Philosophy of Science (2023), **90**, 1407–1416
doi:10.1017/psa.2023.62

PSA PHILOSOPHY OF SCIENCE ASSOCIATION

PHILOSOPHY OF SCIENCE

CONTRIBUTED PAPER

Robustness of Climate Models

Stuart Gluck

U.S. Department of Energy, Office of Science, Washington, DC, USA
Email: stu@jhu.edu

(Received 08 February 2023; revised 20 March 2023; accepted 22 March 2023; first published online 04 April 2023)

Abstract

Robustness with respect to climate models is often considered by philosophers of climate science to be a crucial issue in determining whether and to what extent the projections of the Earth's future climate that models yield should be trusted. Wendy Parker and Lisa Lloyd have introduced influential accounts of robustness for climate models with seemingly conflicting conclusions. I argue that Parker and Lloyd are characterizing distinct notions of robustness and providing complementary insights. Confidence, if warranted, need be by virtue of causally consistent climate models rather than by agreement upon projections by a diverse range of models.

1. Introduction

Robustness with respect to climate models is often considered by philosophers of climate science to be a crucial issue in determining whether and to what extent the projections of the Earth's future climate that models yield should be trusted—and in turn whether society should pursue policies to address mitigation of and adaptation to anthropogenic climate change. As Pirtle et al. (2010) note, climate modelers often use agreement among multiple models as a source of confidence in the accuracy of model projections, hence the priority for philosophers to evaluate this "robustness" as an epistemic strategy.

Central to considerations about the application of robustness to climate models have been the discussions by Parker (2011) and Lloyd (2009, 2015). There seems to be confusion in the literature about how to understand each approach and especially about the relationship between them. Winsberg (2018), for example, leverages yet a third account of robustness, a more generic approach from Schupbach (2018), to try to reconcile them. Reconciliation is unnecessary. Parker and Lloyd address different issues: Parker considers the robustness of the predictive hypotheses, while Lloyd primarily considers the robustness of the models and only secondarily the status that the models thereby confer upon the hypotheses.

I argue that Parker and Lloyd provide complementary insights about related but distinct notions. These insights help illustrate how models can provide confidence in

predictive hypotheses and the conditions that must be met for model development to function as an effective research program. While both accounts are valuable, Lloyd's model robustness is more relevant for evaluating whether climate modeling is a productive research program for advancing science and for providing guidance to policy makers.[1]

2. The concept of robustness

Biologist Richard Levins introduces the concept of robustness in his (1966). Explaining that modelers of biological systems must make simplifying assumptions, he says (1966, 423):

> we attempt to treat the same problem with several alternative models each with different simplifications but with a common biological assumption. Then, if these models, despite their different assumptions, lead to similar results, we have what we call a robust theorem which is relatively free of the details of the model. Hence our truth is the intersection of independent lies.

The idea is that, while the models are literally false in that the mathematics should not be taken as precisely describing the world, when multiple models agree, it seems likely that they are tracking something accurate and important.

As Weisberg (2006) notes, "Levins' original discussion of robustness analysis provides little general characterization of the practice." Numerous scholars have thus attempted to provide more precise characterizations of the concept.[2]

3. Parker's robustness

Parker (2011, 580) introduces her goal in characterizing robustness in the context of climate modeling this way:

> When collections—or ensembles—of these models are used to simulate future climate, it sometimes happens that they all (or nearly all) agree regarding some interesting predictive hypothesis. For instance, two dozen state-of-the-art climate models might agree that, under a particular greenhouse gas emission scenario, the earth's average surface temperature in the 2090s would be more than 2°C warmer than it was in the 1890s. Such agreed-on or robust findings are sometimes highlighted in articles and reports on climate change, but what exactly is their significance? For instance, are they likely to be true?

Parker adds in a footnote:

> I take agreement among modeling results to be synonymous with robustness, as is common in the climate-modeling literature. For Pirtle et al. (2010), by

[1] Parker (2020) provides a framework for addressing this issue using a different concept: adequacy for purpose.

