13

Communicating Uncertainty to Policymakers: The Ineliminable Role of Values

Eric Winsberg

13.1 Introduction

Over the last several years, there has been an explosion of interest and attention devoted to the problem of Uncertainty Quantification (UQ) in climate science—that is, to giving quantitative estimates of the degree of uncertainty associated with the predictions of global and regional climate models. The technical challenges associated with this project are formidable: the real data sets against which model runs are evaluated are large, patchy, and involve a healthy mixture of direct and proxy data; the computational models themselves are enormous, and hence the number of model instances that can be run is minuscule and sparsely distributed in the solution space that needs to be explored; the parameter space that we would like to sample is vast and multidimensional;

E. Winsberg (✉)
Department of Philosophy, University of South Florida,
Tampa, FL, USA

E.A. Lloyd, E. Winsberg (eds.), *Climate Modelling*,
https://doi.org/10.1007/978-3-319-65058-6_13

and the structural variation that exists amongst the existing set of models is substantial but poorly understood. Understandably, therefore, the statistical community that has engaged itself with this project has devoted itself primarily to overcoming some of these technical challenges.

So why is UQ so important in climate science? What goals are we trying to meet with UQ, and are they likely to be met? Those who are interested in these questions might benefit from a close look at some of the recent philosophical literature on the role of social values in science. UQ, I suggest, is first and foremost a tool for communicating knowledge from experts to policymakers. Experts, in this case, climate scientists and climate modelers, have knowledge about the climate. In one sense, therefore, they are the people who ought to be considered best situated to make decisions about what we ought to do in matters related to climate. But in another sense, they are not.

Consider the fact that we often evaluate the wisdom of pursuing various climate adaptation strategies, such as: how to manage the problem of glacial lake outburst floods, one of the many possible dangers of regional climate changes. These floods occur when a dam (consisting of glacier ice and a terminal moraine) containing a glacial lake fails. Should a local community threatened by a possible flood replace the terminal moraine with a concrete dam? The answer to this question depends in part on the likelihood of the glacier melting and the existing (natural) dam bursting, which climate scientists, who have the most expertise about the future of the local regional climate, would be in the best position to address. It also surely depends, however, on the cost of building the dam, and on the likely damage that would ensue if the dam were to break. Just as much, it might depend on how the relevant stakeholders weigh the present costs against the future damages. And so while, on the one hand, we would like the people making the decision to have the most expertise possible, we also, on the other hand, want the decision to be made by people who represent our interests, whoever "we" might be. Making decisions about, for example, climate adaptation strategies, therefore, requires a mixture of the relevant expertise and the capacity to represent the values of the people on whose behalf one is making the decision.[1] But there is rarely any single group of people who obviously possess both of these properties.

UQ, as we will see in what follows, is in principle one way in which these different capacities can be kept separate. One clear motivation for solving the problems of UQ, in other words, is to maintain this division of labor between the epistemic and the normative—between the people who have the pure scientific expertise and the people with the legitimate ability to represent the values of the relevant stakeholders. And so if we want to understand where the need to produce quantitative estimates of uncertainty comes from, we need to delve into the role of social values in the administration of scientific expertise.

13.2 Science and Social Values

What do we mean, first of all, by "social values"? Social values, I take it, are the estimations of any agent or group of agents of what is important and valuable—in the typical social and ethical senses—and of what is to be avoided, and to what degree. What value does one assign to economic growth, on the one hand, and to the degree to which we would like to avoid various environmental risks, on the other? In the language of decision theory, by social values we mean the various marginal utilities one assigns to events and outcomes. The point of the word "social" in "social values" is primarily to flag the difference between these values and what Ernan McMullin once called "epistemic values," like simplicity, fruitfulness, and so forth (1983). But I do not want to beg any questions about whether or not values that are paradigmatically ethical or social can or cannot or should or should not play important epistemic roles. So, I prefer not to use that vocabulary. I talk instead about social and ethical values when I am referring to things that are valued for paradigmatically social or ethical reasons. I do not carefully distinguish, in this chapter, between the social and the ethical.[2]

It is uncontroversial that social and ethical values play a role in science. When we set constraints on experimentation, for example, or when we decide which projects to pursue and which projects to ignore, these decisions uncontroversially reflect social values. But the philosophically controversial question about social and ethical values is about the degree to which they are involved (or better put: the degree to which they are necessarily involved, or inevitably involved, and perhaps most importantly: uncorrectibly involved) in the appraisal of hypotheses or in reaching other conclusions that are internal to science, and that necessarily also involves scientific expertise. This is the question, after all, of the degree to which the epistemic and the normative can be kept apart.

This is a question of some importance because we would like to believe that only experts should have a say in what we ought to believe about the natural world. But we also think that it is *not* experts, or at least not experts qua experts, who should get to say what is important to us, or what is valuable or has utility. Such a division of labor, however, is only possible to the extent that the appraisal of scientific hypotheses, and the consideration of other matters that require scientific expertise, can be carried out in a manner that is free of the influence of social and ethical values.

Philosophers of science of various stripes have mounted a variety of arguments to the effect that the epistemic matter of appraising scientific claims of various kinds cannot be kept free of social and ethical values. Here, we will be concerned only with one such line of argument—one that is closely connected to the issue of UQ—that goes back to the midcentury work of statistician C. West Churchman (1949, 1953) and philosopher of science Richard Rudner (1953).[3] This line of argument is now frequently referred to as the argument from inductive risk. It was first articulated by Rudner in the following schematic form:

1. The scientist qua scientist accepts or rejects hypotheses.
2. No scientific hypothesis is ever completely (with 100% certainty) verified.
3. The decision to either accept or reject a hypothesis depends upon whether the evidence is sufficiently strong.

4. Whether the evidence is *sufficiently* strong is "a function of the *importance*, in a typically ethical sense, of making a mistake in accepting or rejecting the hypothesis."
5. Therefore, the scientist qua scientist makes value judgments.

Rudner's oft-repeated example compared two hypotheses: (1) that a toxic ingredient of a drug is not present in lethal quantity in some resource, (2) that a certain lot of machine stamped belt buckles is not defective. Rudner's conclusion was that "how sure we need to be before we accept a hypothesis will depend upon how serious a mistake it would be" to accept it and have it turn out false (1953, p. 2). We can easily translate Rudner's lesson into an example from climate science: consider a prediction that, given future emissions trends, a certain regional climate outcome will occur. Should we accept the hypothesis, say, that a particular glacial lake dam will burst in the next 50 years? Suppose that if we accept the hypothesis, we will replace the moraine with a concrete dam. But whether we want to build the dam will depend not only on our degree of evidence for the hypothesis, but also on how we would measure the severity of the consequences of building the dam, and having the glacier not melt, vs. not building the dam, and having the glacier melt. Rudner would have us conclude that as long as the evidence is not 100% conclusive, we cannot justifiably accept or reject the hypothesis without making reference to our social and ethical values.

