When Climate Models Agree: The Significance of Robust Model Predictions*

Wendy S. Parker[†‡]

This article identifies conditions under which robust predictive modeling results have special epistemic significance—related to truth, confidence, and security—and considers whether those conditions hold in the context of present-day climate modeling. The findings are disappointing. When today's climate models agree that an interesting hypothesis about future climate change is true, it cannot be inferred—via the arguments considered here anyway—that the hypothesis is likely to be true or that scientists' confidence in the hypothesis should be significantly increased or that a claim to have evidence for the hypothesis is now more secure.

1. Introduction. There is now a broad scientific consensus—underwritten by a substantial and growing body of evidence—that the earth's climate warmed significantly over the last century, that increased atmospheric concentrations of greenhouse gases due to human activities are a major cause of this warming, and that the earth's climate will be still warmer by the end of the twenty-first century (Solomon et al. 2007). Less clear are the quantitative details, especially regarding future climate change. How much will the earth's average surface temperature increase by the end of the twenty-first century if greenhouse gas concentrations continue rising as they have in recent decades? Under that scenario, will the central

*Received September 2009; revised December 2010.

†To contact the author, please write to: Department of Philosophy, Ohio University, Athens, OH 45701; e-mail: parkerw@ohio.edu.

‡Sincere thanks to Dan Steel, Reto Knutti, Kent Staley, Phil Ehrlich, Lenny Smith, Joel Katzav, Charlotte Werndl, and two anonymous referees for helpful suggestions and criticisms. Thanks also to those who provided feedback when earlier versions of this article were presented at Purdue University, University of Colorado at Boulder, University of Toronto, and University of Waterloo.

Philosophy of Science, 78 (October 2011) pp. 579–600. 0031-8248/2011/7804-0003$10.00

United States experience much drier summers as the century unfolds? What will climatic conditions in various locales be like late in the twenty-first century if instead greenhouse gas concentrations are stabilized at 450 parts per million by 2025?

Current scientific understanding suggests that answers to questions like these, about long-term changes in global and regional climate, may depend on the details of complex interactions among many climate system processes—details that cannot be tracked without the help of computer simulation models. Numerous simulation models have been developed, differing in their spatiotemporal resolution, the range of climate system processes that they take into account, and the ways in which they represent those processes. 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.[1] 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.[2] Such agreed-on or *robust* findings are sometimes highlighted in articles and reports on climate change, but what exactly is their significance?[3] For instance, are they likely to be true?

The discussion that follows has two main goals. First, it aims to identify conditions under which robust predictive modeling results—not just from

1. By an interesting predictive hypothesis, I mean a hypothesis about the future that scientists (i) do not already consider very likely to be true or very likely to be false and (ii) consider a priority for further investigation. In climate science today, these are typically, but not always, quantitative hypotheses about changes in global or regional climate on the timescale of several decades to centuries.

2. What does it mean for an ensemble to agree that a hypothesis is true? Assume that modelers have decided on rules for translating statements about (some) model variables into statements about (some) target system properties, so that the values of those variables during a simulation can be understood as indications regarding target system properties. Then a simulation indicates the truth (falsity) of some hypothesis H about a target system if a statement of what the simulation indicates about one or more properties of the target system entails that H is true (false). For example, if H says that temperature will increase by between 1°C and 1.5°C, and each of the simulations in an ensemble indicates an increase between 1.2°C and 1.4°C, then each of those simulations indicates the truth of H, and the ensemble is in agreement that H is true.

3. 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 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.

climate models but from scientific models in general—have special epistemic significance. Classic discussions of robustness include Levins (1966) and Wimsatt (1981/2007), and connections between robustness and prediction have been touched on recently by some authors (e.g., Weisberg 2006; Woodward 2006; Muldoon 2007; Pirtle et al. 2010), but there has been little detailed analysis of the conditions under which robust predictive modeling results have special epistemic significance. Having identified some of these conditions, a second goal is to investigate whether they currently hold in the context of ensemble climate prediction, as a first step toward evaluating the significance of robust predictions from today's climate models.

Section 2 gives a brief introduction to ensemble climate prediction, explaining how and why multiple models are used to investigate future climate change. The next three sections investigate the prospects for inferring from robust modeling results, and from robust climate-modeling results in particular, that

i) an agreed-on predictive hypothesis H is likely to be true (sec. 3),
ii) significantly increased confidence in H is warranted (sec. 4),
iii) the security of a claim to have evidence for H is enhanced (sec. 5).

The findings are disappointing. When today's climate models agree that an interesting hypothesis about long-term climate change is true, it cannot be inferred—via the arguments considered here anyway—that the hypothesis is likely to be true or that scientists' confidence in the hypothesis should be significantly increased or that a claim to have evidence for the hypothesis is now more secure. In closing, section 6 reflects on these findings.

2. Ensemble Climate Prediction. A computer simulation model is a computer-implemented set of instructions for repeatedly solving a set of equations in order to produce a representation of the temporal evolution of selected properties of a target system. In the case of global climate modeling, the target system is the earth's climate system—encompassing the atmosphere, oceans, sea ice, and land surface—and the equations are ones that describe in an approximate way the local rate of change of temperature, wind speed, humidity, and other quantities of interest in response to myriad processes at work in the system. When it comes to formulating such equations, considerable uncertainty remains for several reasons. Although a theory of large-scale atmospheric dynamics (grounded in fluid dynamics) has long been in place and provides the foundation for some parts of today's climate models, some other important climate system processes are less well understood. In addition, for processes that are believed to influence climate in important ways but that occur on scales

finer than those resolved in today's models (e.g., on spatial scales smaller than ~100 km in the horizontal dimension or on time scales shorter than ~1/2 hour), rough representations in terms of larger-scale variables must be developed, and it is rarely obvious how this can best be done. The upshot is that multiple climate models, which differ in various ways in their equations and in the methods they use to estimate solutions, are nevertheless judged to have approximately equal prima facie plausibility as tools for predicting future climate (Parker 2006). Indeed, even after examining how well these different models simulate past and present climate, it is often unclear which would be best for a given predictive task.[4]

Given this uncertainty, how should climate scientists proceed? If it is unclear which of several models will turn out to give the best prediction in a particular case, then it would be unwise to select just one of the models and rely on its prediction, unless all of the models are expected to be so accurate that any would be good enough. Since the latter cannot be expected of today's climate models, ensemble studies present a better option. These studies involve running each of several climate models (or model versions) with the same (or similar) initial conditions and under the same (or similar) emission scenarios (see, e.g., Stainforth et al. 2005; Tebaldi et al. 2005; Murphy et al. 2007). Ensemble studies acknowledge that there is uncertainty about how to represent the climate system and explore how much this uncertainty matters when it comes to predictions of interest (Parker 2006).