[2] E.g., Parker (2011), Lloyd (2009, 2015), Shupbach (2018), Wimsatt (2012), Orzack and Sober (1993), Weisberg (2006, 2012), Woodward (2006), Muldoon (2007), and Pirtle et al. (2010).

https://doi.org/10.1017/psa.2023.62 Published online by Cambridge University Press

> contrast, robustness seems to involve agreement plus some sort of independence among models that warrants increased confidence in the agreed-on result. Reasons for preferring one definition/characterization of robustness to the other will not be pursued here; either way, similar conclusions about the significance of agreement among predictions from today's climate models can be reached.

Parker asks two questions: (1) Under what conditions would special epistemic significance accrue to predictions if multiple models (of any type of scientific model, not just climate models) agree upon them? (2) Do those conditions hold for current climate models?

Because agreement about predictive hypotheses may stem from systemic error, Parker surveys potential epistemic conditions that could combine with agreement to underpin increased confidence. She considers three potential lines and evaluates whether they hold for the case of ensemble climate predictions. The first is that robust predictions from multimodel ensembles are likely to be true, either due to the way in which multimodel ensembles are constructed or due to ensemble performance. Second are three approaches to providing an argument from robustness to significantly increased confidence: a Bayesian perspective, one based on Condorcet's Jury Theorem, and a sampling-based perspective. The third is a view proposed by Staley (2004) that, setting aside whether robustness can increase the strength of evidence for a hypothesis, suggests that robust test results can increase the security of evidence claims.

> Parker (2011, 597) argues that "scientists are not in a position to argue that those conditions hold in the context of present-day climate modeling," and thus "when today's climate models are in agreement that an interesting hypothesis about future climate is true, it cannot be inferred—via the arguments considered here anyway—that the hypothesis is likely to be true or that confidence in the hypothesis should be significantly increased or that a claim to have evidence for the hypothesis is now more secure." Parker's succinct reaction: "This is disappointing."[3]

While the conditions are not met for climate models, according to Parker, the analysis does reveal key goals for the construction and evaluation of ensembles of models in a manner that would enable robustness to be epistemically significant. Parker (2011, 598) puts it this way:

> One goal, for instance, is the identification of a collection or space of models that can be expected to include at least one model that is adequate for indicating the truth/falsity of the hypothesis of interest; sampling from this collection

[3] Parker (2011, 598) remarks "Of course, it does not follow that climate policy decisions should be put on hold. Expectations of a warmer world are well founded; the challenge is rather to make sensible decisions despite remaining uncertainties about the details of future climate change." In Parker (2020), she discusses the notion of adequacy for purpose that provides an avenue for evaluating the capability of models to guide decisions.

https://doi.org/10.1017/psa.2023.62 Published online by Cambridge University Press

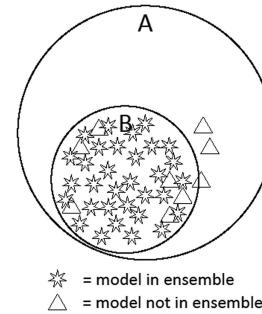

Figure 1. The sets of possible (A) and plausible (B) models. This ensemble effectively samples B.

(in order to construct the ensemble) should then be exhaustive, if possible, or else aimed to produce maximally different results.

So, the collection of models should be numerous and diverse enough that they effectively span the full range of possible models, and the ensemble should be populated by a thorough sampling of these models. In that case, at least one of the models in the ensemble should be adequate for indicating whether the hypothesis in question is true or false since we have covered all the possible bases.

A visual framework helps conceptualize Parker's and Lloyd's insights. Let us call the set of all possible climate models for a specific purpose of our choice **A**.[4] We are generally not interested in the set of all possible models, but rather the set of all plausible models, **B**, which is a proper subset of **A** (Figure 1).[5] Within **A** are numerous models developed by climate scientists, which are denoted schematically by the stars and triangles. The complementary space within **A** represents possible models that have not yet been developed.

What does it mean for a model to be plausible? I have in mind the minimal requirements that the model is consistent with both current scientific knowledge and past observations. For example, the models should not violate the Navier–Stokes equations nor posit that the mean temperature of the Earth last year was 100°C. The models within **B**, of course, could turn out to be empirically adequate (for purpose) or not with respect to future predictions.

