Ensembles as Evidence, not Experts: Why Probabilities are Valuable in the Interpretation of Climate Models

Corey Dethier

[Draft; comments extremely welcome. Please contact at corey.dethier[at]gmail.com before citing or quoting.]

Abstract

Climate scientists frequently interpret climate models as providing probabilistic information, a practice that has come under substantial criticism from philosophers of science. The present paper defends this interpretation. I show that the probabilistic information provided by "ensembles" of climate models is invaluable to climate science; it provides important information about how to distribute confidence over various alternatives. Importantly, this information is best understood and treated as *evidence* (rather than as some kind of deferral-worthy expert function). From this perspective, it becomes clear that the criticisms raised by philosophers only motivate a moderate position according to which some but not all uses of the probabilities generated by "ensemble-based methods" are appropriate.

Contents

0 Introduction

Climate scientists frequently employ groups or "ensembles" of climate models when evaluating hypotheses about the past, present, and future climate. In many cases, they interpret the results given by these ensembles as providing probabilistic information—that is, they treat the variation between the different members of the ensemble as providing evidence about the probability of various alternative scenarios. Philosophers have written extensively about both these "ensemble-based methods" and the probabilities that they generate, most of it quite critical: though the details of the arguments differ, the broad consensus within philosophy seems to be that extant ensembles are not properly "independent" in the way that they would need to be to (e.g.) apply statistics to them, and thus that the probabilistic results of such applications are not worthwhile.[1]

It's important to distinguish between two possible views motivated by these criticisms. More moderately, one might hold that climate scientists shouldn't uncritically accept or adopt the precise probabilities generated by ensemble-based methods. In the terminology of epistemology, we shouldn't take the resulting probabilities to be *expert functions*; they aren't unproblematic guides to the true probabilities or the probabilities that we should adopt. A more extreme view says that climate scientists shouldn't use ensemble-based methods *at all*. Some of the prior literature—see, e.g., Wendy Parker's work—is explicit

[1]For examples, see Betz (2007, 2015), Carrier and Lenhard (2019), Jebeile and Barberousse (forthcoming), Katzav (2014), Katzav et al. (2021), Parker (2010a,b, 2013), Parker and Risbey (2015), Schmidt and Sherwood (2015), and Winsberg (2018). There is to my knowledge only one paper that explicitly defends the practice (Dethier 2022).

in adopting the more moderate position. By contrast, others—such as Katzav et al. (2021)—seem at times to advocate for the extreme view.[2]

Absent from much of the prior literature is a discussion of why the probabilities generated by ensembles might be valuable; the general assumption seems to be that we want probabilities for decision-theoretic purposes. As I'll show in this paper, however, there are other benefits to the probabilistic treatment of climate models besides the (presumptive) decision-theoretic ones discussed in the literature. In particular, the probabilities generated by ensembles of climate models provide valuable *evidence* about how we should distribute our confidence over different hypotheses. I argue that we don't have any good reason to throw away this evidence—or at least, we have *no more* reason to throw away this evidence than we have reason to throw away any information provided by the models. The upshot is an argument in favor of a conditional version of the moderate view: if we have sufficient reason to treat anything the models tell us as evidence—I think we do, but I won't be arguing for that here—then we have sufficient reason to treat the probability functions generated by ensemble-based methods as evidence too.

The first two sections lay my positive argument: the variation between members of an ensemble provides us with information about how to distribute our confidence that a single model does not. As I stress, the benefit here isn't purely theoretical. On the contrary, climate scientists sometimes make use of this information to concretely improve their methods. Section three briefly states my positive position, while the final section addresses various objections that have been raised in the literature. In particular, I argue that many of these arguments are ineffective against the moderate position advanced in this paper; they give us good reason not to treat the probabilities generated by ensembles as experts, but not good reason to throw them out.

One final note. In what follows, I focus on a paradigm case of ensemble-based methods, namely the application of statistics to the set of results generated by simulations run on an ensemble of climate models. As I'll stress, however, what my arguments really motivate is not the use of this particular method but rather merely that we use *some* method that takes the variation between model results into account. There are worthwhile debates to be had about which ensemble-based methods (in this broad sense) climate scientists should use, and I want to leave the door open for other approaches so long as they take account of inter-model variation in some way.

[2]On my reading, Katzav et al. are really interested in a more moderate position—i.e., they're advocating against the use of precise probability functions to represent future uncertainty in the context of decision-making. What they say, however, is that precise probability functions "should not be used in the climate context" (Katzav et al. 2021).

1 Climate modeling and imprecision

1.1 Using climate models to evaluate hypotheses

Here I briefly outline how climate models are used in evaluating hypotheses about the (future) climate. To make the discussion more concrete, consider equilibrium climate sensitivity (ECS), the °C change in temperature that will be observed given a doubling of the atmospheric CO_2 concentration.

Were we Bayesian rational agents in an ideal situation, our estimate for ECS would consist in a precise probability distribution over possible values of ECS, and this distribution would be generated by conditionalizing our prior expectations on the total evidence. In practice, of course, this isn't feasible.[3] We're rarely if ever in a position to directly employ our total evidence in evaluating hypotheses. It's not as though we have access to a complete description of all of the evidence collected up to this point, let alone an understanding of the probabilistic relationships between that description and various hypotheses. Instead of calculating the probability of a hypothesis like ECS = 2.5°C directly on the total evidence, therefore, scientists build theories and models that systematize the evidence as well as possible. These theories and models then tell us what we should believe about the future.

Simulations run using global climate models are one of the methods by which climate scientists estimate ECS.[4] We can think of global climate models as consisting of a number of gridded shells, with each shell representing a layer of the atmosphere and each grid box a location in that layer. Each grid box is assigned a number of climate variables, representing (e.g.) the average temperature and precipitation in that region over the course of a time-step (say, a month). The relationships between the variables found in various grid boxes are given by a series of equations that determine how a change in the climate variables of one box affects other variables in that box as well as the variables in its neighbors. At the simplest level, quantities like heat will simply defuse through the system, but of course there are more complicated effects

[3] It's common for (Bayesian) epistemologists to wave away concerns about feasibility by pointing out that the relevant standards are "evaluative" rather than "normative." Fair enough. Science isn't concerned with the reasons that an agent might have in an abstract evaluative sense, however, but instead with the reasons that can be made intersubjectively salient (Longino 1990): you have to be able to demonstrate to other people that a hypothesis is warranted. Feasibility questions—e.g., can your evidence be communicated?—are thus relevant to philosophy of science in a way that they (arguably) aren't to (ideal) epistemology.