There are two main types of ensemble climate prediction studies today. *Multimodel ensemble studies* produce simulations of future climate using models that differ in a number of ways—in the form of some of their equations, in some of their parameter values, and often in their spatiotemporal resolution, their solution algorithms, and their computing platforms as well. A typical multimodel study requires the participation of research groups at various modeling centers around the world, each running its "in-house" models on local supercomputers, and delivers a total of a few dozen simulations of future climate under a given emission scenario (see, e.g., Meehl et al. 2007). *Perturbed-physics ensemble studies* employ multiple versions of a single climate model whose best parameter values remain uncertain. The model is run repeatedly, leaving the structure of its equations unchanged but allowing its uncertain parameters to take different values on each run. The selection of these parameter values can be made using formal sampling methods or in more informal ways; usually

4. In part, this is because it is difficult to determine what a model's performance in simulating past and present climate indicates about its accuracy in predicting various quantities of interest (see Randall et al. 2007; Gleckler, Taylor, and Doutriaux 2008; Parker 2009).

values are chosen from a range identified by expert judgment. A single perturbed physics study may produce a large number of simulations of future climate, depending on how computationally intensive it is to run a single simulation. Studies carried out by the climateprediction.net project, for example, rely on donated idle processing time on ordinary home computers to produce thousands of simulations using different versions of a (relatively) complex climate model (Stainforth et al. 2005; BBC 2010).

The discussion that follows will focus on results from multimodel ensemble studies. This is because perturbed-physics studies explore such a broad range of parameter values that they deliver a very wide range of results—so wide that the results are not in unanimous (or even near unanimous) agreement regarding interesting predictive hypotheses. It tends to be multimodel ensemble studies, rather, in which such agreement occurs. For instance, in a recent multimodel study that investigated a "high" emission scenario using 17 state-of-the-art climate models, each of the models indicated that, by 2050, global mean surface temperature would be between 1°C and 2°C warmer than during 1980–99 (see Meehl et al. 2007, 763).[5] Likewise, virtually all of the models agreed that, under a "medium" emission scenario, summer rainfall in east Africa would be greater in the late twenty-first century than it was in the late twentieth century (Christensen et al. 2007, 869). The question is whether agreed-on multimodel results like these have special epistemic significance and, if so, what that significance is.[6]

3. Robustness and Truth. Can it be argued that robust predictions from today's multimodel ensembles are likely to be true? More generally, under what conditions can an inference from robustness to likely truth be justified? Consider the following argument, inspired by more general discussions of robustness given by Orzack and Sober (1993) and Woodward (2006):

1. It is likely that one of the models in this collection is true.
2. Each of the models in this collection logically entails hypothesis H.

$\therefore$ It is likely that H.

5. More precisely, average results for individual models were in agreement regarding the hypothesis; some models were run more than once with different initial conditions, and only average results for each model were shown in the main body of the report.

6. Whether the average of results produced by a given multimodel ensemble should be considered current science's "best guess" is a separate matter that will not be addressed here, but see Knutti et al. (2010) for cautionary analysis. Note, however, that an ensemble average can indicate a hypothesis to be true, even when there is substantial discord among individual modeling results, whereas in the case of robust findings, the evidence under consideration is especially concordant regarding the hypothesis.

While its logic is unobjectionable, this argument seems largely inapplicable in science; insofar as a scientific model can be identified with a complex hypothesis about the workings of a target system, there is usually good reason to believe that such a hypothesis is (strictly) false since most scientific models are known from the outset to involve idealizations, simplifications, or outright fictions. So 1 will rarely hold.[7]

Nevertheless, a similar argument with greater potential for applicability might be constructed as follows:

1′. It is likely that at least one simulation in this collection is indicating correctly regarding hypothesis H.
2′. Each of the simulations in this collection indicates the truth of H.
∴ It is likely that H.

Here, reference to the truth of models has been replaced by reference to simulations' indicating correctly regarding a hypothesis. A simulation indicates correctly regarding a hypothesis H if it indicates the correct truth value for H. A model producing such a simulation, while it may rest on various simplifications and idealizations, is nevertheless adequate for the purpose of interest—namely, for indicating whether H is true.[8] Call 1′ the *likely adequacy* condition.

Is there good evidence that the likely adequacy condition is met in today's multimodel climate prediction studies? The answer might be yes in some cases and no in others, depending on the ensembles and the hypotheses. How could climate scientists argue that the condition is met in a particular case? At least two approaches are possible: one that focuses on ensemble construction and one that focuses on ensemble performance.

Taking the former approach, one would argue that an ensemble of models samples so much of current scientific uncertainty about how to represent the climate system (for purposes of the predictive task at hand) that it is likely that at least one simulation produced in the study is indicating correctly regarding H.[9] Can this argument be made for today's multimodel ensembles? It cannot. For these ensembles are *ensembles of*

7. Woodward (2006) notes the limited applicability of a related analysis.

8. An adequate model is one that is sufficient for the purposes of interest not just as a matter of accident (e.g., a one-off accurate prediction) but because the model has properties that make it suitable for those purposes. One way to cash out this suitability is in terms of relevant similarities between the model and the target system, but other approaches are possible too. In the case of climate models, scientists might describe an adequate model as one that "captures enough of the relevant physics" for the purposes at hand.