Parker worries whether climate modelers have sufficiently sampled **B** through an ensemble, **E**, when all or most members of **E** agree on an interesting prediction, **H**. On her account, confidence in **H** requires that **E** include a thorough sample of **B**. We can consider an ensemble of models, **E**, and ask whether the models in **E** effectively sample **B**, which presumably would loosely mean that every sector of **B** is represented by at least one model in **E**. If that is the case (Figure 1), then we would be confident

[4] We may worry that the set of possible models is infinite, either because it's unbounded or because the values for parameters and variables can be continuous with an infinite number of variations between each pair of models. This issue does not matter for our concerns. If the set of models is infinite, in either sense, that strengthens Parker's arguments regarding the failure of ensembles to effectively sample the space of models. Whether it is infinite is irrelevant for Lloyd's model robustness because the space of relevant models spans a defined model type and no sampling is employed.

[5] I introduce this distinction, not Parker.

https://doi.org/10.1017/psa.2023.62 Published online by Cambridge University Press

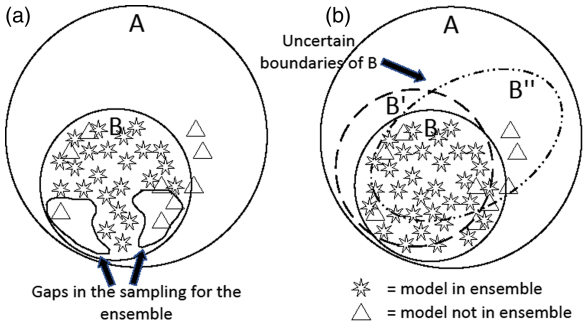

Figure 2. An ensemble failing to effectively sample the set of plausible models (a), and potential failure of effective sampling due to uncertainty about the scope of the set of plausible models (b).

that at least one of the models in the ensemble effectively represents the truth because the ensemble covers the full range of plausible models.

Parker argues that we have reasons to believe that actual climate model ensembles do not meet these conditions. Because ensembles in climate science tend to be "ensembles of opportunity," the models in **E** are likely tightly bunched together (Figure 2a) and thus do not span much of the space of **B** (Tebaldi and Knutti 2007; Sanderson and Knutti 2012).[6] Furthermore, because there is no way to define the boundaries nor extent of **B** (Figure 2b), there is no way to know whether **E** in fact spans **B** (Dethier 2022).

Parker is interested in the epistemic status of the *predictions*. While Parker provides guidance on the evaluation of models elsewhere (Parker 2020), that is not Parker's goal in discussing robustness. The predictions are hypotheses, while the models are (treated as) theories.

4. Lloyd's model robustness

Lloyd, however, provides a strategy for evaluating the epistemic status of *models*, hence the nomenclature of model robustness. On this approach, a collection of features provides a degree of confirmation for a model type, where the models have in common a type of structure, sharing general characteristics, in which "certain parameters are left unspecified" (van Fraassen 1980, 44). Lloyd (2015, 58) states:

> I propose a distinct type of robustness, which I suggest can support a confirmatory role in scientific reasoning, contrary to the usual philosophical claims. In model robustness, repeated production of the empirically successful model prediction or retrodiction against a background of independently supported and varying model constructions, within a group of models containing a shared causal factor, may suggest how confident we can be in the causal factor and predictions/retrodictions, especially once supported by a variety of evidence framework. I present climate models of greenhouse gas global warming of the 20th Century as an example, and emphasize climate scientists' discussions of robust models and causal aspects.

[6] This concern can be seen as a variant of Mill's classic competing hypothesis objection.

https://doi.org/10.1017/psa.2023.62 Published online by Cambridge University Press

A model type is a span of models that share a common "causal core"—causal processes and explanations of interest. Instantiations of models within the type will specify the parameters and details beyond the causal core.

The model type is confirmed by direct and indirect empirical support for the assumptions and core structure of the model as well as the accuracy of common predictions/retrodictions that can be tested empirically. Evidence that has been collected supporting the fundamental physics underlying the causal core of present-day greenhouse gas (GHG) models helps endorse the theory represented by the model type, as does the model type getting empirically accessible predictions/retrodictions correct. Lloyd is a proponent of complex empiricism, which is a variant of pragmatism. One therefore should not read "confirmation" here in the strong, formal sense in which it was used by advocates of hypothetico-deductivism. Lloyd means something more like the model type has been useful and empirically adequate and scientists have accumulated reasons for thinking that it is a productive research program.