The best-known reply to Rudner's argument came from logician and decision theorist Richard Jeffrey (1956). Jeffrey argued that the first premise of Rudner's argument—that it is the proper role of the scientist qua scientist to accept and reject hypotheses—is false. The proper role of scientists, he urged, is to assign probabilities to hypotheses with respect to the currently available evidence. Others—for example, policymakers—can attach values or utilities to various possible outcomes or states of affairs and, in conjunction with the probabilities provided by scientists, decide how to act.

In providing this response to Rudner, Jeffrey was making it clear that an important purpose of probabilistic forecasts is to separate practice from theory and the normative from the epistemic, so that social values can be relegated entirely to the domain of practice, and cordoned off from the domain of scientific expertise. If the scientist accepts or rejects a hypothesis, then Rudner has shown that normative considerations cannot be excluded from that decision process. In contrast, if scientists don't have to bring any normative considerations to bear when they assign probabilities to a hypothesis, then the normative considerations can be cordoned off. It should now be clear why I said at the beginning that UQ is first and foremost a tool for communicating knowledge from experts to policymakers. It is a tool for dividing our intellectual labor. If we were entirely comfortable simply letting experts qua experts decide for us how we should act, then we would not have such an acute need for UQ.

It is clear, however, that Jeffrey did not anticipate the difficulties that modern climate science would have with the task that he expected to be straightforward and value free, the assignment of probability with respect to the available evidence. There are many differences between the kinds of examples that Rudner and Jeffrey had in mind and the kinds of situations faced by climate scientists. For one, Rudner and Jeffrey discuss cases in which we need the probability of the truth or falsity of a single hypothesis, but climate scientists generally are faced with having to assign probability distributions over a space of possible outcomes. I believe, however, that the most significant difference between the classic kind of inductive reasoning Jeffrey had in mind (in which the probabilities scientists are meant to offer are their subjective degrees of belief based on the available evidence) and the contemporary situation in climate science is the extent to which epistemic agency in climate science is distributed across a wide range of scientists and tools.

Here, I am pursuing a theme that is at the heart of much of my work on computationally intensive science (2010): that this new kind of science requires of philosophers new ways of thinking about old epistemological issues. These kinds of claims can also be found in the work of Elisabeth Lloyd (in this volume and elsewhere, 2012, 2015), where she argues that recent developments in science require that we adopt what she calls "complex empiricism."

I will return to the issue of how climate science differs from the kind of science envisioned by Jeffrey later (especially in Sect. 13.6), but for now, we should turn to what I would claim are typical efforts in climate science to deliver probabilistic forecasts and see how they fare with respect to Jeffrey's goal of using probabilities to divide labor between the epistemic and the normative.

13.3 Uncertainty in Climate Science

Where do probabilistic forecasts in climate science come from? We should begin with a discussion of the sources of uncertainty in climate models. There are two main sources that concern us here: *structural model uncertainty* and *parameter uncertainty*. While the construction of climate models is guided by basic science—science in which we have a great deal of confidence—these models also incorporate a barrage of auxiliary assumptions, approximations, and parameterizations, all of which contribute to a degree of uncertainty about the predictions of these models. This source of uncertainty is often called "structural model uncertainty."

Next, complex models involve large sets of parameters or aspects of the model that have to be quantified before the model can be used to run a simulation of a climate system. We are often highly uncertain about what the best value for many of these parameters is, and hence, even if we had at our disposal a model with ideal (or perfect) structure, we would still be uncertain about the behavior of the real system we are modeling, because the same model structure will make different predictions for different values of the parameters. Uncertainty from this source is called "parameter uncertainty."[4]

Most efforts in contemporary climate science to measure these two sources of uncertainty focus on what one might call "sampling methods." In practice, in large part because of the high computational cost of each model run, these methods are extremely technically sophisticated, but in principle they are rather straightforward.

I can best illustrate the idea of sampling methods with an example regarding parameter uncertainty: consider a simulation model with one parameter and several variables.[5] If one had a data set against which to benchmark the model, one could assign a weighted score to each value of the parameter based on how well it retrodicted values of the variables in the available data set. Based on this score, one could then assign a probability to each value of the parameter. Crudely speaking, what we are doing in an example like this is observing the frequency with which each value of the parameter is successful in replicating known data—how many of the variables does it get right? with how much accuracy? over what portion of the time history of the data set?—and then weighting the probability of the parameter taking this value in our distribution in proportion to how well it had fared in those tests.

The case of structural model uncertainty is similar. The most common method of estimating the degree of structural uncertainties in the predictions of climate models is a set of sampling methods called "ensemble methods," which examine the degree of variation in the predictions of the existing set of climate models. By looking at the average prediction of the set of models and calculating their standard deviation, one can produce a probability distribution for every value that the models calculate.

13.4 Some Worries About the Standard Methods

There are reasons to doubt, however, that these simple methods for estimating structural model uncertainty and parameter uncertainty are conceptually coherent. Signs of this are visible in the results that have been produced. These signs have been particularly well noted by climate scientists Claudia Tebaldi and Reto Knutti (2007). Tebaldi and Knutti have noted, in the first instance, that many studies founded on the same basic principles produce radically different probability distributions. One of their very illustrative charts shows a comparison of four different attempts to quantify the degree of uncertainty associated with the predictions of climate models for a variety of scenarios, regions, and predictive tasks. Tebaldi and Knutti note the wide range of the various estimates.

Beyond the graphical display of the wide variety of possible results one can get from ensemble averages, there are various statistical analyses one can perform on ensemble sample characteristics that cast doubt on their reliability for naïve statistical analysis. These are summarized in Tebaldi and Knutti. I quote their conclusions here:

> Recent coordinated efforts, in which numerous general circulation climate models have been run for a common set of experiments, have produced large data sets of projections of future climate for various scenarios. Those multimodel ensembles sample initial conditions, parameters, and structural uncertainties in the model design, and they have prompted a variety of approaches to quantifying uncertainty in future climate change … This study outlines the motivation for using multimodel ensembles and discusses various challenges in interpreting them. Among these challenges are that the number of models in these ensembles is usually small, their distribution in the model or parameter space is unclear, and that extreme behavior is often not sampled … While the multimodel average appears to still be useful in some situations, these results show that more quantitative methods to evaluate model performance are critical to maximize the value of climate change projections from global models. (2007, p. 2053)

Indeed, I would argue that there are four reasons to suspect that ensemble methods are not a conceptually coherent set of methods:

1. Ensemble methods either assume that all models are equally good, or they assume that the set of available methods can be relatively weighted.
2. Ensemble methods assume that, in some relevant respect, the set of available models represent something like a sample of independent draws from the space of possible model structures.
3. Climate models have shared histories that are very hard to sort out.
4. Climate modelers have a herd mentality about success.