9. This is reminiscent of Michael Weisberg's claim: "The key comes in ensuring that a sufficiently heterogeneous set of situations is covered in the set of models subjected to robustness analysis" (2006, 739).

opportunity, assembled from existing climate models and only insofar as research groups are willing to participate (Meehl et al. 2007; Tebaldi and Knutti 2007); they are "not designed to span an uncertainty range" (Knutti et al. 2008, 2653). For instance, while each state-of-the-art model in an ensemble includes some representation of clouds, no attempt is made to ensure that the ensemble as a whole does a good job of sampling (or spanning) current scientific uncertainty about how to adequately represent clouds, and likewise for various other subgrid processes and phenomena. Indeed, when it comes to discerning the truth/falsity of quantitative hypotheses about long-term climate change, climate scientists today are not in a position to specify a small set of models that can be expected to include at least one adequate model. In part, this is because it remains unclear whether processes and feedbacks that will significantly shape long-term climate change have been overlooked (so-called unknown unknowns). But it also reflects the challenge of anticipating how recognized simplifications, approximations, and omissions will affect the accuracy of predictions produced by complex, nonlinear models (see also Parker 2009).

On a performance approach to justifying the likely adequacy condition, an ensemble is viewed as a tool for indicating the truth/falsity of hypotheses of a particular sort, of which the predictive hypothesis H is an instance; the ensemble's past reliability with respect to H-type hypotheses is cited as evidence that it is likely that at least one of its simulations is indicating correctly regarding this particular H.[10] Assuming that H concerns the value of a given variable, this is tantamount to arguing that it is likely that the range of values spanned by the ensemble's predictions will either include the true value of that variable or else come within some specified distance of that value. For instance, consider H: under this emission scenario, global mean surface temperature (GMST) for 2080–89 would be between 1.5°C and 2.0°C warmer than GMST for 1980–89. Suppose that all of the climate models in an ensemble indicate the truth of this hypothesis and, specifically, that their predicted changes all fall between 1.6°C and 1.9°C. Then the likely adequacy condition will be met only if it is likely that the range of predictions delivered by the ensemble will either include the true temperature change or else come within 0.1°C of doing so (extending just to the edges of the hypothesized range).

Does the performance of today's multimodel ensembles up to now provide good evidence that, for a given climate variable of interest, it is likely that the range of values predicted by those ensembles will either include the true value of the variable or else come within some specifiable,

10. So while we may not know which member(s) of the ensemble will indicate correctly regarding a given H-type hypothesis, we have evidence that there is usually at least one such member.

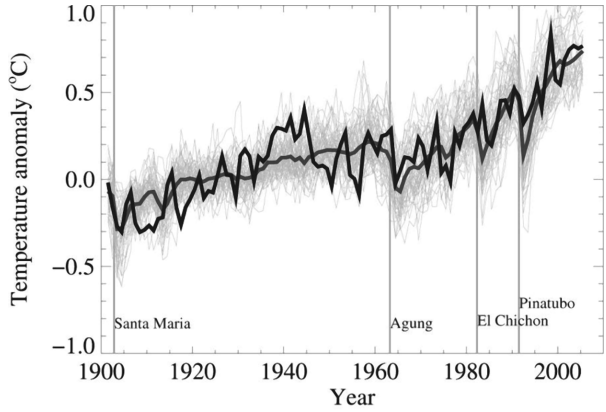

Figure 1. Anomalies in global mean surface temperature (GMST) over the twentieth century as obtained from observations (*black*) and from 58 simulations produced by 14 different climate models (*thin light gray*). Anomalies for a given simulation are relative to that simulation's average GMST for 1901–50. Observed anomalies are relative to observed average GMST for 1901–50. Average values for the simulated anomalies are also shown (*dark gray*). Adapted from *Climate Change 2007: The Physical Science Basis*. Working Group I Contribution to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, figure 9.5 (a). Cambridge University Press. See Solomon et al. (2007).

small distance of it? That is, is there good evidence that today's ensembles reliably "capture truth"—or come close enough to capturing it—when it comes to the predictive variables that interest scientists and decision makers?[11]

In practice, careful investigation of the truth-capturing performance of today's ensembles has been carried out for rather few variables thus far. A striking example, however, is shown in figure 1. Glancing at the figure, it appears that, for almost every year in the twentieth century, the observed global temperature anomaly for the year is within the range of values spanned by the ensemble.[12] Nevertheless, it can be difficult to determine what findings like those depicted in figure 1 indicate about the future truth-capturing abilities of today's ensembles, in part because of the complicated model-data relationships that often obtain in this context; as

11. The "capturing truth" terminology is taken from Judd, Smith, and Weisheimer (2007), who include a related technical definition of the "bounding box" of an ensemble.

12. Temperature anomalies are differences from some reference temperature. In the case of fig. 1, the reference temperature is different for each time series (see caption).

discussed below, climate data sets are often model filtered, and climate models are often data laden (Edwards 1999, 2010).

It is widely recognized that scientific analysis often appeals not to raw observational data but to cleaned-up depictions of those data, known as *data models* (Suppes 1962; Harris 2003). This is certainly true in climate science. However, the production of data models in climate science often goes beyond the sort of correcting for instrumental error and noise that is typical in other sciences. In particular, cleansed observational data may be synthesized with output from weather-forecasting models. This is done in part to fill in gaps—to provide values for locations in the atmosphere (or even entire fields/variables) for which few if any raw observations are available—and delivers data sets that include values for chosen variables on a regular spatial grid and at regular time intervals. Known as *reanalysis data sets*, they often are used to evaluate climate model performance.[13] But interpreting the results of such model-data comparisons is complicated since weather-forecasting models include a number of assumptions about the physics of the atmosphere that are similar, if not identical, to those included in state-of-the-art climate models—assumptions that to varying degrees involve idealization and simplification. This raises the worry that the fit between reanalysis data sets and simulations of past climate, and thus the frequency with which ensembles are found to capture truth, will be artificially inflated. So far, however, evaluations of climate models have not been accompanied by estimates of the extent to which this inflation may be occurring for different variables and time periods.[14]

A climate model can become data laden in several ways, most notably via *tuning*. Tuning a climate model involves making ad hoc changes to its parameter values or to the form of its equations in order to improve the fit between the model's output and observational/reanalysis data. Tuning of models occurs in many scientific fields and is not necessarily bad since after a model has been tuned to a data set it may perform better with respect to as-yet-unseen data as well. But given the ad hoc nature of the tuning process, and the fact that today's climate models are far from perfect in their representation of the climate system, it cannot be assumed that the performance of a tuned climate model with respect to as-yet-unseen data will be similar to its performance with respect to the

13. Gleckler et al. (2008) investigate the performance of climate models in simulating approximately two dozen fields (e.g., wind speeds at a particular atmospheric level, latent heat flux at the ocean surface), and data sets for about half of those fields are reanalysis data sets.