[4] For a comparison of the various sources of estimates for ECS, see IPCC (2013, 922–923, 1110, box 12.2).

as well.[5]

To use a global climate model in estimating a quantity like ECS, climate scientists run computer simulations in which the model is "forced" to take on a new state by an exogenous change; in the case of ECS, for instance, one standard procedure is to rapidly double the amount of CO_2 in the (simulated) atmosphere. Comparing the end-state of the simulation to the initial state yields a point-value quantity for the change in average temperature *in the model*. So suppose that the in-model change in average temperature, represented by $\Delta\bar{T}$, is 2.5°C. In what follows, I'll speak of a model "saying" or "reporting" that $\Delta\bar{T} = 2.5$°C. The idea here is that the "model report" is akin to an "instrumental reading": the quantity that our computer simulation spits out is *like* the quantity that we read off a thermometer. It's a data point to be recorded and interpreted and which our hypotheses about the climate will be expected to account for (compare Parker 2020a).[6]

In what follows, I'll often speak about the interpretation of model reports in Bayesian language (though nothing hangs on this particular choice of framework). In this framework, "interpreting" model reports means conditionalizing on them in accordance with Bayes' rule. In our ECS example, that means that the probability that we (should) assign to a hypothesis like ECS = 2.5°C on the basis of the model report that $\Delta\bar{T} = 2.5$°C is given by

$$\begin{aligned} Pr^*(\text{ECS} = 2.5) &= Pr(\text{ECS} = 2.5 \mid m : \Delta\bar{T} = 2.5) \\ &= \frac{Pr(\text{ECS} = 2.5)Pr(m : \Delta\bar{T} = 2.5 \mid \text{ECS} = 2.5)}{Pr(m : \Delta\bar{T} = 2.5)} \end{aligned}$$

where "$m : \Delta\bar{T} = 2.5$" indicates that the model m is reporting that $\Delta\bar{T} = 2.5$°C. The crucial point is that what we take from the model report depends on how well (we think) the model is tracking the truth, or, in the Bayesian framework, on the likelihood ratio.

[5]The picture I've presented here is simplified in a number of ways. For a deeper discussion, see a climate modeling primer such as Gettelman and Rood (2016) and McGuffie and Henderson-Sellers (2014). For a philosophical introduction, see Winsberg (2018, 27–54).

[6]I suspect that this assumption is a key point of divergence between my own approach and that of those who are more skeptical of ensemble-generated probabilities. On the picture offered by Thompson and Smith (2019), for instance, a model needs to clear some sort of bar for accuracy or reliability before we can treat it as providing information about the world, whereas I'm simply building the inaccuracy or unreliablility of the model into the likelihood ratio. I think Thompson and Smith are probably right about the practice—climate scientists both should and in fact do only take models into consideration when they meet some minimal standard of accuracy—but this doesn't affect my central contention. Roughly, if the models don't clear the bar, we shouldn't take anything they say as evidence; if they do clear the bar, we should treat the inter-model variation as evidence.

1.2 The problem of imprecise evidence

It is uncontroversial that climate models are not perfect: they misrepresent or idealize some real climate processes, omit or paramaterize others, and rely on assumptions that are risky or arbitrary in the sense that we don't know whether they're true.[7] It is also uncontroversial that climate models are (relatively) "opaque" in the sense that it is hard to tell how any one idealization affects the accuracy of the model with respect to a variable of interest.[8]

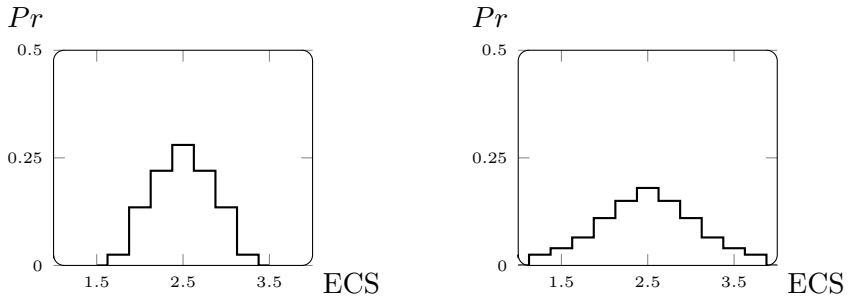
The upshot of these two facts is that climate scientists are rarely (if ever) in a position to know exactly how to interpret a given model report—in our Bayesian framework, they don't know the likelihood of a given model report on different hypotheses. It's helpful to be slightly more concrete. So suppose that our model generates a report of $\Delta\bar{T} = 2.5°\text{C}$. For the sake of simplicity, suppose further that we know that the likelihood of observing different values for $\Delta\bar{T}$ given some hypothesis h is given by a normal distribution centered on the truth. Essentially: we know that the hypothesis on which our model report has the highest likelihood is ECS $= 2.5°\text{C}$, and that the likelihood of the model report falls off as we move to hypotheses that assign more distant values to ECS. The *sole* question in this simplified example is how quickly the likelihood falls off. If the normal distribution has a standard deviation of .25, our confidence in different values of ECS will look like the graph pictured in figure 1a. And if the normal distribution has a standard deviation of .5, our confidence in different values of ECS will look like the graph pictured in figure 1b. Insofar as we're uncertain about which assumption about the model's accuracy we ought to adopt, we'll be equally uncertain about which distribution we should prefer.

Situations like this one are sometimes described by epistemologists as involving "imprecise evidence" (see, e.g., Carr 2019). That terminology can be confusing, however. It's not the case that the datum that we're conditioning on is itself imprecise; on the contrary, the model's reported value for $\Delta\bar{T}$ can be calculated with as much precision as we like. Instead, the problem is that we're not in a position to justify a precise interpretation of the model report—we can't pick out a single probability function as *the* assignment of probabilities that the report supports.

Here's another way at getting at this contrast. Evidence can warrant more

[7] See IPCC (2013, chapter 9). The point is also widely acknowledged by philosophers: see Carrier and Lenhard (2019), Jebeile and Barberousse (forthcoming), and Parker (2010a).

[8] The terminology of "opacity" is owed to Humphreys (2004); for discussions of opacity in the context of climate models, see Carrier and Lenhard (2019), Lenhard and Winsberg (2010), and Parker and Winsberg (2018).



(a) Expected model accuracy is given by $Pr(m : \Delta\bar{T} = x \mid \text{ECS} = 2.5) \sim \mathcal{N}(2.5, .25)$.

(b) Expected model accuracy is given by $Pr(m : \Delta\bar{T} = x \mid \text{ECS} = 2.5) \sim \mathcal{N}(2.5, .5)$.

Figure 1: Posterior probability distributions for values of ECS at the $1/4^{\text{th}}$ of a °C level induced by different views about likelihoods. Priors assumed to be identical and uniform.

or less precise conclusions in at least two different senses. On the one hand, the evidence can warrant more or less precise conclusions in the sense of ruling out possible values for a quantity. To illustrate, contrast learning the proposition that [[ECS falls between 0.5 and 4.5°C]] with learning the proposition that [[ECS falls between 1.5 and 3.5°C]]. The latter rules out more possible values for ECS and is thus more precise in what we might call a first-order sense.

On the other hand, the evidence can warrant more or less precise conclusions in the sense of ruling out possible *distributions* over values. The most familiar (but not the only) way to understand this higher-order sense of precision is in terms of what are called imprecise probability distributions.[9] So, in a standard Bayesian framework, updating your priors $Pr(\cdot)$ on a piece of evidence E yields a single preferred posterior probability function $Pr^*(\cdot) = Pr(\cdot|E)$. That is: the standard Bayesian framework treats all evidence as maximally precise in that it only allows for a single probability distribution. In the example we saw above, however, our uncertainty about likelihood functions meant that we were uncertain about which of two posterior probability functions to adopt—rather than a single probability function Pr^*, we had a set of them $\{Pr^*_1, Pr^*_2\}$. In this example, E is less than maximally precise in a higher-order sense: it doesn't rule out all but one distribution over the possible values.

[9]For an overview, see Bradley (2019) and Mahtani (2019). The alternative that I have in mind replaces imprecise probability's sets of functions with a modal frame and the accompanying access relations (see Dorst 2019; Dorst et al. forthcoming). The differences between these two approaches shouldn't matter for the present discussion.