Model robustness is primarily an evaluation of the model type;[7] if the model type is robust, then the common predictions of the future are more likely to be true in virtue of the confidence that has accrued to the model type that generated them, not merely because the models agree upon them. Lloyd (2015, 59) explains: "'model robustness' involves all this direct and indirect empirical support for the *explanatory causal core* of the model type, and by means of the causal core, the model prediction/retrodiction is also supported." Whereas Parker asks whether model agreement increases confidence in a prediction, Lloyd asks whether scientists should continue to investigate the climate through a research program using the model type. Lloyd (2015) argues that GHG global warming models constitute a robust model type.

Lloyd is aware that model robustness is unlike many other approaches to robustness (2015, 59): "Note that this is very different from other philosophical meanings of 'robustness,' which are usually solely defined in terms of the convergent *predictions* of a group of possibly (or even preferably) unrelated models" (emphasis in original).

We can see that Lloyd's account works in a different manner from that of Parker. Lloyd constructs a space of models, which we can denote as **C**, which spans those models which have the structure consistent with the model type (Figure 3a). The relationship between **C** and **B** is likely heterochronic—it depends on the maturity of the modeling efforts. Initially, multiple model types may be plausible, so **C** will be a proper subset of **B**. As the model type gains confirmation, the scientific community may come to think that the model type is the only currently plausible theory for the domain and purpose, in which case B will no longer extend beyond **C**. Lloyd suggests that to pursue the research program scientists engage in counterfactual probing of the assumptions of the model type, for example, by intentionally varying the plausible values of a parameter one model at a time. Lloyd argues that process credibility (Bukovsky et al. 2017; Knutti 2018; Lloyd, Bukovsky, and Mearns 2021), in addition to testing of agreement and empirical adequacy, should play a crucial role in this exploration. Some of the models within the model type may thus be ruled out as implausible. As the research program proceeds, **B** may in this way become a subset of **C** (Figure 3b).

[7] "Model robustness involves both the causal core of the individual models and the convergence of the model outcomes" (Lloyd 2015, 64), but when indicated by these factors, it provides a positive epistemic evaluation of the model type.

https://doi.org/10.1017/psa.2023.62 Published online by Cambridge University Press

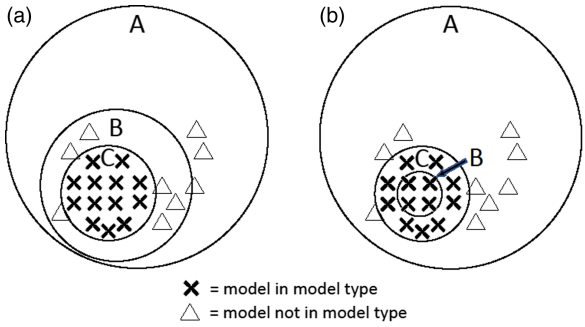

Figure 3. Models in a model type C (a) and progress through a research program based on a model type C leading to the plausible models B as a subset of C (b).

Robustness and model robustness clearly aim to address different questions. Parker is asking whether predictive hypotheses are more likely to be true because multiple models agree upon them. Lloyd is asking whether agreement on predictions, along with independent empirical support for the assumptions and causal structure of the model type, provide confirmation for the theory represented by the model type. Nevertheless, it may seem surprising at first that they arrive at different answers because one might think that if contemporary climate models represent a theory that has accumulated confirmatory value (as Lloyd suggests), they will agree on predictions that are likely to be accurate, while if we do not have confidence in the predictive hypotheses (as Parker suggests), then there must be something wrong with the models. But Parker and Lloyd are each providing valuable, and complementary, insights.

5. Clarifying types of robustness

Let us consider a basic intuition pump. Suppose we have a collection of models for predicting the outcome of an election. Each predicts that candidate A will defeat candidate B. Should we have more confidence that candidate A will win because of the agreement? Can we make use of a model or group of the models in future elections?

If all the models are essentially the same but for an adjustment in a parameter value, then the agreement should do little for our confidence in the prediction. Imagine each model in the collection relies upon an economic cause for voter decisions. The only variation is a slight adjustment in the weight of GDP in the last quarter leading up to the election. We would likely think that we have consulted the same model repeatedly, so agreement is expected and is thus epistemically unremarkable.