14. Detailed discussion of how data can be processed legitimately vs. illegitimately if they are to be used for model evaluation is beyond the scope of this article, but the topic is clearly worth pursuing.

data to which it is tuned.[15] Moreover, when today's climate models are tuned, it is often difficult to adequately test their out-of-sample performance, both because reliable observations of past climate are limited and because most observations that are available are for time periods in which greenhouse gas concentrations were significantly lower than they are expected to be in the future.

Do these complications arise in the case of figure 1 in particular? The values plotted as "observations" in figure 1 were calculated from a data set (Brohan et al. 2006) whose production did not involve synthesizing observational data with output from weather-forecasting models, so the concern about model-filtered data does not seem to be in play here.[16] However, because accounting for twenty-first-century changes in GMST has been a major focus of modeling efforts in recent decades, it seems very likely that for most of today's state-of-the-art climate models, including those in the figure 1 ensemble, at least some tuning has been done with these temperature changes in mind.[17] This makes it harder to discern what figure 1 says about the future truth-capturing ability of its ensemble, even with respect to future GMST anomalies, much less other predictive variables.

A closer look at the simulations from which figure 1 was produced complicates matters further. It turns out that the temperature anomalies plotted in figure 1 were derived from simulated temperature values with biases of several degrees Celsius in many regions (see Randall et al. 2007, supplementary material; Knutti et al. 2010). So while the models roughly track the way estimated GMST has changed over the last century, some of them show significant errors when it comes to the temperatures from which those changes are calculated. From the point of view of dynamical systems theory, this means that the trajectories of those simulations through a high-dimensional state space (defined by the models' variables) differ substantially from the trajectory of the real climate system as estimated from observations. Given nonlinear feedback in the climate sys-

15. If a model is thought to provide a very accurate representation of (relevant) aspects of a target system, with parameters that have clear physical correlates, then scientists might justifiably expect that the model will perform best when its parameters are set to values very near their measured values, with the optimal values found via tuning. But this is not the situation in climate modeling today; because of significant errors elsewhere in the model, parameter values that give the best model performance might be noticeably different from measured values—if a clear physical interpretation of the parameter can be given at all.

16. Edwards (1999) notes that other types of models, such as models of instruments, are also (at least implicitly) relied on in producing climate data sets. The data set of fig. 1 is model filtered in this broader sense, but so is virtually every data set in science.

17. Knutti (2008) suggests something similar.

tem, this raises concern that model trajectories for the twenty-first-century climate (and beyond) might rapidly diverge from the observed trajectory for a given emission scenario. This is yet another reason why it is risky to assume that the frequency with which an ensemble captures truth in simulations of recent climate is representative of how frequently it will do so in the future.

In summary, whether ensemble construction or ensemble performance is considered, there is not yet good evidence that today's multimodel ensembles meet the likely adequacy condition for interesting hypotheses about future climate change. So it is not yet possible to make the argument (presented above) from robustness to likely truth.[18]

4. Robustness and Confidence. Even if an inference from robustness to the likely truth of an agreed-on predictive hypothesis cannot be justified in a particular case, it still might be argued that robustness warrants significantly increased confidence in the hypothesis. Indeed, a recent analysis by Pirtle et al. (2010) suggests that climate scientists often do assume that agreement warrants this. In what follows, three general approaches to providing an argument from robustness to significantly increased confidence are identified, but each runs into problems in the context of ensemble climate prediction.

4.1. A Bayesian Perspective. Within a standard Bayesian framework, one's confidence (or degree of belief) in a hypothesis H is the subjective probability that one assigns to H, and Bayes' Theorem provides a rule for updating that assignment in light of new evidence e. According to the rule, one's new probability assignment, $p(H|e)$, should be set as follows: $p(H|e) = p(H) \times p(e|H)/p(e)$, where $p(H)$ is one's probability assignment for H before obtaining e, $p(e|H)$ is the probability that one assigns to e under the assumption that H is true, and $p(e)$ is the probability that one assigned to e before actually encountering e. Given this updating rule, confidence in H should increase in light of evidence e if and only if (iff)

18. This is fully compatible with there being some hypotheses about future climate that scientists can, with justification, consider likely to be true. For example, the expectation that the global climate will continue to warm in the twenty-first century is grounded not just in agreement among predictions from complex climate models but also in basic understanding of physical processes, theoretical analysis, observational data, and results from simpler models. When it comes to discerning the truth/falsity of relatively precise quantitative hypotheses about long-term changes in global and regional climate, however, complex simulation models are the primary means of investigation since only they are intended to track in significant detail the complex interactions among climate system processes (see sec. 1).

$p(e|H) > p(e|\sim H)$.[19] That is, e will increase confidence in H iff the occurrence of e is more probable if H is true than if H is false. Similarly, e will significantly increase confidence in H iff the occurrence of e is substantially more probable if H is true than if H is false, that is, iff $p(e|H) \gg p(e|\sim H)$, where what counts as "significant" and "substantial" is context relative.

Thus, a Bayesian argument from robustness to significantly increased confidence might go as follows:

1. e warrants significantly increased confidence in predictive hypothesis H if $p(e|H) \gg p(e|\sim H)$.
2. e = all of the models in this ensemble indicate H to be true.
3. The observed agreement among models is substantially more probable if H is true than if H is false; that is, $p(e|H) \gg p(e|\sim H)$.