To summarize, climate modeling—or at least the project of using climate models to estimate quantities like ECS—faces a problem. Due to the heavily idealized and relatively opaque nature of climate models, climate scientists often don't know the likelihood of a given model report on different hypotheses about quantities of interest. Their uncertainty about likelihoods renders the evidence provided by the model report *imprecise* in the higher-order sense just sketched: the evidence allows for a variety of possible distributions over different values for ECS.

Generally speaking, imprecision in our evidence is undesirable: we prefer to be in situations where the evidence warrants more precise hypotheses rather than those in which it only warrants less precise ones. This general preference holds regardless of what sense of precision is at issue and is particularly acute in climate science. As is widely discussed in the scientific literature, the available evidence places much tighter bounds on the low end for ECS than on the high end. Given that higher values for ECS represent relatively disastrous scenarios, however, practical questions concerning (e.g.) what CO_2 concentrations we should aim to stay beneath are highly sensitive to the probability distribution over various unlikely high-end options (Weitzman 2012). In climate science, therefore, imprecise evidence is not just undesirable in an abstract sense—it presents a genuine practical problem.

In the next section, I'll argue that ensembles of models *help*: they provide evidence that is *more* precise than the evidence provided by a single model.

2 From a model to an ensemble

2.1 Ensemble-based methods: a primer

As we saw above, one way that climate scientists estimate quantities like ECS is by running a simulation on a climate model to generate what I've called a "model report," which are like instrumental readings in the sense that they are evidence that needs be "interpreted." (We modeled this "interpretation" step with Bayesian updating, but we could represent in other ways.) "Ensemble-based methods" proceed along largely the same lines: the same simulation is run on each of the models in the ensemble, generating a set of model reports. The crucial difference is that scientists do not reason from or interpret the individual model reports directly; instead, they reason using the features of the set of model reports as a whole.[10]

[10] As Parker (2010b) notes, there are important differences between different kinds of ensembles—and the practice has grown more complex in recent years with the continued

The standard method for turning the set of reports generated by an ensemble into evidence involves employing statistics. In short, this means assuming that the set of reports behaves *as though* it were drawn from some sort of population according to a given sampling procedure. In the simplest case, for instance, climate scientists might assume that the reports behave as though they were randomly drawn from a population centered on the truth. Or (more realistically), they might assume that each of the members of the ensemble is an equally realistic representation of the true climate and thus that the ensemble behaves like a random sample from a population that contains the truth as one of its members.[11] These assumptions are essentially qualitative ways of fixing what's called a "statistical model," a set of assumptions about the probabilistic relationship between various hypotheses and the observed data (i.e. the model reports). Given a statistical model, the observed reports can be used to generate a probability distribution over various alternatives, and these probability distributions (again, as opposed to the individual model reports) are then what climate scientist employ in making judgments about how much confidence we should assign to various hypotheses.

It's worth being a little bit more concrete here. So consider the the assumption that the ensemble behaves like a random sample from a population that contains the truth as one of its members. Essentially, this assumption amounts to the stipulation that for any temperature x, the probability that $\text{ECS} = x$ is equivalent to the probability that an arbitrary model generates a report that $\Delta\bar{T} = x$. Given this stipulation, the likelihood of observing a given set of model reports on the assumption that $\text{ECS} = x$ is just the probability of drawing a report that $\Delta\bar{T} = x$ from the same distribution that characterizes the (imagined) population. So, for instance, if the underlying population is normally distributed, then the likelihood of observing a sample with mean $\Delta\bar{T}$ of 2.5 and standard deviation of .25 on the assumption that ECS falls between 2.4 and 2.6 is given by:

$$Pr(\mu = 2.5, \sigma = .25 \mid 2.4 < \text{ECS} < 2.6) = \int_{2.4}^{2.6} \frac{1}{.25\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-2.5}{.25}\right)^2} dz$$

We can then use these likelihood assignments as part of either a Bayesian updating procedure or a classical hypothesis test—climate scientists use both approaches, though the classical one is currently much more popular.

development of the Coupled Model Intercomparison Project. I'm going to ignore these differences for present purposes: while they are certainly relevant to the evaluation of a particular ensemble-generated probability distribution, they shouldn't affect my central conclusion.

[11] The contrast between these two options is discussed at length in Annan and Hargreaves (2010, 2011) and Sedláček and Knutti (2013).

There are three points that I want to stress before moving on to a discussion of how the use on ensembles *helps* address the problem identified in the last section. First, what I'll be arguing below is that the proper parallel to draw here is not between the reports generated by the ensemble and the individual model report but rather between the probability distribution generated by the ensemble-based method and the individual model report. That is: the probability distributions need to be "interpreted" in the same way that individual model reports do. In this respect, the ensemble-based method is no different from a method based on a single model. What differs between the two cases is that the probability distribution provides us with information that a single model report doesn't. In particular, there's no analogue of the variance (the second moment of the distribution of ensemble results) in the single-model case.

Second, to reiterate a point from the introduction, the ensemble-based method I've sketched here is simply a paradigm case of the most popular approach, and other approaches are possible. Some climate scientists have experimented with interpreting ensembles by weighting the different members and taking their weighted average (Knutti et al. 2017; Sanderson, Knutti, and Caldwell 2015); alternatively, some philosophers have suggested a pooling approach that employs imprecise probabilities and explicitly accounts for the stakes in interpreting the ensemble (Roussos, Bradley, and Frigg 2021). Which of these approaches is best is an interesting question that I don't want to address here; as we'll see, my contention is solely that there are good reasons for climate scientists to use some ensemble-based method that takes account of the variation between the different model reports. In other words, climate scientists should interpret ensembles in a "probabilistic" manner; there's room for disagreement about how to calculate (and use) the relevant probabilities, when and where to coarse-grain or invoke "imprecise" probabilities, but not as to whether the interpretation should be probabilistic.

Finally, as should already be clear, the move to an ensemble doesn't solve the problem of the last section. Recall: in a Bayesian framework, the problem is that we don't know the likelihood relationship between the observed data and various hypotheses. Exactly the same problem arises here, as illustrated above: the debates about which statistical model we should employ in interpreting ensembles are essentially debates about the proper assumptions to make about the likelihood of observing a particular distribution of model reports. The upshot is that the probability distribution generated by an ensemble-based method counts as "imprecise evidence" in the same sense that a single model report does: both allow for a variety of posterior probability distributions over different values for ECS.

2.2 Ensembles and precise evidence

Nevertheless, moving to an ensemble *helps*. The easy way to illustrate this point is by considering the simplified example of the last section. There, we stipulated that the likelihood function—that is, the probability of observing a model report of $\Delta\bar{T} = x$ given a hypothesis ECS $= y$—was given by a normal distribution centered on the truth. Unfortunately, even given this strong assumption, a single model report just doesn't provide us with any information to narrow down the class of possible distributions. In other words, even when we know that the distribution is normal, a single model doesn't tell us how wide or narrow we should expect the normal distribution to be.

An ensemble does. Given the assumption that the likelihood function is given by a normal distribution centered on the truth, an ensemble will allow us to pick out a preferred distribution over possible values for ECS. The crucial difference between the two cases is the inter-model variation, which provides information about the width of the likelihood function: the more variation there is in the sample, the wider we should expect the normal distribution that represents the likelihood function to be. So, just to be concrete, in this case, the likelihood function for an arbitrary model report m_i would be given by a probability density function, meaning that we can calculate the likelihood of any observed report as follows

$$Pr(x < m_i < y \mid \text{ECS} = \mu) = \int_x^y \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-\bar{m}}{\sigma}\right)^2} dz$$

where $\bar{m}$ is the ensemble mean and σ is given by:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(m_i - \bar{m})^2}$$

with n being the number of models in the ensemble.