I discuss each of these four reasons in what follows. But, first, consider a simple example that mirrors all four: suppose that you would like to know the length of a barn. You have one tape measure and many carpenters. You decide that the best way to estimate the length of the barn is to send each carpenter out to measure the length and then take the average. There are four problems with this strategy. First, it assumes that each carpenter is equally good at measuring. But what if some of the carpenters have been drinking on the job? Perhaps you could weight the degree to which their measurements play a role in the average in inverse proportion to how much they have had to drink. But what if, in addition to drinking, some have also been sniffing from the fuel tank? How do you weight these relative influences? Second, you are assuming that each carpenter's measurement is independently scattered around the real value. But why think this? What if there is a systematic error in their measurements? Perhaps there is something wrong with the tape measure that systematically distorts them. Third (and relatedly), what if all the carpenters went to the same carpentry school, and they were all taught the same faulty method for what to do when the barn is longer than the tape measure? And fourth, what if, before recording their value, each carpenter looks at the running average of the previous measurements, and if theirs deviates too much, they tweak it to keep from getting the reputation as a poor measurer?

All of these sorts of problems play a significant role—both individually, but especially jointly—in making ensemble statistical methods in climate science conceptually troubled. I will now discuss the role of each of them in climate science in detail:

1. *Ensemble methods either assume that all models are equally good, or they assume that the set of available methods can be relatively weighted.*

If you are going to use an ensemble of climate models to produce a probability distribution, you ought to have some grounds for believing that all of them ought to be given equal weight in the ensemble. Failing that, you ought to have some principled way to weight them. But no such thing seems to exist. While there is widespread agreement among climate scientists that some models are better than others, quantifying this intuition seems to be particularly difficult. It is not difficult to see why.

As Peter Gleckler et al. (2008) note, no single metric of success is likely to be useful for all applications. Their beautiful illustrations show the success of various models for various prediction tasks. It is fairly clear that while there are some unambiguous flops on the list, there is no unambiguous winner, nor a clear way to rank them.

2. *Ensemble methods assume that, in some relevant respect, the set of available models represent something like a sample of independent draws from the space of possible model structures.*

This is surely the greatest problem with ensemble statistical methods. The average and standard deviation of a set of trials is only meaningful if those trials represent a random sample of independent draws from the relevant space—in this case, the space of possible model structures. Many commentators have noted that this assumption is not met by the set of climate models on the market. In fact, I would argue, it is not clear what this would even mean in this case. What, after all, is the space of possible model structures? And why would we want to sample randomly from this? After all, we want our models to be as physically realistic as possible, not random. Perhaps we are meant to assume, instead, that the existing models are randomly distributed around the ideal model, in some kind of normal distribution, on analogy to measurement theory. But modeling isn't measurement, and so there is very little reason to think this assumption holds.[6]

3. *Climate models have shared histories that are very hard to sort out.*

Large clusters of the climate models on the market have shared histories, which is one reason for doubting that existing models are randomly distributed around an ideal model.[7] Some of them share code. Scientists move from one lab to another and bring ideas with them. Various parts of climate models come from a common toolbox of techniques, and so forth. Worse still, we do not even have a systematic understanding of these interrelations. So, it is not just the fact that most current statistical ensemble methods are naive with respect to these effects; it's also that it is far from obvious that we have the background knowledge we would need to eliminate this naïveté and therefore account for them statistically.

4. *Climate modelers have a herd mentality about success.*

Most climate models are highly tunable with respect to some of their variables, and to the extent that no climate lab wants to be the oddball on the block, there is significant pressure to tune one's model to the crowd. This kind of phenomenon has historical precedent.[8] In 1939, Walter Shewhart published a chart of the history of measurement of the speed of light. The chart shows a steady convergence of measured values that is not well explained by their actual success. Myles Allen puts the point like this: "If modeling groups, either consciously or by 'natural selection,' are tuning their flagship models to fit the same observations, spread of predictions becomes meaningless: eventually they will all converge to a delta-function" (2008).

13.5 The Inevitability of Values: Douglas *contra* Jeffrey

What should we make of all of these problems from the point of view of the Rudner–Jeffrey debate? This much should be clear: Jeffrey's goal of separating the epistemic from the normative cannot be achieved using UQ based on statistical ensemble methods. But Heather Douglas's (2000) discussion of the debate about science and values should have made this clear from the beginning.[9]

Douglas noted a flaw in Jeffrey's response to Rudner: scientists often have to make methodological choices that do not lie on a continuum. Suppose I am investigating the hypothesis that substance X causes disease D in rats. I give an experimental group of rats a large dose of X and then perform biopsies to determine what percentage has disease D. How do I perform the biopsy? Suppose that there are two staining techniques I could use. One is more sensitive and the other is more specific—one produces more false positives and the other more false negatives. Which one should I choose? Douglas notes that which one I choose will depend on my inductive risk profile. To the extent that I weigh more heavily the consequences of saying that the hypothesis is false if it is in fact true, I will chose the stain with more false positives, and vice versa. But that, of course, depends on my social and ethical values. Social and ethical values therefore play an inevitable role in science.

Now, inevitability is always relative to some fixed set of background conditions, and the set of background conditions Douglas assumes include the use of something like classical statistical methods. If I have some predetermined level of confidence, *alpha*, say .05, then which staining method I use will raise or lower, respectively, the likelihood that the hypothesis will be accepted. What if, on the other hand, all toxicologists were good Bayesians of the kind that Jeffrey almost surely had in mind? What is the argument that they could not use their expert judgment, having chosen whatever staining method they like, to factor in the specificity and sensitivity of the method when they use the evidence they acquire to update their degrees of belief about the hypothesis? In principle, surely they could. By factoring the specificity and sensitivity of the method into their degrees of belief, they are essentially eliminating or "screening out" the influence of the social or ethical values that otherwise would have been present. And if they could do this, social and ethical values, at least the kind that normally play a role in the balance of inductive risks, would not *have* to play a role in their assessments of the probabilities.[10] Let us call this the Bayesian response to the Douglas challenge (BRDC).

Back to climate science: another way to look at the problem with ensemble statistical methods is that they have no hope of skirting Douglas's challenge and hence no hope of fulfilling their intended role— to divide the epistemic from the normative. To the extent that we use sampling methods and ensemble averages, we are doomed to embed past methodological choices of climate modelers into our UQ. And, for just the reasons that Douglas highlights, along with some others, methodological choices often *need* to reflect judgments of social and ethical values.

There are at least two ways in which methodological choices in the construction of climate models will often ineliminably reflect value judgments in the typically social or ethical sense.

1. Model choices have reflected balances of inductive risk.
2. Models have been optimized, over their history, to particular purposes, and to particular metrics of success.