Significant diversity between the models would be more epistemically impactful. Suppose that one relies on economic factors, another on the personal popularity of the candidates, another on the popularity of policy proposals they have made, another on cultural factors, and so forth. If all these models predict victory for candidate A, then we would likely gain confidence that candidate A will emerge victorious. Whatever reason(s) are in fact at play, they all point in the same direction.

Concerning the usefulness of these models for future elections, which we might call model projectability, quite distinct considerations are operative. The diversity

https://doi.org/10.1017/psa.2023.62 Published online by Cambridge University Press

that was helpful for confidence in the prediction can now be counterproductive. To have confidence in applying a model or collection of models to future elections, we should desire confidence in the causal story or stories that they are telling. Are economic factors influential in determining elections? Are policy prescriptions? We want to include only models whose causal narratives have some degree of confirmation. In fact, we would ideally want an accounting of their relative contributions and weight them accordingly. To the extent that any independent variables are uninformative, they should be excluded, and we would not want to include mutually exclusive "causal cores," even if they agree on the predicted outcome. That is to say that we want not just the right prediction, but we also want it to be for the right reason(s).[8]

A research program would search for independent warrant for the assumptions and causal structure of the models and collect empirical evidence about the accuracy of the predictions of the models. Models that lack independent warrant and/or prove to be inaccurate would be weeded out. In this way, the collection of models may hopefully coalesce into a model type that gets ever more accurate. Because our intuition pump addresses politics, we may not end up with a single model type and being able to achieve a high level of predictive accuracy. In the case of a research program in the natural sciences, hopes would be higher of achieving these features, though given the complexity present in many domains of modeling, different model types may be needed to provide adequacy for different features.

Hopefully, it is apparent how Parker and Lloyd are providing complementary insights. Parker is making the point that agreement by multiple models on a prediction should not in and of itself provide us with confidence that the prediction is likely to be true unless, for example, the collection of models is large enough and diverse enough. Three economic models with slightly different weights for final quarter GDP predicting that candidate A will win the election should not fill us with confidence that candidate A will indeed be victorious. Likewise, an ensemble of opportunity of climate models that share significant code, and that were developed by overlapping groups of scientists all agreeing that the earth's average surface temperature in the 2090s would be more than 2°C warmer than it was in the 1890s, should not fill us with confidence that the value will indeed be more than 2°C warmer.

Lloyd is making the point that scientific research programs aim to develop projectable models—model types in which we have confidence about its causal story. Such a model type has accumulated confirmation using an independent empirical warrant for its assumptions and causal core and its predictions have tended to be fairly accurate. A program of counterfactual probing of details should be employed to progressively increase the accuracy of the model type. Lloyd argues that GHG global warming models qualify as such a robust model type. Because this model type is robust, if the models within it agree that the earth's average surface temperature in the 2090s would be more than 2°C warmer than it was in the 1890s, we should have increased confidence that the value will indeed be more than 2°C warmer.

Model robustness is more germane to the trustworthiness of contemporary climate modeling than Parker's account of robustness. Both provide valuable insights, but Parker's point is abstract—scientists and policy makers should not have

[8] This is the motivation for process credibility in the case of climate models.

https://doi.org/10.1017/psa.2023.62 Published online by Cambridge University Press

confidence in the projections of climate models based on the reason that multiple models agree on the prediction—whereas Lloyd's point is practical—scientists and policy makers should have some confidence that the GHG global warming model type has accumulated confirmation. In line with her complex empiricism, Lloyd is not suggesting that a specific model within the type necessarily represents a completely accurate description of the real world but is arguing for optimism about scientists and policymakers consulting the projections of this model type in considering, for example, mitigation and adaptation strategies.

6. Conclusion

Climate modelers often use agreement among multiple models as a source of confidence in the accuracy of model projections. Philosophers of climate science have therefore prioritized analyzing this "robustness" as an epistemic strategy. The most influential accounts of robustness with regards to climate modeling are those due to Parker (2011) and Lloyd (2009, 2015). I have argued that these accounts address different questions. Both provide valuable insights about models and their predictions. I have argued that it is model robustness that is more relevant for determining the productivity and trustworthiness of climate modeling. A diverse collection of models with inconsistent causal narratives may increase confidence in a specific predictive hypothesis but cannot serve as the foundation of a productive research program. A robust model type with a consistent (and confirmed) causal narrative, is what is needed for a productive and trustworthy research program.