Of course, we know that extant ensembles aren't actually like random samples from a normal distribution centered on the truth—indeed, they don't approximate such random samples terribly well (Knutti et al. 2010)—and so the assumption just outlined is not in fact warranted. That's why, as stressed above, we cannot say that ensembles solve the problem of imprecise evidence. But they do help: the variation between ensembles allows us to narrow the range of plausible likelihood functions and thus to rule out as implausible some possible posterior probability distributions over values of ECS. Returning to the language of imprecise probabilities, we can think of single-model methods

as delivering a large set of permissible probability functions and ensemble-based methods as delivering a (strictly) smaller set of such functions. Insofar as we desire less imprecision (all other things being equal), we should prefer the ensemble-based method.

What's going on here is basically that inter-model variation provides what epistemologists term *higher-order evidence*. Both a single model and an ensemble provide us with an estimate for the true value of ECS—the model report on the one hand and the mean of the distribution on the other. What the ensemble provides, in addition, is variation between ensemble members, which serves as higher-order evidence concerning how accurate we should expect this estimate to be; all other things being equal, the more variation, the less we should trust the mean as an estimator. The information provided by this variation is what's gained by shifting from the single point-value report to the distribution generated by the ensemble-based method. Of course, the relevant distribution is itself imperfect—it may be misleading or inaccurate in the same way that the first-order evidence may be—but the mere possibility of these sorts of problems doesn't mean that it's not useful.

Here's a slightly different way of making the point. Suppose there's a set of mutually exclusive propositions $\{P, Q, R, ...\}$ and that we don't know exactly what probability we should assign to each of them. One way to proceed here is to adopt the one that seems most likely to be true and to treat it as true—e.g., to work under the assumption that $Pr(P)$ (say) is equal to 1. Later on, we can qualify our results to address the fact that our work was carried out under this risky assumption. Obviously, this isn't an ideal strategy, but it's essentially the strategy that we employ when using a single model. We're taking our "best guess" at how to represent the world, proceeding as though it's entirely accurate, and then keeping our concerns about reliability in mind when updating on the report that it outputs.

In this analogy, using an ensemble-based method is akin to adopting a more equitable distribution of probabilities over the set. So, for example, if our models are equally divided between P, Q, and R and we weight each model equally, that's equivalent to assigning each proposition a probability of 1/3. Since we don't know what probability distribution our total evidence warrants in this case, the resulting probability distribution may be misleading—just as in the single model case, we shouldn't update on the results without taking our concerns about reliability into account. Importantly, however, the ensemble allows us to build *some* of our concerns about reliability into the method itself: the ensemble-based approach accounts for the possibility of error due to assuming P rather than Q or R. As a consequence, the ensemble and the probability functions that it generates are likely to be better at capturing what

we should believe than the single-model and its point-prediction. Essentially, the latter requires us to assign 100% of our confidence to one option, while the former allows us to adopt confidence distributions that more closely track our true confidence. The end result is information that requires less qualification than that provided by a single model. Adopting a particular distribution over the set is risky, but the risk is less substantial than in the single-model case.

The important takeaway is that ensemble-based methods mitigate the problem of imprecise evidence outlined in the last section in virtue of the fact that there's variation between ensemble members that isn't present in a single model. Or, in plainer English, methods that make use of the differences between models provide us more guidance about how to distribute our confidence than methods that don't.

2.3 The concrete benefits of variation

The advantages of ensemble-based methods are not merely philosophical advantages. Of course, it's widely recognized that ensemble means are generally more accurate than the report generated by any single model. The point is made explicitly by the empirical work on ensembles that is commonly cited in the philosophical literature (see, e.g., Knutti et al. 2010) and is in some sense unsurprising: basically any average of a set of estimates will have a higher expected accuracy than the individual estimates (Roussos 2020, 119–20).[12]

In saying that the advantages of ensembles are not merely philosophical, however, I don't have this kind of increase in accuracy in mind. After all, my claim in this section is that the variation between ensemble members provides valuable information—it's open for an opponent to argue that ensemble means are valuable but that the variation between ensemble members isn't. I'm going to wrap up this section by arguing that that position is wrong: variation between ensemble members provides climate scientists with concrete advantages over and above the advantage of having a point-value estimate with higher expected accuracy. Briefly, the reason why is that the estimates generated by climate models aren't just treated as ends in themselves but are frequently used as parts of longer and more complex strings of reasoning. In these con-

[12]That said, I think the improvement in accuracy gained by averaging is underappreciated. As Annan and Hargreaves (2011) argue, the degree to which ensemble means outperform individual models is not fixed by abstract mathematical considerations and demands explanation (see also Rougier 2016). To me, this looks like a problem for the critic of ensemble-based methods: the *surprising* accuracy of ensemble averages looks like empirical disconfirmation of the view that ensembles are too "opportunistic" to be useful (see §4.2 for further discussion).

texts, the variation between ensemble members is crucial: even adopting the ensemble mean in these longer chains of reasoning introduces an additional source of error that climate scientists avoid through the use of ensemble-based methods—in short, it's like rounding in the middle of a calculation.

Our running example can be used to illustrate this point. As noted above, ECS is estimated in a wide variety of ways. So far, we've focused on direct estimates generated by running a simulation on a climate model or set of models that generates a value for the change in temperature. One of the other ways that climate scientists estimate ECS, however, involves using temperature data to estimate the effect that past increases in CO_2 have had on temperature and then extrapolating from those results.[13]

Speaking roughly, this method of estimating ECS works in the following way. Climate scientists collect substantial data on past changes to temperature and then run complex regressions to determine how much of the past temperature change can be attributed to CO_2 and how much to other factors such as the interval variability of the climate system. To run these regressions, we need a quantified understanding of how different factors affect the climate. So, for example, we consistently observe that while the planet as a whole is warming, the upper atmosphere is actually cooling. To determine how much of the observed warming is caused by CO_2 and how much by other factors, we need to know how these different factors affect the distribution of heat throughout the atmosphere. This information—what's sometimes called the "signature" or "fingerprint" of a particular factor—is usually provided by climate models.

Simplifying and abstracting substantially, the resulting regression equation looks like this:

$$Y = \sum_{i}^{n} \beta_i X_i + v_Y$$

where Y is the observed data; β_i and X_i are the percentage of the increase due to the i^{th} factor and the signature of that factor, respectively; and v_Y is the internal variability of the climate. Standard least squares algorithms are then used to estimate the β terms. The results indicate how much of observed warming a particular factor is responsible for; if the least squares analysis yields a result that $\beta_{GHG} = .95$, for example, that would indicate that greenhouse

[13] Stott et al. (2006) is the oldest paper that I'm aware of to estimate quantities like ECS in this way; many, perhaps most, contemporary papers on the attribution of climate change to humans now include sections in which ECS and other variables are estimated using the methods described below.

gases are responsible for 95% of observed warming.[14] Climate scientists can then use results that the regression spits out for CO_2 to estimate ECS.[15]

The point of this example is that the methodology relies on the accuracy of the "signatures" for the different factors—the X terms—and these are estimated using climate models. Standard regression techniques require the assumption that the signatures are given (that is, perfectly accurate). Since our climate models are not perfectly accurate and thus cannot be expected to deliver perfectly accurate estimates for the X terms, this presents a concrete problem for climate scientists: when using standard regression methods, errors in the estimation of the X terms will lead to errors in the estimation of the β terms and thus errors in the estimation of ECS (Carroll et al. 2006).