The first point should be obvious from our discussion of Douglas. When a climate modeler is confronted with a choice between two ways of solving a modeling problem, she may be aware that each choice strikes a different balance of inductive risks with respect to a problem that concerns her at the time. Choosing which way to go, in such a circumstance, will have to reflect a value judgment. This will always be true so long as a methodological choice between methods A and B are not epistemologically *forced* in the following sense: while option A can be justified on the grounds that it is *less* likely to predict, say, outcome O, than B is when O *will not* in fact occur, option B could also be preferred on the grounds that it is *more* likely to predict O if O *will* in fact occur. So, to return to our old example, if the central question is whether or not some glacial dam will burst, there will often be a modeling choice that will make it less likely to predict that the dam will burst when it fact it won't, and a different modeling choice that will make it less likely to predict that the dam won't burst when in fact it will. In such a situation, neither choice will be "objectively correct," since the correct choice will depend on which of the above two situations is deemed more undesirable.

As to the second point, when a modeler is confronted with a methodological choice, she will have to decide which metric of success to use when evaluating the likely success of the various possibilities. And it is hard to see how choosing a metric of success will not reflect a social or ethical value judgment, or possibly even a response to a political pressure, about which prediction task is more "important" (in a not purely epistemic sense.) Suppose choice A makes a model that looks better at matching existing precipitation data, but choice B better matches temperature data. A modeler will need to decide which prediction task is more important in order to decide which method of evaluation to use and that will influence the methodological choice she makes.

13.6 Three Features of Climate Models

The discussion thus far should make two things clear. First, ensemble sampling approaches to Uncertainty Quantification (UQ) are founded on conceptually shaky ground. Second, and perhaps more importantly, they do not enable UQ to fulfill its primary function, namely, to divide the epistemic from the normative in the way that Jeffrey expected probabilistic forecasts to do. And they fail for just the reasons that Douglas has made perspicuous: because they ossify past methodological choices (which themselves can reflect balances of inductive risk and other social and ethical values) into "objective" probabilistic facts.

This raises, of course, the possibility that climate UQ could respond to these challenges with something akin to the BRDC: by adopting a thoroughly Bayesian approach to quantifying probabilities. Recall the problem faced by Douglas' hypothetical toxicologist. If she is looking for statistically significant evidence that substance X is causing disease D at some predetermined level of "statistical significance," then a particular choice of staining method will either raise or lower the probability of finding that result. But if she has some prior probability for the hypothesis, and updates it in response to the evidence acquired in the biopsies, then the choice of staining method needn't influence those probabilities. Similarly, one might hope, the Bayesian climate scientist might avoid the fundamental problem of any approach founded on "objective" ensemble averaging: that past methodological choices become features of the ensemble and hence exert a pull on the estimated uncertainties.

Indeed, this approach has been endorsed by several commentators.[11] Unfortunately, the role of genuinely subjective Bayesian approaches to climate UQ has been primarily in theoretical discussions of what to do; they have not been widely drawn on to produce actual estimates that one sees published and that are delivered to policymakers. Here, I identify some of the difficulties that might explain why these methods are not used in the field. Genuinely Bayesian approaches to UQ in climate science, in which the probabilities delivered reflect the expert judgment of climate scientists rather than observed frequencies of model outputs, face several difficulties. In particular, the difficulties arise as a consequence of three features of climate models: their massive size and complexity; the extent to which epistemic agency in climate modeling is distributed, in time and space, and across a wide range of individuals; and the degree to which methodological choices in climate models are generatively entrenched. Let me take each of these features in turn.

Size and Complexity

Climate models are enormous and complex. Take one of the state-of-the-art American models, NOAA's GFDL CM2.x. The computational model itself contains over a million lines of code. There are over a thousand different parameter options. It is said to feature modules that are “constantly changing” and as well as hundreds of initialization files that contain “incomplete documentation” (Dunne 2006, p. 00). It is also said to contain novel component modules written by over 100 different people. Just loading the input data into a simulation run takes over two hours. Using over 100 processors running in parallel, it takes weeks to produce one model run out to the year 2100 and months to reproduce thousands of years of paleoclimate (Dunne 2006). Storing the data from a state of the art global climate model (GCM) every five minutes can produce tens of terabytes per model year.

Another aspect of the models’ complexity is their extreme “fuzzy modularity” (Lenhard and Winsberg 2010). In general, a modern state-of-the-art climate model is a model with a theoretical core that is surrounded and supplemented by various submodels that themselves have grown into complex entities. Their overall interaction determines the dynamics—and these interactions are themselves quite complex. The coupling of atmospheric and oceanic circulation models, for example, is recognized as one of the milestones of climate modeling (leading to so-called coupled general circulation models). Both components had an independent modeling history, including an independent calibration of their respective model performance. Putting them together was a difficult task because the two submodels now interfered dynamically with each other.[12]

Today, atmospheric GCMs have lost their central place and given way to a deliberately modular architecture of coupled models that comprise a number of highly interactive submodels, like atmosphere, oceans, or ice cover. In this architecture, the single models act (ideally!) as interchangeable modules.[13] This marks a turn from a reliance on one physical core—the fundamental equations of atmospheric circulation dynamics—to the development of a more networked picture of interacting models from different disciplines (see Küppers and Lenhard 2006).

In sum, climate models are made up of a variety of modules and submodels. There is a module for the general circulation of the atmosphere, a module for cloud formation, for the dynamics of sea and land ice, for effects of vegetation, and many more. Each of them, in turn, includes a mixture of principled science and parameterizations. And it is the interaction of these components that generates the overall observable dynamics in simulation runs. The results of these modules are not first gathered independently and then only after that synthesized. Rather, data are continuously exchanged between all modules during the runtime of the simulation.[14] The overall dynamics of one global climate model is the complex result of the interaction of the modules—not the interaction of the results of the modules. This is why I modify the word "modularity" with the warning flag "fuzzy" when I talk about the modularity of climate models: due to interactivity and the phenomenon of "balance of approximations," modularity does not break down a complex system into separately manageable pieces.[15]

Distributed Epistemic Agency

Climate models reflect the work of hundreds of researchers working in different physical locations and at different times. They combine incredibly diverse kinds of expertise, including climatology, meteorology, atmospheric dynamics, atmospheric physics, atmospheric chemistry, solar physics, historical climatology, geophysics, geochemistry, geology, soil science, oceanography, glaciology, paleoclimatology, ecology, biogeography, biochemistry, computer science, mathematical and numerical modeling, time series analysis, and so forth.

Epistemic agency in climate science is not only distributed across space (the science behind model modules comes from a variety of labs around the world) and domains of expertise but also across time. No state-of-the-art, coupled atmosphere-ocean GCM (AOGCM) is literally built from the ground up in one short surveyable unit of time. They are assemblages of methods, modules, parameterization schemes, initial data packages, bits of code, coupling schemes, and so forth that have been built, tested, evaluated, and credentialed over years or even decades of work by climate scientists, mathematicians, and computer scientists of all stripes.[16]

No single person, indeed no group of people in any one place, at one time, or from any one field of expertise, is in a position to speak authoritatively about any AOGCM in its entirety.[17]

Methodological Choices are Generatively Entrenched

In our (2010), Johannes Lenhard and I argued that complex climate models acquire an intrinsically historical character and show path-dependency. The choices that modelers and programmers make at one time about how to solve particular problems of implementation have effects on what options will be available for solving problems that arise at a later time. And they will have effects on what strategies will succeed and fail. This feature of climate models, indeed, has lead climate scientists such as Smith (2002) and Palmer (2001) to articulate the worry that differences between models are concealed in code that cannot be closely investigated in practice.

