Considerations from the Philosophy of Simulation.” In *The Routledge Handbook of Scientific Realism*, edited by Juha Saatsi, 250–260. London: Routledge.