To address this problem, climate scientists employ ensemble-based methods. There are a couple of different approaches that they have adopted. The first, developed by Huntingford et al. (2006), replaces the X terms with a probability distribution over possible values for X estimated using an ensemble of climate models. The more recent approach, first outlined by Schurer et al. (2018), runs a standard regression for each estimate of X given by the different models to generate probability distributions for the relevant β terms and then uses a Bayesian updating procedure to generate an ensemble probability distribution for the β terms based on the set of distributions generated by each model. In both cases, the end result is a probability distribution over the β terms that can then be used to estimate ECS. Unsurprisingly, tests against data with known properties indicate that both methods generate results that both more accurate and more reliable than those generated by regressions that employ either a single model or just the ensemble mean (Hannart, Ribes, and Naveau 2014; Schurer et al. 2018).

The key takeaway is the following. To estimate ECS in the manner sketched above, we need *some* representation of the signature of factors like CO_2 and thus some estimate for the X terms. We can either (a) adopt a point-value estimate for each X term (generated either by a single model or, better, by taking the mean of an ensemble of estimates) or (b) adopt the probability distribution generated by an ensemble-based method. Both options are vulnerable to misrepresentation: the first might assign the wrong value to an X term; the second might assign the wrong probability to a possible value for

[14] This description of attribution methodology is really only adequate as anything other than a rough approximation for a brief period in the late 90s are early 2000s; the paradigmatic paper here is probably Allen and Tett (1999). For discussion, see Dethier (2022).

[15] For various reasons, this isn't quite as simple as taking the observed change in temperature, multiplying it by β_{CO_2} and dividing by the observed increase in CO_2, but we can forego the details of this last step for present purposes.

an X term. In this sense, they're analogous. Nevertheless, as stressed above, there's an important sense in which the latter is *less* of a misrepresentation because even if it only loosely approximates the confidence that we should assign to each possible value for X, it approximates that distribution better than the first option. After all, the first option can be thought of as adopting a probability distribution that assigns probability 1 to a particular estimate for X. And in providing this more accurate representation of the actual state of our uncertainty, ensemble-based methods allow us to generate more accurate and reliable estimates of related quantities like ECS.

In short: when estimating some quantity of interest (ECS), climate scientists often find themselves needing to rely on model-generated estimates of some other quantity (the X terms). In these contexts, employing a probability distribution over the other quantity can improve the estimate of the quantity of interest. Since the variation found in ensembles provides higher-order evidence about what distribution to adopt, methods that take account of this variation in generating probability distributions allow us to more accurately and reliably estimate the quantity of interest. The upshot is that the higher-order evidence provided by ensembles is not just valuable in an abstract philosophical sense; its presence concretely improves the science.

3 Ensemble-based methods: a positive view

So far, I've argued that climate science faces a problem owing to imprecise evidence and that ensemble-based methods help mitigate that problem, This section briefly states the positive position that I think the above arguments motivate. The next addresses various objections.

The positive position is the following. Given that the variation between ensembles members is (a) useful for the reasons described in the last section and (b) potentially misleading due to the imperfect nature of extant ensembles, climate scientists should treat the probability distributions generated by ensemble-based methods that take account of inter-model variation as potentially misleading evidence. Or, more precisely, if any aspects of the models are treated as evidence, the probability distributions generated by ensemble-based methods should be too. In particular, when estimating quantities like ECS, climate scientists should make use of ensemble-based methods rather than those that don't take inter-model variation into account. This doesn't mean that they should take the probabilities generated by said methods to be the "true probabilities." Plausibly, for instance, it would be a mistake to plug the precise probability distributions calculated using ensemble-based methods

directly into a decision matrix. Instead, such results should be carefully hedged and presented in a more coarse-grained manner like that actually employed by the IPCC.

The motivation for this view is straightforward. The probabilities generated by ensemble-based methods are alike to the point-value estimates generated by individual models in the sense that both are more precise than is warranted. We're assuming that scientists shouldn't ignore the precise point-estimates generated by individual models; by analogy, they shouldn't ignore the precise probabilities generated by ensemble-based methods. At the same time, given that we know that these outputs are more precise than is warranted, it would be irrational to accept either the precise point-values or the precise probability distributions on the say-so of the model(s). Hence my claim that the outputs of ensemble-based methods should be treated as evidence: like the precise point-values, these precise probabilities provide information that must be interpreted. In a slogan: the probability functions generated by ensemble-based methods are evidence, not experts.

This suggestion is amenable to some of the positions found in the literature. Wendy Parker, for instance, has long defended the view that the precise probabilities generated by ensemble-based methods should only be presented to decision-makers or reported as results under specific and demanding criteria (Parker 2010b; Parker and Risbey 2015).[16] Nothing that I have said above conflicts with this position. Indeed, a view on which the probability functions generated by ensemble-based methods are evidence neatly explains why said functions should only be counted as results fit for public consumption under special conditions—after all, most "raw" evidence is exactly the same. Further, as Parker herself has emphasized (e.g. Parker 2020b), which representational tools are appropriate depends on contextual factors, meaning that it shouldn't be surprising that there are some contexts in which it is appropriate to use the precise probability distributions generated by ensemble-based methods to represent our uncertainty in the range of outcomes and other contexts in which it isn't.

Not everyone shares Parker's moderate view here, however, and number of commentators have taken a harder line towards the use of (precise) probability distributions in climate science and to ensemble-based methods more broadly. Winsberg (2018, 98), for example, describes the application of statistics to ensembles as "conceptually troubled." Stainforth et al. (2007, 2155) deny that the probability functions generated by the application of statistical tools to

[16]To be clear, neither Parker nor I am suggesting that these probabilities should be hidden from the public, simply that they shouldn't be (e.g.) touted as results in press releases.

ensemble results are "meaningful." Betz (2015) and Katzav (2014) argue for "possibilist" interpretations of ensembles according to which climate scientists should simply *ignore* the distribution of models within an ensemble. And, most recently, Katzav et al., after explicitly considering a moderate position like the one defended in this paper, argue that precise probability functions "should not be used in the climate context" (Katzav et al. 2021).

The idealized character of extant ensembles is what provides the explicit motivation for these views. From the present perspective, however, rejecting a probabilistic interpretation due to possibility of misrepresentation is—at minimum—overly hasty: whether or not climate scientists should interpret ensembles probabilistically depends on the balance between the benefits of doing so and the potential risks or costs. As we've seen, there are important and concrete benefits to the probabilistic interpretation of ensembles; insofar as extant discussions fail to account for these benefits, they haven't made a complete case for abandoning the use ensemble-based methods and the probabilities that they generate.

Nevertheless, it may be true that in climate science generally speaking the (potential) costs outweigh the benefits. In the next section, I'll argue that they don't—or at least that none of the objections raised in the literature provide compelling reasons for adopting anything more extreme than a moderate and cautious position that is compatible with what I've argued for here.