Of course the modelers could—in principle—re-work the entire code. The point is, however, that in even moderately complex cases, this is not a viable option for practical reasons. At best, this would be far too tedious and time-consuming. Conceivably, we would not even know how to proceed. So in the end, each step in the model building process, and how successful it might be, could very well depend on the particular way previous steps were carried out—because the previous steps are unlikely to be completely disentangled and redone.

This is the sense in which modeling choices are generatively entrenched. Modeling choices that are made early in the model construction process have effects on the models at later times in unpredictable ways. And the success of modeling choices at later times depends in unpredictable ways on earlier modeling choices.

13.7 Summary

To summarize then, state-of-the-art global climate models are highly complex, they are the result of massively distributed epistemic labors, and they arise from a long chain of generatively entrenched methodological choices whose effects are epistemically inscrutable. These three features, I would now argue, make the BRDC very difficult to pull off with respect to climate science.

13.8 The Failure of the BRDC in Climate Science

Recall how the BRDC is meant to go: Rudner argues that the scientist who accepts or rejects hypotheses has to make value judgments. Jeffrey replies that she should only assign probabilities to hypotheses on the basis of the available evidence, and, in so doing, avoid making value judgments. Douglas argues that scientists make methodological choices, and that these choices will become embedded in the mix of elements that give rise to estimates of probabilities that come from classical, as opposed to Bayesian, statistics. Since those methodological choices will involve a balance of inductive risks, the scientist cannot avoid value judgments. The BRDC suggests that scientists avoid employing any deterministic algorithm that will transmit methodological choices into probabilities (like employing a classical statistical hypothesis test in the toxicology case, or employing ensemble averages in the climate case), and should instead rely on their expert judgment to assess what the appropriate degree of belief in a hypothesis is given that a particular methodological choice is made and resultant evidence acquired. The probabilities such a scientist would offer should be the scientist's subjective degree of belief, one that has been conditionalized on the available evidence.

Unfortunately, large groups of individuals, distributed across space and time, do not possess subjective degrees of belief. Subjective Bayesian probabilities need to be "owned" by one individual epistemic agent (Parker 2011), or, at the very least, by manageably small epistemic groups.[18] But the three features of global climate models I have pointed to—that they are highly complex, are the result of massively distributed epistemic labors, and arise from a long chain of generatively entrenched methodological choices—make it seem implausible, at least to me, that any individual epistemic agent[19] will ever be in good position to have a useful degree of expert judgment of the kind required to implement the BRDC.[20] The BRDC precisely requires that *one epistemic agent* be capable of making an informed judgment about how every single methodological choice on which a climate model is built ought to influence his or her degree of belief in a hypothesis that he or she is evaluating with the use of that model. But how can we expect any individual, or well-defined group of experts, to do this successfully when faced with massively complex models, built over large expanses of space and time, and built on methodological choices that have become generatively entrenched, and hence epistemically inscrutable?

The argument thus far, then, can be summarized as follows. Climate science, and the construction of climate models, like almost all of science, is full of unforced methodological choices. And like in the rest of science, these choices often reflect priorities with respect to predictive power, and balances of inductive risk. There is nothing new here. It is plausible to suppose, moreover, that Jeffrey understood this to be a feature of much of science, and still believed, pace Douglas, that the subjective Bayesian had available a defense of value-free science: once the methodological choices are made, the scientists qua scientist can update her degree of belief in any relevant hypothesis in light of the evidence that comes from those methodological choices—and that updating can be free of the canonically social or ethical values that guided those methodological choices in the first place. Or at least, so a modern Jeffrien is entitled to maintain. So why is climate science different? It is different because of the size, complexity, socially cooperative origin, and historical path dependency of climate modes. And it is different because climate experts, in light of the individually limited role that they play in the socially extended activity of building climate knowledge, can only arrive at posterior degrees of belief in ways that are fundamentally mediated by the complex models that they build. And they are incapable of sorting out the ways in which past methodological choices are influencing, through their entrenchment in the very models that mediate their inferences, the ways in which they could possibly arrive at those posterior degrees of belief. Their judgments about climate uncertainties, therefore, whether they come from "objective" ensemble methods, or from their subjective judgments, cannot be free from the social values that guide methodological choices everywhere in the sciences.

13.9 Values in the Nooks and Crannies

At this point in the discussion, it might be natural for a reader to ask for a specific example of a social, political, or ethical value that has influenced a methodological choice in the history of climate modeling. It is easy to give a couple of potted examples. In previous work, I have focused on the extent to which climate models have been optimized, over their history, to particular purposes, and to particular metrics of success.[21] I gave the example that, in the past, modelers had perhaps focused on the metric of successfully reproducing known data about global mean surface temperature, rather than other possible metrics. I speculated that they might have done so because of a social and political climate in which the concern was about "global warming," a phrase that is now being supplanted by the phrase "anthropogenic climate change."

But I now think it was a mistake to focus on particular historical claims about specific motives and choices. I want to focus instead on the fact that climate modeling involves literally thousands of unforced methodological choices.[22] Many crucial processes are poorly understood, many compromises in the name of computational exigency need to be made, and so forth. All one needs to see is that, as in the case of the biopsy stain, no unforced methodological choice can be defended in a value vacuum. If one asks, "Why parameterize this process rather than try to resolve it on the grid?" or "Why use this method for modeling cloud formation?" it will rarely be the case that the answer can be "because that choice is objectively better than the alternative." Rather, most choices will be better in some respects and worse in other respects than their alternatives, and the preference for the one over the other will reflect the judgment that this or that respect is more important. Some choices will invariably increase the probability of finding a certain degree of climate variation, while its alternative will do the opposite—and so the choice that is made can be seen as reflecting a balance of inductive risks.

All we need to argue here is that many of the choices made by climate modelers had to have been unforced in the absence of a relevant set of values—that in retrospect, such choices could only be defended against *some set of predictive preferences* and *some balance of inductive risks.* In other words, any rational reconstruction of the history of climate science would have to make mention of predictive preferences and inductive risks at pain of making most of these choices seem arbitrary. But what I want to be perfectly clear about here (in a way that I think I have not been in earlier work) is that I do not mean to attribute to the relevant actors these psychological motives, nor any particular specifiable or recoverable set of interests.[23] I am not in the business of making historical, sociological, or psychological claims. I have no idea why individual agents made the choices that they made—and indeed it is part of my argument that these facts are mostly hidden from view. In fact, for many of the same reasons that these methodological choices are immune from the BRDC, they are also relatively opaque to us from a historical, philosophical and sociological point of view. They are buried in the historical past under the complexity, epistemic distributiveness, and generative entrenchment of climate models.