$\therefore$ e warrants significantly increased confidence in H.

The argument has a valid form. But are its premises true in the case of ensemble climate prediction?

Premise 1 is part and parcel of the Bayesian framework, as just discussed; 2 is simply a statement of robustness/agreement; 3 is where the real action of the argument will be in any particular case and also where the potential weakness of this Bayesian approach becomes clear. For 3 concerns the probability assignments made by a particular epistemic agent (individual or group), and if those assignments do not reflect substantial evidence, then the move from robustness to increased confidence in H could come very cheaply. If the argument above is to have much persuasive force, 3 should be given some substantive justification.

Once again, at least two justificatory approaches are possible, focusing on ensemble construction and ensemble performance, respectively. Taking the former approach, scientists might argue that, given the conditions under which the individual models in the ensemble can be expected to err—inferred from information about how the models are constructed, such as the sorts of idealizations that they include—the models are substantially more likely to agree that H is true when it is true than when it is false. A performance-based justification, by contrast, might demonstrate that, in a large set of trials up to now, the requisite agreement among ensemble members' indications regarding H-type hypotheses occurred

19. From the updating rule, we see that $p(H|e) > p(H)$ iff $p(e|H)/p(e) > 1$. But $p(e|H)/p(e) > 1$ iff $p(e) < p(e|H)$. When is $p(e) < p(e|H)$? By the law of total probability, $p(e) = p(e|H) \times p(H) + p(e|\sim H) \times p(\sim H)$. Since $p(H) + p(\sim H) = 1$, $p(e)$ is in effect a weighted average of $p(e|H)$ and $p(e|\sim H)$; it takes a value between $p(e|H)$ and $p(e|\sim H)$. So $p(e)$ will be smaller than $p(e|H)$ iff $p(e|H) > p(e|\sim H)$. So $p(H|e) > p(H)$ iff $p(e|H) > p(e|\sim H)$.

much more often when H-type hypotheses were true than when they were false.

Unfortunately, neither sort of justification is readily supplied in the case of ensemble climate prediction today, for reasons already discussed in section 3. Current understanding (of today's models and of the climate system itself) is not extensive enough to allow for a construction-based justification. Performance data are limited (because observational data are limited) and, in addition, are generally difficult to interpret due to tuning, model-filtered data, and so on. Moreover, there are reasons to worry that simulations from today's state-of-the-art climate models might not so infrequently agree that a predictive hypothesis of interest is true, even though it is false.[20] First, there are climate system features and processes—some recognized and perhaps some not—that are not represented in any of today's models but that may significantly shape the extent of future climate change on space and time scales of interest. In addition, when it comes to features and processes that are represented, different models sometimes make use of similar idealizations and simplifications. Finally, errors in simulations of past climate produced by today's models have already been found to display some significant correlation (see, e.g., Knutti et al. 2010; Pennell and Reichler 2011). Thus, in general, the possibility should be taken seriously that a given instance of robustness in ensemble climate prediction is, as Nancy Cartwright once put it, "an artifact of the kind of assumptions we are in the habit of employing" (1991, 154).[21] Perhaps with additional reflection and analysis, persuasive arguments for $p(e|H) \gg p(e|\sim H)$ can be developed in some cases, but at present such arguments are not readily available.

4.2. Condorcet's Jury Theorem. Another possible approach draws on Condorcet's Jury Theorem. According to the traditional version of this theorem, if each of $n > 1$ voters has the same probability $p > .5$ of voting correctly regarding which of two options is "better" (on some criterion) and if the votes are statistically independent, then the probability that at least a majority of voters will choose the "better" option exceeds p and, moreover, exceeds p to a greater extent with increasing n (see, e.g., Ladha 1995). Treating the indications of individual simulations regarding the

20. The reasons given here are discussed independently by Tebaldi and Knutti (2007).

21. Wimsatt (1981/2007) discusses a case in biology in which an apparently robust modeling result turned out to be grounded in erroneous assumptions shared by the models. See also Orzack and Sober (1993, 539).

truth of a predictive hypothesis as votes, an argument from robustness to increased confidence might be made as follows:[22]

1. The probability that the majority of simulations in a collection indicates correctly regarding hypothesis H exceeds the probability that any given individual simulation indicates correctly if (*a*) the indications are statistically independent and (*b*) each simulation has the same probability $p > .5$ of giving the correct indication regarding H.[23]
2. *a* and *b* hold for this collection of simulations.
3. If the probability that the majority of simulations in this collection indicates correctly regarding hypothesis H exceeds the probability that any given individual simulation indicates correctly, then if all of the simulations in the collection indicate that H is true, increased confidence in H (beyond the confidence had in light of just one of the simulations' indicating H) is warranted.
4. All of the simulations in this collection indicate that H is true.

$\therefore$ Increased confidence in H is warranted.

When it comes to ensemble climate prediction, the most obvious difficulties with this argument arise in connection with 2. First, while including a model in an ensemble study aimed to discern the truth/falsity of a particular predictive hypothesis would presumably imply a belief that $p > .5$ for that model, in many cases (i.e., for many predictive hypotheses of interest) the basis for such a belief may not be very strong, for reasons already discussed. Second, while climate scientists often do assume that the predictions of different state-of-the-art climate models carry approximately equal evidential weight, it is doubtful that all of these models would have the same probability of indicating correctly regarding the predictive hypothesis of interest.

In addition, the assumption of independence is clearly questionable. In traditional applications of Condorcet's Jury Theorem, independence is assumed to require that voters do not confer with one another, do not base their votes on shared information, do not have similar training and experience, and are not influenced by opinion leaders (see Ladha 1995, 354). How independence should be evaluated in the context of climate

22. For the sake of simplicity, the argument given here targets increased confidence, rather than significantly increased confidence. It is relatively easy to imagine how an analogous argument for significantly increased confidence might be given, once what counts as "significant" is defined in the case of interest.