4 Objections to ensemble-based methods

4.1 Ensembles misrepresent, part 1

The oldest and most frequently repeated objection to the use of ensemble-based methods is that ensembles don't accurately represent the full spread of possibilities, and (thus) that the probabilities that they generate don't accurately represent the uncertainty that we either do have or should in fact have. In what follows, I'm going to distinguish between two ways of running this objection. First, it can be run in a non-specific way: the problem is that there is *some* respect in which the ensembles / probabilities misrepresent. Second, we can run the objection by pointing towards empirical work that shows specific respects in which extant ensembles misrepresent. I treat the first of these here and the second in what follows.

It is widely recognized that extant ensembles are imperfect. Carrier and Lenhard nicely summarize the various problems:

> First, the models are not independent of each other in the sense

> that they only share physical principles and other trustworthy assumptions but are different otherwise. ... Second, errors are correlated between different models and are not random for this reason. ... Third, the ensemble cannot be expected to represent the entire space of possibility. (Carrier and Lenhard 2019, 3–4)

The upshot: ensembles are unlikely to accurately represent—or even serve as a representative sample from—the set of possible climate systems that we should take into account when reasoning about the future.[17] As such, the probabilities generated by ensemble-based methods are unlikely to accurately represent either our "true" uncertainty or the uncertainty that we should have.[18]

It is a substantial step from the fact of misrepresentation to a normative conclusion regarding whether or not we should use the ensembles and methods in question, however. Indeed, philosophers of science have roundly rejected this inference in its general form: the received wisdom is that *all* scientific representations misrepresent their targets in some ways but that many (if not most) are nevertheless useful and informative (see Teller 2004). As the above discussion illustrates, ensemble-based methods require climate scientists to make idealized assumptions about the nature of the ensemble and its relationship to the space of possible representations of the climate. But it's no good to object to the use of idealizations *here* when we happily accept them in other cases; you can't consistently argue against ensemble-based methods on the grounds that extant ensembles are idealized unless you're also willing to argue against the use of climate models—or indeed, all models—on the same grounds.[19]

To be fair, were we aiming to treat the probabilities in question as experts rather than evidence—were we simply taking the probabilities generated by ensemble-based methods to be the objective chances (say)—the existence of idealizations would be a genuine problem, and much of the literature takes

[17] As Dethier (2022) has argued, the technical aspect of these objections largely misses the mark; the application of statistics doesn't require assumptions of genuine independence, full coverage, or even uncorrelated errors. Nevertheless (as he admits), the conclusion that extant ensembles misrepresent is unaffected.

[18] A different version of this argument has been advanced by Smith and various co-authors (see, e.g., Frigg and Smith forthcoming), in which it's argued on similar grounds that the forecasts generated by models are unlikely to line up with the true frequencies and thus that relying on them will lead to non-ideal decisions. I take it that the arguments rehearsed below apply equally well to this version of the argument, however.

[19] There are a number of other potential objections that share this problem. So, one might worry that the probabilities generated by ensemble-based methods might take on a life of their own, even when carefully hedged. But the same is true of any number that climate scientists might generate; there's no reason to think this problem is more damning for ensemble-based methods than it is for quantitative science generally.

this kind of treatment of ensembles as its target. As I argued above, however, this isn't how we should understand the role of ensembles or ensemble-based methods. Instead, the outputs generated by ensemble-based methods should be treated as (potentially misleading) evidence, and I would argue that this is the way that climate scientists (or at least the IPCC) usually treat the probabilities that are generated by ensemble-based methods. So, for instance, while ensembles are often used to generate precise probabilities, it's relatively rare to see these precise probabilities reported directly to the public or to policy-makers. Instead, climate scientists coarse-grain and qualify these results—and they're often explicit that the reason that they do so is to account for idealizations present in the ensemble-based method (see, e.g., IPCC 2013, 883). In other words: climate scientists take the results of ensemble-based methods to indicate what confidence we should have in various hypotheses, but not to do so definitively or with perfect accuracy. And the mere presence of idealizations doesn't serve to undermine this approach.

4.2 Ensembles misrepresent, part 2

The idealized character of extant ensembles isn't just an abstract philosophy problem, however. Indeed, there's been quite a bit of empirical work aimed at evaluating how accurately ensembles represent those targets that we can test them against.[20] Early ensembles were too narrow—they under-sampled from the extremes—while recent ones have included a large number of models that are unrealistically extreme.

This kind of misrepresentation is certainly not innocent. As noted above, the extreme scenarios matter in reasoning about climate policy. The misrepresentation of such extremes thus provides a clear argument against the use ensemble-based methods: since extant ensembles fail to properly account for extreme scenarios and failing to properly account for extreme scenarios will lead us to make the wrong kinds of decisions about climate policy, we shouldn't use extant ensembles.

I think this is the most important objection to the use of extant ensembles. Whether the argument succeeds largely depends on an empirical question, namely: in practice, how predictable are the deficiencies in extant ensembles, and how well can climate scientists account for them via adjusting the assumptions embedded in ensemble-based methods. After all, as Horowitz (2019) stresses, evidence that is predictably misleading is not really misleading

[20] See, e.g., Annan and Hargreaves (2011), Knutti et al. (2008, 2010), Tebaldi and Knutti (2007), and Tokarska et al. (2020).

at all—you simply have to correct for known errors. To my knowledge, however, no one has yet even attempted a systematic demonstration that climate scientists can't account for the known deficiencies in ensembles, and the success of ensemble-based methods relative to those methods that don't employ ensembles (see §2.3 and note 12) provides at least face-value evidence that the practical problems here are not insuperable.

So at the very least this argument against an ensemble-based method depends on empirical questions that haven't yet been answered. I think that there's an even stronger response to this objection, however, namely that insofar as we're worried that extant ensembles fail to accurately represent some important possibilities, ensemble-based methods and the probabilities that they generate are usually going to be our best methods for investigating this possibility and gaining a better understanding of what these extreme scenarios look like. This is true both in a relatively trivial sense—our evidence that extant ensembles misrepresent the extremes depends on empirical comparisons between the distribution of model results and known data in which the variation between model reports plays a crucial role—and also in a deeper one.

To illustrate the deeper point, consider the spread between model results, a quantity that has often been suggested (and used) as an alternative to the precise probabilities generated by ensemble-based methods (see, e.g., Carrier and Lenhard 2019; Jebeile and Barberousse forthcoming). Because the problem here concerns the models, however, and not our means of interpreting them, model spread is at least as likely to misrepresent as the probabilities generated by ensemble-based methods: if the ensemble doesn't include models that accurately represent extreme possibilities, then model spread won't cover those possibilities; if it includes models that are unrealistically extreme, then model spread will cover unrealistic extremities.