Acknowledgments. I would like to thank Lisa Lloyd, Wendy Parker, Ryan O'Loughlin, Zach Pirtle, and Steve Elliott for their valuable feedback. This paper is partially based on research that was funded by the National Science Foundation Award Number 1754740: A Methodological Study of Big Data and Atmospheric Science.

References

Bukovsky, M. S., R. R. McCrary, A. Seth, and L. O. Mearns. 2017. "A Mechanistically Credible, Poleward Shift in Warm-Season Precipitation Projected for the US Southern Great Plains?" *Journal of Climate* 30 (20):8275–98.

Dethier, C. 2022. "When Is an Ensemble Like a Sample? 'Model-Based' Inferences in Climate Modeling." *Synthese* 200 (1):52.

Knutti, R. 2018. "Climate Model Confirmation: From Philosophy to Predicting Climate in the Real World." *Climate Modelling: Philosophical and Conceptual Issues*, edited by Elisabeth A. Lloyd and Eric Winsberg, 325–59. Cham: Palgrave Macmillan.

Levins, R. 1966. "The Strategy of Model Building in Population Biology." *American Scientist* 54 (4):421–31.

Lloyd, E. A. 2009. "I—Elisabeth A. Lloyd: Varieties of support and confirmation of climate models." *Aristotelian Society Supplementary Volume* 83 (1):213–32.

Lloyd, E. A. 2015. "Model Robustness as a Confirmatory Virtue: The Case of Climate Science." *Studies in History and Philosophy of Science Part A* 49:58–68.

Lloyd, E. A., M. Bukovsky, and L. O. Mearns. 2021. "An Analysis of the Disagreement about Added Value by Regional Climate Models." *Synthese* 198:11645–72.

Muldoon, R. 2007. "Robust simulations." *Philosophy of Science* 74 (5):873–83.

Orzack, S. H., and E. Sober. 1993. "A Critical Assessment of Levins's the Strategy of Model Building in Population Biology (1966)." *The Quarterly Review of Biology* 68 (4):533–46.

Parker, W. S. 2011. "When Climate Models Agree: The Significance of Robust Model Predictions." *Philosophy of Science* 78 (4):579–600.

Parker, W. S. 2020. "Model Evaluation: An Adequacy-for-Purpose View." *Philosophy of Science* 87 (3):457–77.

https://doi.org/10.1017/psa.2023.62 Published online by Cambridge University Press

Pirtle, Z., R. Meyer, and A. Hamilton. 2010. "What Does It Mean When Climate Models Agree? A Case for Assessing Independence among General Circulation Models." *Environmental Science & Policy* 13 (5):351–61.
Sanderson, B. M., and R. Knutti. 2012. "On the Interpretation of Constrained Climate Model Ensembles." *Geophysical Research Letters* 39 (16):L16708.
Schupbach, J. N. 2018. "Robustness Analysis as Explanatory Reasoning." *The British Journal for the Philosophy of Science* 69 (1):275–300.
Staley, K. W. 2004. "Robust Evidence and Secure Evidence Claims." *Philosophy of Science* 71 (4):467–88.
Tebaldi, C., and R. Knutti. 2007. "The Use of the Multi-model Ensemble in Probabilistic Climate Projections." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365(1857): 2053–75.
Van Fraassen, B. C. 1980. *The Scientific Image*. Oxford: Oxford University Press.
Weisberg, M. 2006. "Robustness Analysis." *Philosophy of Science* 73 (5):730–42.
Weisberg, M. 2012. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.
Wimsatt, W. C. 2012. "Robustness, Reliability, and Overdetermination (1981)." *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science*, 61–87.
Winsberg, E. 2018. *Philosophy and Climate Science*. Cambridge: Cambridge University Press.
Woodward, J. 2006. "Some Varieties of Robustness." *Journal of Economic Methodology* 13 (2):219–40.

Cite this article: Gluck, Stuart. 2023. "Robustness of Climate Models." *Philosophy of Science* 90 (5): 1407–1416. https://doi.org/10.1017/psa.2023.62