23. Note that when *a* and *b* obtain, it follows that $p(e|H) > p(e|{\sim}H)$, so a Bayesian argument from robustness to increased confidence (similar to that of sec. 4.1) can also be made.

modeling is still a matter of some discussion (see, e.g., Abramowitz 2010; Pirtle et al. 2010). But many modeling groups do have similar training and experience, and predictions from today's climate models clearly are based on substantial shared information, including but not limited to previously published predictions, which may influence modeling groups as they develop and fine-tune their models (see also Tebaldi and Knutti 2007, 2067–68). Moreover, as noted above, recent investigations have found that errors in simulations of past and present climate produced by today's state-of-the-art climate models show significant correlation (see Knutti et al. 2010; Pennell and Reichler 2011).

There are generalizations of Condorcet's Jury Theorem that have more relaxed assumptions about the competence of voters (e.g., Owen et al. 1989) or that allow certain kinds of dependence among votes (e.g., Ladha 1992, 1995). For instance, while still assuming that voters have the same probability $p > .5$ of voting correctly, Ladha (1992) argues that the probability that the majority vote is correct exceeds p if the average correlation among the voters' choices remains small enough. Perhaps these generalized versions of the theorem hold some promise when it comes to developing a sound argument from robust climate-modeling results to significantly increased confidence in agreed-on predictive hypotheses.[24] But once again, such arguments will require information that is not so easy to come by, such as information about how reliably today's models indicate correctly the truth/falsity of hypotheses of a relevant class.

4.3. A Sampling-Based Perspective. Although it is commonly assumed that ensemble studies somehow involve sampling, it is not obvious how a sampling-based argument from robust model predictions to significantly increased confidence might best be constructed. What follows is one good-faith attempt.

Let q be a set of criteria that can be used to rate any given model's perceived quality as a tool for correctly indicating the truth/falsity of some particular predictive hypothesis H. Assume that today's scientists construct this quality metric q in light of current scientific understanding and computing power—it might take into account whether a model includes particular physical assumptions, how it performs in simulating the behavior of the target system up to now, its spatiotemporal resolution, and so on. Let M_B be the collection of all models, whether already constructed by scientists or not, whose score on q would exceed some chosen threshold;

24. Odenbaugh (forthcoming) considers how a relaxed version of the Condorcet Jury Theorem might be used to analyze the significance of scientific consensus (among experts, rather than models) regarding the existence and causes of global climate change. He too finds that further investigation of (in)dependence is needed.

the models in M_B have features such that they are considered to be, at present, the best models for the predictive purposes at hand. Then the following argument from robustness to increased confidence might be given (see n. 22):

1. In the absence of other overriding evidence, the degree of confidence assigned to predictive hypothesis H should equal f, the fraction of models in M_B whose simulations indicate that H is true.
2. If all of the simulations produced by models in a random sample from M_B are found to agree in indicating that H is true, then an increase in the current estimate of f—and correspondingly an increase in the confidence assigned to H—is warranted.
3. This collection of today's models is a random sample from M_B.
4. The simulations produced by models in this collection all indicate that H is true.

∴ Increased confidence in H is warranted.

Compared to previous arguments, the logic of this one is less tight. While a number of concerns about the argument might be raised, in the context of ensemble climate prediction the most obvious problem is 3, which asserts that some particular ensemble of today's models is a random sample from M_B. This suggests that the scope of some M_B has been identified—that scientists have some sense of the space of models that it encompasses—and that a randomizing procedure was employed when selecting today's models from M_B. But this is not so.

As noted in section 3, today's multimodel ensembles are widely acknowledged to be ensembles of opportunity; any "sampling" by which they are assembled "is neither systematic nor random" (Tebaldi and Knutti 2007, 2068). In fact, according to some climate scientists, "it is not clear how to define a space of possible model configurations of which [today's multimodel ensemble] members are a sample" (Murphy et al. 2007, 1995; see also Parker 2010). Given current uncertainty about how to adequately represent the climate system, any reasonable quality metric that today's climate scientists might specify would allow that many climate models that differ significantly (in their construction) from today's models would qualify for inclusion in M_B. Indeed, today's models may well differ from one another much less than random samples from M_B typically would, which in turn might make them biased estimators of f.[25]

To sum up, various arguments from robustness to significantly increased confidence in an agreed-on predictive hypothesis of interest are possible, but none of the arguments considered above are readily appli-

25. This is assuming that f can be defined for M_B; this issue is not addressed here. If f cannot be defined, then premise 1 is also problematic.

cable in the context of ensemble climate prediction today. Arguments invoking a Bayesian perspective or a generalized version of the Condorcet Jury Theorem show some promise, but further information is needed before these arguments can be advanced.

5. Robustness and Security. A third view regarding the significance of robustness can be found in recent work by Kent Staley (2004). He sets aside the question of whether robustness can increase the strength of evidence for a hypothesis and instead focuses on the *security* of evidence claims—the degree to which an evidence claim is immune to defeat when there is a failure of one or more auxiliary assumptions relied on in reaching it (468). Staley argues that robust test results can increase the security of evidence claims in several ways, one of which will be developed in greater detail here.[26]

Suppose that in light of the results of some test procedure, such as a laboratory experiment or a computer simulation, scientists arrive at an evidence claim, E: "We have evidence of at least strength S for hypothesis H." The strength S might be expressed qualitatively (e.g., weak, strong, conclusive) or perhaps quantitatively.[27] In order to arrive at E, the scientists rely on a set of auxiliary assumptions, A, which includes assumptions about the test procedure (e.g., that the apparatus involved did not malfunction, that the test procedure is of a moderately reliable kind). These auxiliary assumptions are ones that the scientists believe to be true.[28] If any one of the assumptions turns out to be mistaken, the inference from the results of the test procedure to E will need to be reconsidered. Now suppose the scientists conduct a second test of H, and the results of the second test, in conjunction with a set of auxiliary assumptions, A', lead the scientists to the same evidence claim E. That is, as with the first test results, the scientists consider the second test results to provide evidence of at least strength S for hypothesis H. Then as long as A' is at least *partially logically independent* of A—that is, as long as there is at least one assumption in A such that, even if that assumption is false, all assumptions in A' could still be true—then the security of the scientists' evidence claim E will be enhanced since in effect they will have discovered

26. The current analysis expands on the insightful but brief discussion given by Staley (2004, 474–75).

27. Important questions about how the strength of evidence is defined and determined remain to be addressed; for the sake of discussion, it is assumed here that some reasonable and coherent analysis can be given.