In fact the situation is even worse: the inter-model variation found in a given ensemble provides climate scientists with a means of "filling in" information from missing models or issuing corrections where the models are unrealistic. (This is true, for what it's worth, regardless of whether we assume that the sample is normally distributed around the truth or shares some other relationship with it.) So, for instance, if the ensemble under-samples from the extremes, empirical information and the actual distribution of models can be used to estimate what the tails of the distribution look like—that is, how "far" into extreme territory we should treat as realistic. Model spread provides with no similar principled means of making this demarcation—there's no principled way of extrapolating extant model spread to determine what spread a more

realistic ensemble of models would cover.[21]

In cases where the ensemble includes unrealistic models, similarly, probability distributions are essential both for evaluating whether (and to what degree) these unrealistic models bias the ensemble *as a whole* and for determining whether a given method of correction is effective. Tokarska et al. (2020, 1), for example, motivate their claim that extant ensembles misrepresent the extremes in part by pointing to the fact that the distribution of models within the ensemble is skewed in a way that model spread alone can't capture. And they evaluate the effects of their proposed correction in part by comparing the probability distributions generated by the corrected ensemble to past estimates (Tokarska et al. 2020, 8). Probability distributions are valuable in these comparisons because they contain information—information about inter-model variation, in particular—that qualitative measures like model spread do not.[22]

To reiterate from above, none of this is to say that ensemble-based methods are perfect or that the probabilities that they generate should be taken as the final word on a subject. Even empirically corrected ensemble-generated probability distributions may misrepresent the probability of extreme scenarios—the inter-model variation may be misleading. The point of the present section is that in cases where we're worried that the ensemble itself misrepresents a certain group of possibilities in particular (extreme warming scenarios, in this case), we're likely to be better off using of the variation between ensembles to generate probability distributions, because these distributions allow us to extrapolate from areas of relative confidence in a way that more qualitative approaches do not. In other words, ensemble-based methods provide more tools for investigating extreme scenarios than we would otherwise have. Rejecting ensemble-based methods because extant ensembles misrepresent certain scenarios is thus not just under-motivated, it's counter-productive.

4.3 Probabilities are too precise

The final objection that I want to discuss here is that the probabilities generated by ensemble-based methods are "too precise." Given the discussion

[21] Well, there are principled ways of doing so, but they work by piggybacking on the probabilities.

[22] Katzav et al. (2021) allege that precise probability functions "lose" information about uncertainty. That's true insofar as the contrast class is an (idealized) more complex probabilistic representation (Bradley and Drechsler 2014). It's at least not clear that it's true for any alternative to precise probability distributions on the table, however, and the opposite is true for any non-probabilistic alternative.

of the prior sections, it's easy to motivate this objection. As we've already seen, ensemble-based methods don't solve the problem of imprecise evidence. Ensembles allow us to make *more* precise judgments—in the setting of imprecise probabilities, they justify adopting a strictly smaller set of probability functions—but they don't warrant adopting the single precise probability function generated by the application of statistical tools. Hence the positive position outlined above: climate scientists should make use of the probabilities generated by these methods but not adopt them as though they were confirmed results.

Some philosophers, including most explicitly Katzav et al. (2021), seem to be inclined to a more extreme stance: their view is that climate science shouldn't make use of precise probability distributions at all, and they motivate their position on the grounds that such distributions misrepresent in virtue of being more precise than is warranted. So, as they put it: "When is a [probability density function] appropriate? A simple answer is: when it represents what our subjective probability *ought* to be given available evidence, including evidence concerning our uncertainty" (Katzav et al. 2021). They then go on to argue that the probability functions found in climate science don't meet this criterion and therefore are not appropriate.

The standard employed here is unrealistically strict, however. A probability function can be the *best* means of representing our uncertainty in a given context even if it is not a perfect one. The rejoinder I'm pushing here is in effect simply a more specific version of the response to idealizations that I made above: given the uncontroversial fact that all of our representational tools are imperfect, the question that we should be asking is not whether precise probability functions misrepresent in some way but whether they misrepresent in ways that undermine their utility for a given task or function. Or, in other words, the question is whether (in a given context) the best way to represent our uncertainty is with the a precise probability function—where "best" is to be judged not according to some standard of ideal accuracy but rather in terms of what representation best facilitates reliable inferences in the given context.

The arguments of philosophers are not well-suited for answering this kind of question; in keeping with the moderate view sketched in the last section, we should expect that whether precise probabilities are the best tool for representing uncertainty will depend on the details of the situation including empirical facts that we're not in the best position to judge. Unless Katzav et al. want to commit to the idea that any misrepresentation in this domain is damning—and I suspect they agree with me here that this position is non-starter—their arguments simply don't establish the extreme claim that climate scientists shouldn't ever use the precise probabilities generated by ensemble-

based methods.

It's unfair to attribute the extreme position just rebutted to Katzav et al., however. A more accurate reading would see them not as arguing for the rejection of the probabilities generated by ensemble-based methods but instead as advocating that climate scientists develop and use methods that don't yield precise probabilities. This suggestion is entirely amenable to thesis of this paper so long as the imagined methods take account of inter-model variation. *Perhaps*, for instance, it would generally be better to adopt ensemble-based methods that output imprecise probabilities rather than precise ones. It seems to me to be an open question whether such methods would in fact be preferable in practice, however, where this question—to reiterate—is one about the costs and benefits of adopting this different representational tool (compare Bradley 2019, §3.5). Perhaps imprecise probabilities buy us an increase in accuracy, but only a marginal one and only at substantial costs in terms of processing power, complexity, or the amount of data required. In such circumstances, it may not be worthwhile to use imprecise methods as opposed to precise ones. To be clear, however, this is an almost entirely unexplored area: it's one thing to say it would be good in principle for climate scientists to employ imprecise probabilities; it's another thing entirely to show that there are imprecise methods that can replace the precise ones discussed in section 2 or that these imprecise methods are preferable to the precise ones in the context of real problems.

Climate scientists need to represent aspects of the climate that are not perfectly understood. There are many desiderata for such representations. One is that they should accurately capture our uncertainty with respect to the feature in question. But others include that they should be as constrained by the empirical evidence as possible; that they should be mathematically tractable; that they should be reliable, accurate, and trustworthy in realistic (as opposed to heavily idealized) conditions; and that they should be informative and easily understood. I think that it's likely that the precise probabilities generated by ensemble-based methods will, in many contexts, be the best representational tool according to this suite of desiderata. Strictly speaking, however, my position is a weaker one, namely that taking account of the variation between models should be treated as one desideratum. Ignoring this variation amounts to ignoring a powerful source of evidence that offers concrete and empirically-demonstrated benefits to climate science. Since probabilistic (whether precise or imprecise) interpretations are the only ones that achieve this goal, we should prefer them to non-probabilistic approaches, all other things being equal.

5 Conclusion

This paper offers a conditional defense of the use of ensemble-based methods in climate science and the probabilities that they generate. There are three key takeaways. First, that climate modeling faces a problem due to what epistemologists call "imprecise evidence": we don't know (precisely) how to interpret the evidence produced by climate models. Second, that ensemble-based methods are able to mitigate this problem by making use of inter-model variation. Importantly, the value added by these methods is not merely philosophical; as we saw, there are least some cases where employing ensemble-based methods improves the accuracy and reliability of the results. Third, and finally, that the extant objections to ensemble-based methods (or the use of the precise probabilities that they generate) only really tell against views that would treat these methods as delivering direct access to the truth. Once we recognize that the probabilities in question should be (and often are) treated as evidence that must be interpreted rather than as expert functions that should be deferred to, it becomes clear that the question of whether they should be preferred to other representational tools is a largely practical question about costs and benefits.

References

Allen, Myles R. and Simon F. B. Tett (1999). Checking for Model Consistency in Optimal Fingerprinting. *Climate Dynamics* 15: 419–34.

Annan, James D. and Julia C. Hargreaves (2010). Reliability of the CMIP3 Ensemble. *Geophysical Research Letters* 37: 1–5.

— (2011). Understanding the CMIP3 Model Ensemble. *Journal of Climate* 24: 4529–38.