Some readers may find that this makes my claim about the value-ladenness of climate models insufficiently concrete to have any genuine bite. One might ask: "Where are the actual values?" Some readers, in other words, might be craving some details about how agents have been specifically motivated by genuine concrete ethical or political considerations. They might be tempted to think that I have too abstractly identified the role of values here to be helpful. But this is to miss the dialectical structure of my point. The very features that make the BRDC implausible make this demand unsatisfiable. No help of the sort that "finds the hidden values" can be forthcoming on my account. The social, political, and ethical values that find their way into climate models cannot be recovered in bite-sized pieces.

Recall that we began this whole discussion with a desire to separate the epistemic from the normative. But we have now learned that, with respect to science that relies on models that are sufficiently complex, epistemically distributed, and generatively entrenched, it becomes increasingly difficult to tell a story that maintains that kind of distinction. And without being able to provide a history that respects that distinction, there is no way to isolate the values that have been involved in the history of climate science.

One consequence of the blurred distinction between the epistemic and the normative in our case is that the usual remarks philosophers often make about the value-ladenness of science do not apply here. Those who make the claim that science is value laden often follow up with the advice that scientists ought to be more self-conscious in their value choices and that they ought to ensure that their values reflect those of the people they serve. Or they suggest implementing some system for soliciting public opinions or determining public values and making that the basis for these determinations. But on the picture I am painting, neither of these options is really possible. The bits of value-ladenness lie in all the nooks and crannies; they might very well have been opaque to the actors who put them there, and they are certainly opaque to those who stand at the end of the long, distributed, and path-dependent process of model construction. In the case of the biopsy stains I can say "consumer protection is always more important than corporate profits! Even in the absence of epistemologically forcing considerations, the toxicologist should choose the stain on the left!" But in the climate case, the situation is quite different. We can of course ask for a climate science that does not reflect systematic biases, unlike one cynically paid for by the oil industry. But this demand for a science that reflects the "right values" cannot go "all the way down" into all those nooks and crannies. In those relevant respects, it becomes terribly hard to ask for a climate science that reflects "better" values.[24]

13.10 Conclusion

So what could Climate Science—its practitioners, its public consumers, and the policymakers who rely on it—do? One very sensible response to a state of affairs in which there is no principled and value-neutral way to assign a precise probability distribution to climate outcomes is to refrain from giving one—certainly from giving one that is derived in a simplistic way from the distribution of modeling results that come from the set of models on the market. This is what Wendy Parker has urged in response to some of my earlier work. Perhaps, she argues, what we have learned is that a probability density function over all the possible outcomes is too detailed and precise a depiction. Perhaps what climate scientists ought to deliver to the public, and to policymakers, is something coarser.

> In practice, coarser depictions of uncertainty are what we actually get from expert groups like the Intergovernmental Panel on Climate Change (IPCC). Even for GMST, IPCC uncertainty estimates reached on the basis of expert judgment assign only a portion of the probability mass and, moreover, in some cases assign it to a predictive range that extends significantly beyond that delineated by predictions from today's state-of-the-art models and/or by more formal probabilistic methods.[25]

She gives the following example from the IPCC report (see Fig. 13.1):

In Fig. 13.1, the gray bars give the ranges of values (for each emissions scenario), inside which the IPCC experts deemed that there was at least a 66% chance that the actual value of global mean surface temperature would fall. As we can see, this range is significantly larger than any of the formal methods for calculating probability would give us. This reflects their judgment, as Parker puts it, that "today's state-of-the-art models do not thoroughly or systematically sample existing uncertainty about how to adequately represent the climate system. More specifically, it reflects the judgment that these models are more likely to underestimate rather than overestimate changes in GMST and that, while they may well be off

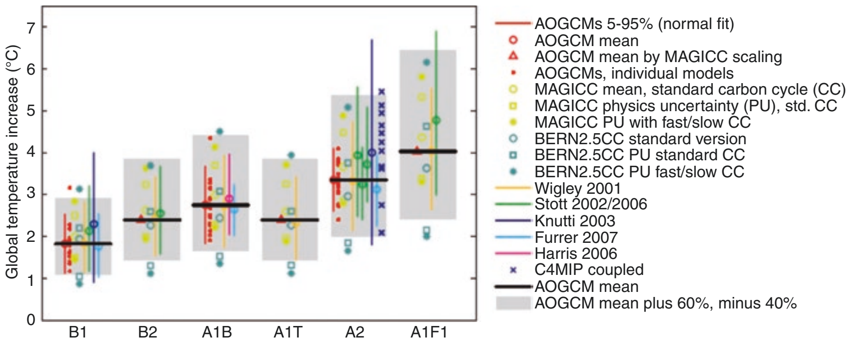

Fig. 13.1 Projections and uncertainties for global mean temperature increase in 2090–2099 (rel. to 1980–1999 avg.) for the six SRES marker scenarios (Source: IPCC AR4 WG1 2007)

by something like 50% in their projections for 2050, there is not a high probability that they are off by something like 500%" (Ibid., Parker).

I agree with Parker (and the IPCC!) that this is an excellent strategy. What I do not agree with is that it is a value-free strategy. Notice what Parker correctly notes is the justification for this report: that there is a (relatively) "high probability" that the models may be off by something like 50%, but a (relatively) "low probability" that they are off by something like 500%. I agree that these are the correct sorts of judgments to be making, but these are classic Rudnerian judgments—they reflect a balance of inductive risks. Deciding to omit some chunk of possibility space from covering a range of values because there is a sufficiently low (second order) probability that it belongs there is exactly the kind of judgment that Rudner was talking about—only elevated to the level of second order probabilities.[26] It can only be made with a combined judgment of the probability that the real value lies in that space, and of the moral, social, or political cost of being wrong. But this is exactly what the IPCC is doing when they leave off those tails on the grounds that they have a "low probability." It is a logical possibility, after all, that one might make the judgment that, even though the probability that the models are off by 500% is extremely small, that the seriousness ("in the typically ethical sense"—Rudner) of neglecting that possibility and having it actually be the case would outweigh that very small probability.

I would like to emphasize that I am not criticizing the IPCC here. I agree with Parker that this is the correct thing to do in light of the present situation with climate models, and in light of the situation that is likely to exist under any practicable state of affairs. But I insist that it is not value free. It is only a slight twist on the classic Rudnerian decision to decide that the (second order) probability that less than 66% of the (first order) probability lies in the gray bar is sufficiently low to be safely ignored. To decide that second order probabilities are sufficiently low to be ignored is to choose a balance of inductive risks. It reflects a judgment that the risk of sticking their necks out further and being wrong is equally balanced by the risk of not sticking it out far enough. As long as there is no principled PDF to be offered, some amount of neck sticking is required. And how far out one should stick one's neck is a classic balance of inductive risks.