28. In fact, scientists may only believe that these assumptions are true enough. For the sake of simplicity, this is ignored in the discussion above; including it would complicate but not undermine the argument.

that there is a "backup route" to E that might remain intact, even if their original inference to E turns out to involve a mistaken assumption (see also Staley 2004, 474–75).[29]

A generalized version of this argument for the case of robust model predictions is as follows:

1. A modeling result r_n enhances the security of an evidence claim E if
   a) E is derivable from r_n in conjunction with a set of auxiliary assumptions, A_n, and
   b) E is derivable from each of modeling results $r_1 \ldots r_{n-1}$, respectively, in conjunction with sets of auxiliary assumptions $A_1 \ldots A_{n-1}$, respectively, and
   c) A_n is partially logically independent of each of $A_1 \ldots A_{n-1}$.
2. 1*a*–1*c* are met in the present case.

∴ The security of E is enhanced.

If 1 is accepted as an analysis of the minimal conditions for increasing security, then the question is whether 1*a*–1*c* are met in the context of ensemble climate prediction today.[30]

Working backward, it seems that 1*c* often is met. In reaching an evidence claim E from any given simulation result, climate scientists will make use of a number of auxiliary assumptions. Assuming that these concern the appropriateness of the model's physical assumptions and numerical solution techniques, the absence of significant programming errors, the reliability of the computing platform on which the model is run, and so on, then the sets of auxiliary assumptions used in conjunction with different simulation results can be expected to differ from one another in various ways since the models producing the simulations will not all reflect the same assumptions about the climate system, will not all be run on the same computing platform, and so on. It seems clear that each set of auxiliary assumptions will be at least partially logically independent of each of the other sets.

29. The mathematical logician typically uses a somewhat different notion of logical independence.

30. Security can be enhanced more or less. Ceteris paribus, the closer the sets of auxiliary assumptions come to being fully logically independent of one another, the more security is enhanced. A set of assumptions A' is *fully logically independent* of another set A iff every assumption in A is such that, if that assumption is false, all of the assumptions in A' could still be true. Security is also enhanced more, ceteris paribus, to the extent that it is not only possible that all of the assumptions in A' could be true, even while some assumption in A is false, but likely that all of the assumptions in A' will be true if some assumption in A is false. Due to space constraints, the discussion above does not consider this quantitative aspect of enhanced security.

For 1*a* and 1*b*, the situation is less clear. In practice, it is often assumed that results from different state-of-the-art climate models each constitute weak (positive) evidence regarding the truth/falsity of interesting predictive hypotheses. (Only together might they even possibly provide strong evidence.) This suggests that, when results from these climate models agree that predictive hypothesis H is true, climate scientists might conclude on the basis of each result, in conjunction with various auxiliary assumptions, that E: there is weak evidence for H.

Unfortunately, it is not clear that the key underlying assumption—that each simulation result has positive evidential relevance—can be given solid justification.[31] The reasons are by now familiar: uncertainty about the importance of various climate system processes, constraints on model construction due to limited computing power, relatively few opportunities to test climate model performance, and difficulty in interpreting the significance of model-data fit in cases where comparisons can be made. While it is true that today's state-of-the-art climate models are constructed using an extensive body of knowledge about the climate system and that they generally deliver results that are (from a subjective point of view) quite plausible in light of current scientific understanding, their individual reliability in indicating the truth/falsity of quantitative predictive hypotheses of the sort that interest today's scientists and decision makers remains significantly uncertain; indeed, it is in part because of this uncertainty that the move to ensembles is made in the first place (see sec. 2).[32] So in the end, even claims of enhanced security seem out of reach in the context of ensemble climate prediction today.

6. Concluding Remarks. The foregoing analysis revealed that, while there are conditions under which robust predictive modeling results have special epistemic significance, scientists are not in a position to argue that those conditions hold in the context of present-day climate modeling; in general, 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

31. Note that even if results from each climate model in an ensemble do have positive evidential relevance, this is not necessarily enough for the argument of sec. 4.1 to work. That argument also depends on the correlations among erroneous indications from the models, and even models that individually are more reliable than chance may nevertheless be more likely to agree that H is true when it is false than when it is true. Thanks to Dan Steel for reminding me to attend to connections between the discussion here and in sec. 4.1.

32. The claim here is not that individual modeling results have negative evidential relevance but that their evidential status (with regard to interesting hypotheses about long-term climate change) is largely unknown.

or that confidence in the hypothesis should be significantly increased or that a claim to have evidence for the hypothesis is now more secure. This is disappointing.

Nevertheless, the analysis did reveal goals for the construction and evaluation of ensembles—whether in the study of climate change or in any other context—such that robust results will have desired epistemic significance. 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 (in order to construct the ensemble) should then be exhaustive, if possible, or else aimed to produce maximally different results. In other cases, when ensembles are not carefully constructed in this way, the goal might be to obtain extensive error statistics regarding the past performance of an ensemble in indicating the truth/falsity of hypotheses of the relevant sort; this in turn will require careful consideration of which hypotheses are relevant.

When it comes to ensemble climate prediction, the prospects for reaching these goals in the near future seem slim. Certainly the design of multimodel ensemble studies could be improved, aiming to better sample recognized uncertainty about how to adequately represent the climate system for a given predictive task, but the specification and deployment of ensembles that can (with justification) be expected to include adequate models—while still giving robust results—seems likely to remain beyond scientific understanding for some time. Likewise, in the near term it will be difficult to obtain desired error statistics for climate ensembles, given the long-term nature of the predictions of interest, the limited time span for which reliable observational data are available, the lack of comprehensiveness of these data (leading to reanalysis), and the prior practice of tuning.[33]

That said, prospects seem substantially brighter in some other predictive modeling contexts. For instance, when it comes to hypotheses about the next opportunities to see solar eclipses from various locations on earth, today's physicists might well have sufficient background knowledge to design ensemble studies that can be expected to meet the likely adequacy condition (e.g., studies that explore parameter and initial condition uncertainty, perhaps even using a single model structure). Likewise, today's weather forecasters might collect extensive error statistics on the performance of ensemble weather-forecasting systems, providing good evidence

33. 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.

that $p(e|H) \gg p(e|\sim H)$ for quantitative hypotheses about next-day high temperatures in a given locale. In cases like these, robust model predictions may well have special epistemic significance.