Betz, Gregor (2007). Probabilities in Climate Policy Advice: A Critical Comment. *Climatic Change* 85.1-2: 1–9.

— (2015). Are Climate Models Credible Worlds? Prospects and Limitations of Possibilistic Climate Prediction. *European Journal for Philosophy of Science* 5.2: 191–215.

Bradley, Richard and Mareile Drechsler (2014). Types of Uncertainty. *Erkenntnis* 79.6: 1225–48.

Bradley, Seamus (2019). Imprecise Probabilities. In: *Standford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. URL: https://plato.stanford.edu/entries/imprecise-probabilities/.

Carr, Jennifer Rose (2019). Imprecise Evidence Without Imprecise Credences. *Philosophical Studies* (online first).

Carrier, Martin and Johannes Lenhard (2019). Climate Models: How to Assess Their Reliability. *International Studies in the Philosophy of Science* 32.2: 81–100.

Carroll, Raymond J. et al. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd edition. Boca Raton: Chapman & Hall/CRC.

Dethier, Corey (2022). When is an Ensemble Like a Sample? 'Model-Based' Inferences in Climate Modeling. *Synthese* 200.52: 1–20.

Dorst, Kevin (2019). Higher-Order Uncertainty. In: *Higher-Order Evidence: New Essays*. Ed. by Mattias Skipper and Asbjørn Steglich-Petersen. Oxford: Oxford University Press: 35–61.

Dorst, Kevin et al. (forthcoming). Deference Does Better. *Philosophical Perspectives*.

Frigg, Roman and Leonard A. Smith (forthcoming). An Ineffective Antidote for Hawkmoths. *European Journal for Philosophy of Science*.

Gettelman, Andrew and Richard B. Rood (2016). *Demystifying Climate Models: A Users Guide to Earth System Models*. Springer.

Hannart, Alexis, Aurélien Ribes, and Phillippe Naveau (2014). Optimal Fingerprinting under Multiple Sources of Uncertainty. *Geophysical Research Letters* 41: 1261–68.

Horowitz, Sophie (2019). Predictably Misleading Evidence. In: *Higher-Order Evidence: New Essays*. Ed. by Mattias Skipper and Asbjørn Steglich-Petersen. Oxford: Oxford University Press: 105–23.

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

Huntingford, Chris et al. (2006). Incorporating Model Uncertainty Into Attribution of Observed Temperature Change. *Geophysical Research Letters* 33.L05710: 1–4.

IPCC (2013). *Climate Change 2013: The Physical Science Basis*. Ed. by Thomas F. Stocker et al. Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press.

Jebeile, Julie and Anouk Barberousse (forthcoming). Model Spread and Progress in Climate Modelling. *European Journal for Philosophy of Sciece*.

Katzav, Joel (2014). The Epistemology of Climate Models and Some of its Implications for Climate Science and the Philosophy of Science. *Studies in History and Philosophy of Science Part B* 46: 228–38.

Katzav, Joel et al. (2021). On the Appropriate and Inappropriate Uses of Probability Distributions in Climate Projections, and Some Alternatives. *Climatic Change* 169.15: 1–20.

Knutti, Reto et al. (2008). A Review of Uncertainties in Global Temperature Projections over the Twenty-First Century. *Journal of Climate* 21.11: 2651–63.

Knutti, Reto et al. (2010). Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate* 25.10: 2739–58.

Knutti, Reto et al. (2017). A Climate Model Projection Weighting Scheme Accounting for Performance and Interdependence. *Geophysical Research Letters* 44: 1909–18.

Lenhard, Johannes and Eric Winsberg (2010). Holism, Entrenchment, and the Future of Climate Model Pluralism. *Studies in History and Philosophy of Science Part B* 41.3: 253–62.

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

Mahtani, Anna (2019). Imprecise Probabilities. In: *The Open Handbook of Formal Epistemology*. Ed. by Richard Pettigrew and Johnthan Weisberg. PhilPapers Foundation: 107–30.

McGuffie, Kendal and Ann Henderson-Sellers (2014). *The Climate Modeling Primer*. 4th edition. Chichester: Wiley Blackwell.

Parker, Wendy S. (2010a). Predicting Weather and Climate: Uncertainty, Ensembles and Probability. *Studies in the History and Philosophy of Modern Physics* 41: 263–72.

— (2010b). Whose Probabilities? Predicting Climate Change with Ensembles of Models. *Philosophy of Science* 77.5: 985–97.

— (2013). Ensemble Modeling, Uncertainty and Robust Predictions. *Wiley Interdisciplinary Reviews: Climate Change* 4: 213–23.

— (2020a). Evidence and Knowledge from Computer Simulation. *Erkenntnis* (online first).

— (2020b). Model Evaluation: An Adequacy-for-Purpose View. *Philosophy of Science* 87.3: 457–77.

Parker, Wendy S. and James S. Risbey (2015). False Precision, Surprise and Improved Uncertainty Assessment. *Philosophical Transactions of the Royal Society Part A* 373.3055: 20140453.

Parker, Wendy S. and Eric Winsberg (2018). Values and Evidence: How Models Make a Difference. *European Journal for Philosophy of Science* 8.1: 125–42.

Rougier, Jonathan (2016). Ensemble Averaging and Mean Squared Error. *Journal of Climate* 29.24: 8865–70.

Roussos, Joe (2020). Policymaking Under Scientific Uncertainty. PhD dissertation. London School of Economics.

Roussos, Joe, Richard Bradley, and Roman Frigg (2021). Making Confident Decisions with Model Ensembles. *Philosophy of Science* 88.3: 439–60.

Sanderson, Benjamin M., Reto Knutti, and Peter M. Caldwell (2015). A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble. *Journal of Climate* 28: 5171–94.

Schmidt, Gavin A. and Steven C. Sherwood (2015). A Practical Philosophy of Complex Climate Modelling. *European Journal for Philosophy of Science* 5.2: 149–69.

Schurer, Andrew P. et al. (2018). Estimating the Transient Climate Response from Observed Warming. *Journal of Climate* 31.20: 8645–63.

Sedláček, Jan and Reto Knutti (2013). Evidence for External Forcing on 20th-century Climate from Combined Ocean-atmosphere Warming Patterns. *Geophysical Research Letters* 29.20: 1–5.

Stainforth, David A. et al. (2007). Confidence, Uncertainty and Decision-Support Relevance in Climate Predictions. *Philosophical Transactions of the Royal Society Series A* 365.1857: 2145–61.

Stott, Peter A. et al. (2006). Observational Constraints on Past Attributable Warming and Predictions of Future Global Warming. *Journal of Climate* 19.13: 3055–69.

Tebaldi, Claudia and Reto Knutti (2007). The Use of the Multi-Model Ensemble in Probabilistic Climate Projections. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 365.1857: 2053–75.

Teller, Paul (2004). What is a Stance? *Philosophical Studies* 121.2: 159–70.

Thompson, Erica L. and Leonard A. Smith (2019). Escape from Model-land. *Economics* 13.1: 1–15.

Tokarska, Katarzyna B. et al. (2020). Past Warming Trend Constrains Future Warming in CMIP6 Models. *Science Advances* 6.12: 1–13.

Weitzman, Martin L. (2012). GHG Targets as Insurance Against Catastrophic Climate Damages. *Journal of Public Economic Theory* 14.2: 221–44.

Winsberg, Eric (2018). *Philosophy and Climate Science*. Cambridge: Cambridge University Press.