If one is uncomfortable with second order probabilities, there are other ways to interpret what the IPCC is doing. But none of them change the conclusion. It is clear that the IPCC cannot be perfectly confident that exactly 66% of the probability mass lies precisely in the grey bars. If they were perfectly confident of this, than they would have a principled precise first order probability—and this is what we have argued, above, they cannot have. But this means that could have made a more precise estimate with less confidence, or a less precise estimate with more confidence. And choosing the right balance of precision and confidence here is a value judgment.

Of course, when values enter into the picture in *this* kind of way—when the experts at the IPCC make a determination about what kinds of minimum probabilities to report—the points I made earlier about the inscrutability of the values no longer apply. At this point in the process, one might even say that the values are being applied fairly self-consciously. And so vis-à-vis this part of the process, I think the ordinary lessons about the role of values in science (that scientists ought to be more self-conscious in their value choices, and that they ought to ensure that their values reflect those of the people they serve, etc.) do apply. And I have no reason to doubt that the IPCC does a reasonably good job of this. But we should not let this conceal the fact that the fundamental science on which IPCC bases its judgments (all the color-coded action inside the gray bars) conceals, in all the ways I described in the last section, an opaque, inscrutable tapestry of values.

Acknowledgements Thanks to Kevin Elliot, Rebecca Kukla, Elisabeth Lloyd, Wendy Parker, Isabelle Peschard, Bas van Fraassen, and Jessica Williams for helpful comments, criticisms, and suggestions as I worked on this manuscript. And thanks to all the participants at conferences and colloquia where I have presented earlier versions of this work, including at the Technical University Eindhoven, San Francisco State University, Georgetown University, the 2010 AGU meeting in San Francisco, and the University of South Florida, and at the 2011 Eastern APA Author Meets Critics session. Too many helpful suggestions, comments, and criticisms have been made to keep track of. Thanks to Justin Biddle and Johannes Lenhard for working with me on previous projects (see the bibliography) that have contributed immeasurably to my understanding of these topics.

Notes

1. Of course one might have worries about whether elected representatives generally represent the values of their constituents but that is the subject of a different discussion.
2. I variously use the expressions "social values," "ethical values," or "social and ethical values" which should not be read as flagging important philosophical differences.
3. See also (Frank 1954; Neurath 1913; Douglas 2000; Howard 2006; Longino 1990, 1996, 2002; Kourany 2003a, b; Solomon 2001; Wilholt 2009; Elliott 2011a, b).
4. Many discussions of UQ in climate science will also identify data uncertainty. In evaluating a particular climate model, including both its structure and parameters, we compare the model's output to real data. Climate modelers, for example, often compare the outputs of their models to records of past climate. These records can come from actual meteorological observations or from proxy data—inferences about past climate drawn from such sources as tree rings and ice core samples. Both of these sources of data, however, are prone to error, and so we are uncertain about the precise nature of the past climate. This, in turn, has consequences for our knowledge of the future climate. While data uncertainty is a significant source of uncertainty in climate modeling, I do not discuss this source of uncertainty here. For the purposes of this discussion, I make the crude assumption that the data against which climate models are evaluated are known with certainty. Notice, in any case, that data uncertainty is part of parameter uncertainty and structural uncertainty, since it acts by affecting our ability to judge the accuracy of our parameters and our model structures.
5. A parameter for a model is an input that is fixed for all time, while a variable takes a value that varies with time. A variable for a model is thus both an input for the model (the value the variable takes at some initial time) and an output (the value the variable takes at all subsequent times). A parameter is simply an input.
6. Some might argue that if we look at how the models perform on past data (for, say, mean global surface temperature), they often are distributed around the observations. But, first, these distributions do not display anything like random characteristics (i.e., normal distribution). And, second, this feature of one variable for past data (the data for which the models have been tuned) is a poor indicator that it might obtain for all variables and for future data.

7. Masson and Knutti (2011) discuss this phenomenon and its effects on multimodel sampling, in detail.
8. Shewhart (1939).
9. Which, inter alia, did much to bring the issue of "inductive risk" back into focus for contemporary philosophy of science and epistemology.
10. Whether they would do so in fact is not what is at issue here. Surely that would depend on features of their psychology and of the institutional structures they inhabit, about which we would have to have a great deal more empirical evidence before we could decide. What is at stake here is whether their social and ethical values would *necessarily* play a role in properly conducted science.
11. See, for example, Goldstein and Rougier (2006).
12. For an account of the controversies around early coupling, see Shackley et al. (1999); for a brief history of modeling advances, see Weart (2010).
13. As, for example, in the earth system modeling framework. See, e.g., Dickenson et al. (2002).
14. Because data are being continuously exchanged one can accurately describe the models as parallel rather than serial in the sense discussed in Winsberg (2006).
15. "Balance of approximations" is a term introduced by Lambert and Boer (2001) to indicate that climate models sometimes succeed precisely because the errors introduced by two different approximations cancel each other out.
16. There has been a move, in recent years, to eliminate "legacy code" from climate models. Even though this may have been achieved in some models (this claim is sometimes made about CM2), it is worth noting that there is a large difference between coding a model from scratch and building it from scratch, that is, devising and sanctioning from scratch all of the elements of a model.
17. See Rougier and Crucifix, this volume.
18. I do not have the space to talk about what "manageably small" might mean here. But see our discussion of "catch and toss" group authorship in the work mentioned in the next note.
19. One might reasonably wonder whether, in principle, a group could be an epistemic agent. In fact, this is the subject of a forthcoming paper by Bryce Huebner, Rebecca Kukla, and me. I would argue here, however, and hope that we will argue in more detail in that paper, that the analytic impenetrability of the models made by the groups involved here is an obstacle to these groups being agents with subjective degrees of belief.

20. One can think of the contribution to this volume by Rougier and Crucifix as a recognition of, and attempt to address, this problem: that complex climate models are too complex to help climate scientists develop subjective degrees of belief.
21. See especially Biddle and Winsberg (2009), and also Winsberg (2010, ch. 6).
22. Here, my point is very well supported by Elisabeth Lloyd's contribution to this volume. Her chapter chronicles in detail a very nice example of the kind of unforced methodological choice I am talking about: the choice of how to calibrate the relevant satellite data. The way Lloyd tells the story, the process involved a whole host of data-processing decisions and choices. I am simply adding to Lloyd's narrative the observation that each of the decisions and choices she chronicles can be understood as being underwritten by balances of inductive risk and prediction preferences.
23. One might complain that if the decisions do not reflect the explicit psychological motives or interests of the scientist, then they do not have a *systematic* effect on the content of science, and are hence no different than the uncontroversial examples of social values I mentioned in the introduction (such as attaching greater value to AIDS research than to algebraic quantum field theory). But though the effect of the values in the climate case might not have a *systematic* effect on the content of science, it is nonetheless an effect *internal* to science in a way that those other examples are not.
24. Again, Elisabeth Lloyd's contribution to this volume illustrates this point.
25. This comes from Parker's remarks at the 2011 meeting of the Eastern division of the American Philosophical Association during an author meets critic session for my (2010).
26. The probability that less than 66% of the probability mass lies inside the gray bar is a second order probability because it talks about the probability of a probability.