REFERENCES

Abramowitz, Gab. 2010. "Model Independence in Multi-Model Ensemble Prediction." *Australian Meteorological and Oceanographic Journal* 59:3–6.

BBC (British Broadcasting Corporation). 2010. "Climate Change Experiment Results," http://www.bbc.co.uk/sn/climateexperiment.

Brohan, Philip, J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones. 2006. "Uncertainty Estimates in Regional and Global Observed Temperature Changes: A New Data Set from 1850." *Journal of Geophysical Research* 111:D12106, doi:10.1029/2005JD006548.

Cartwright, Nancy. 1991. "Replicability, Reproducibility, and Robustness: Comments on Harry Collins." *History of Political Economy* 23:143–55.

Christensen, Jens Hesselbjerg, et al. 2007. "Regional Climate Projections." In Solomon et al. 2007, 847–940.

Edwards, Paul N. 1999. "Global Climate Science, Uncertainty and Data: Data-Laden Models, Model-Filtered Data." *Science as Culture* 8:437–72.

———. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Gleckler, Peter J., Karl A. Taylor, and Charles Doutriaux. 2008. "Performance Metrics for Climate Models." *Journal of Geophysical Research* 113:D06104.

Harris, Todd. 2003. "Data Models and the Acquisition and Manipulation of Data." *Philosophy of Science* 70:1508–17.

Judd, Kevin, Leonard A. Smith, and Antje Weisheimer. 2007. "How Good Is an Ensemble at Capturing Truth? Using Bounding Boxes for Forecast Evaluation." *Quarterly Journal of the Royal Meteorological Society* 133:1309–25.

Knutti, Reto. 2008. "Why Are Climate Models Reproducing the Observed Global Surface Warming so Well?" *Geophysical Research Letters* 35:L18704, doi:10.1029/2008GL034932.

Knutti, Reto, et al. 2008. "A Review of Uncertainties in Global Temperature Projections over the Twenty-first Century." *Journal of Climate* 21:2651–63.

———. 2010. "Challenges in Combining Projections from Multiple Climate Models." *Journal of Climate* 23:2739–58.

Ladha, Krishna K. 1992. "The Condorcet Jury Theorem, Free Speech, and Correlated Votes." *American Journal of Political Science* 36 (3): 617–34.

———. 1995. "Information Pooling through Majority-Rule Voting: Condorcet's Jury Theorem." *Journal of Economic Behavior and Organization* 26:353–72.

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

Meehl, Gerald A., et al. 2007. "Global Climate Projections." In Solomon et al. 2007, 747–846.

Muldoon, Ryan. 2007. "Robust Simulations." *Philosophy of Science* 74:873–83.

Murphy, James M., et al. 2007. "A Methodology for Probabilistic Predictions of Regional Climate Change from Perturbed Physics Ensembles." *Philosophical Transactions of the Royal Society* A 365:1993–2028.

Odenbaugh, Jay. Forthcoming. "Consensus, Climate, and Contrarians." In *Topics in Contemporary Philosophy: Environment*. Cambridge, MA: MIT Press.

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

Owen, Guillermo, Bernard Grofman, and Scott L. Feld. 1989. "Proving a Distribution-Free Generalization of the Condorcet Jury Theorem." *Mathematical Social Sciences* 17:1–16.

Parker, Wendy S. 2006. "Understanding Pluralism in Climate Modeling." *Foundations of Science* 11:349–68.

———. 2009. "Confirmation and Adequacy-for-Purpose in Climate Modelling." *Proceedings of the Aristotelian Society* 83 (Suppl.): 233–49.
———. 2010. "Whose Probabilities? Predicting Climate Change with Ensembles of Models." *Philosophy of Science* 77 (5): 985–97.
Pennell, Christopher, and Thomas Reichler. 2011. "On the Effective Number of Climate Models." *Journal of Climate* 24:2358–67.
Pirtle, Zachary, Ryan Meyer, and Andrew Hamilton. 2010. "What Does It Mean When Climate Models Agree? A Case for Assessing Independence among General Circulation Models." *Environmental Science and Policy* 13:351–61.
Randall, David A., et al. 2007. "Climate Models and Their Evaluation." In Solomon et al. 2007, 589–662.
Solomon, Susan, et al., eds. 2007. *Climate Change 2007: The Physical Science Basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. New York: Cambridge University Press.
Stainforth, David A., et al. 2005. "Uncertainty in Predictions of the Climate Response to Rising Levels of Greenhouse Gases." *Nature* 433:403–6.
Staley, Kent. 2004. "Robust Evidence and Secure Evidence Claims." *Philosophy of Science* 71:467–88.
Suppes, Patrick. 1962. "Models of Data." In *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, ed. Ernest Nagel, Patrick Suppes, and Alfred Tarski, 252–61. Stanford, CA: Stanford University Press.
Tebaldi, Claudia, and Reto Knutti. 2007. "The Use of the Multi-Model Ensemble in Probabilistic Climate Projections." *Philosophical Transactions of the Royal Society* A 365: 2053–75.
Tebaldi, Claudia, Richard L. Smith, Doug Nychka, and Linda O. Mearns. 2005. "Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach." *Journal of Climate* 18:1524–40.
Weisberg, Michael. 2006. "Robustness Analysis." *Philosophy of Science* 73:730–42.
Wimsatt, William C. 1981/2007. "Robustness, Reliability, and Overdetermination." In *Re-engineering Philosophy for Limited Beings*. Repr., Cambridge, MA: Harvard University Press.
Woodward, James. 2006. "Some Varieties of Robustness." *Journal of Economic Methodology* 13:219–40.