Works Cited

Allen, Myles. What Can Be Said About Future Climate? ClimatePrediction.net, June. Available at http://www.climateprediction.net/science/pubs/allen_Harvard2008.ppt. Accessed 3 July 2008.

Biddle, Justin, and Eric Winsberg. 2009. Value Judgments and the Estimation of Uncertainty in Climate Modeling. In *New Waves in the Philosophy of Science*, ed. P.D. Magnus and Jacob Busch. New York: Palgrave Macmillan.

Churchman, C. West. 1949. *Theory of Experimental Inference*. New York: Macmillan.

———. 1953. Science and Decision Making. *Philosophy of Science* 23 (3): 247–249.

Clark, Andy. 1987. The Kluge in the Machine. *Mind and Language* 2 (4): 277–300.

Dickenson, Robert E., Stephen E. Zebiak, Jeffery L. Anderson, et al. 2002. How Can We Advance Our Weather and Climate Models as a Community? *Bulletin of the American Meteorological Society* 83 (3): 431–434.

Douglas, Heather. 2000. Inductive Risk and Values in Science. *Philosophy of Science* 67 (4): 559–579.

Dunne, John. 2006. Towards Earth System Modelling: Bringing GFDL to Life. Paper Presented at the ACCESS 2006 BMRC Workshop. Available at http://www.cawcr.gov.au/bmrc/basic/wksp18/papers/Dunne_ESM.pdf. Accessed 11 Jan 2011.

Elliot, Kevin. 2011a. Direct and Indirect Roles for Values in Science. *Philosophy of Science* 78: 303–324.

———. 2011b. *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Research*. New York: Oxford University Press.

Frank, Philipp G. 1954. The Variety of Reasons for the Acceptance of Scientific Theories. In *The Validation of Scientific Theories*, ed. Phillipp Frank, 3–17. Boston: Beacon Press.

Gleckler, Peter J., Karl E. Taylor, and Charles Doutriaux. 2008. Performance Metrics for Climate Models. *Journal of Geophysical Research* 113: D06104. https://doi.org/10.1029/ 2007JD008972.

Goldstein, Matthew, and Jonathan C. Rougier. 2006. Bayes Linear Calibrated Prediction for Complex Systems. *Journal of the American Statistical Association* 101 (475): 1132–1143.

Howard, Don A. 2006. Lost Wanderers in the Forest of Knowledge: Some Thoughts on the Discovery-Justification Distinction. In *Revisiting Discovery and Justification: Historical and Philosophical Perspectives on the Context Distinction*, ed. Jutta Schickore and Friedrich Steinle, 3–22. New York: Springer.

IPCC (Intergovernmental Panel on Climate Change). 2001. *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third*

Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press.

———. 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*. Cambridge: Cambridge University Press.

Jeffrey, Richard C. 1956. Valuation and Acceptance of Scientific Hypotheses. *Philosophy of Science* 23: 237–246.

Kourany, Janet. 2003a. A Philosophy of Science for the Twenty-First Century. *Philosophy of Science* 70 (1): 1–14.

———. 2003b. Reply to Giere. *Philosophy of Science* 70 (1): 22–26.

Küppers, Günter, and Johannes Lenhard. 2006. Simulation and a Revolution in Modeling Style: From Hierarchical to Network-like Integration. In *Simulation: Pragmatic Construction of Reality, Sociology of the Sciences*, ed. Johannes Lenhard, Günter Küppers, and Terry Shinn, 89–106. Dordrecht: Springer.

Lambert, Steven, and George Boer. 2001. CMIP1 Evaluation and Intercomparison of Coupled Climate Models. *Climate Dynamics* 17 (2–3): 83–106.

Lenhard, Johannes, and Eric Winsberg. 2010. Holism, Entrenchment, and the Future of Climate Model Pluralism. *Studies in History and Philosophy of Modern Physics* 41: 253–262.

Lloyd, Elisabeth. 2012. The Role of 'Complex' Empiricism in the Debates About Satellite Data and Climate Models. *Studies in History and Philosophy of Science* 43: 390–401.

———. 2015. *Model Robustness* as a Confirmatory Virtue: The Case of Climate Science. *Studies in History and Philosophy of Science* 49: 58–68. https://doi.org/10.1016/j.shpsa.2014.12.002.

Longino, Helen. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press.

———. 1996. Cognitive and Non-cognitive Values in Science: Rethinking the Dichotomy. In *Feminism, Science, and the Philosophy of Science*, ed. Lynn Hankinson Nelson and Jack Nelson, 39–58. Dordrecht: Kluwer.

———. 2002. *The Fate of Knowledge*. Princeton: Princeton University Press.

Masson, David, and Reto Knutti. 2011. Climate Model Genealogy. *Geophysical Research Letters* 38 (8): L08703. https://doi.org/10.1029/2011GL046864.

Neurath, Otto. 1913. Die Verirrten des Cartesius und das Auxiliarmotiv: Zur Psychologie des Entschlusses. In *Jahrbuch der Philosophischen Gesellschaft an der Universität Wien*, 45–59. Leipzig: Johann Ambrosius Barth.

Palmer, Tim N. 2001. A Nonlinear Dynamical Perspective on Model Error: A Proposal for Non-local Stochastic–Dynamic Parameterization in Weather and Climate Prediction Models. *Quarterly Journal of the Royal Meteorological Society* 127 (572): 279–304.

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

Rudner, Richard. 1953. The Scientist *Qua* Scientist Makes Value Judgments. *Philosophy of Science* 20 (3): 1–6.

Shackley, Simon, James Risbey, Peter Stone, and Brian Wynne. 1999. Adjusting to Policy Expectations in Climate Change Science: An Interdisciplinary Study of Flux Adjustments in Coupled Atmosphere Ocean General Circulation Models. *Climatic Change* 43 (3): 413–454.

Shewhart, Walter A. 1939. *Statistical Method from the Viewpoint of Quality Control*. New York: Dover.

Smith, Leonard A. 2002. What Might We Learn from Climate Forecasts? *Proceedings of the National Academy of Sciences USA* 4 (99): 2487–2492.

Solomon, Miriam. 2001. *Social Empiricism*. Cambridge, MA: MIT Press.

Tebaldi, Claudia, and Reto Knutti. 2007. The Use of the Multimodel Ensemble in Probabilistic Climate Projections. *Philosophical Transactions of the Royal Society A* 365 (1857): 2053–2075.

Weart, Spencer. 2010. The Development of General Circulation Models of Climate. *Studies in History and Philosophy of Modern Physics* 41 (3): 208–217.

Wilholt, Torsten. 2009. Bias and Values in Scientific Research. *Studies in History and Philosophy of Science* 40: 92–101.

Wimsatt, William. 2007. *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, MA: Harvard University Press.

Winsberg, Eric. 2006. Handshaking Your Way to the Top: Simulation at the Nanoscale. *Philosophy of Science* 73 (5): 582–594.

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