Statistically-indistinguishable ensembles and the evaluation of climate models

Abstract
Over the last decade, Annan and Hargreaves (2010, 2011b, 2017) have argued for a new “paradigm” for the interpretation of ensembles climate models that promises to allow us to better justify drawing inferences from them. I argue that the apparent benefits of their view are illusory: the assumptions required by the view are strong enough to justify drawing inferences from an ensemble regardless of the choice of interpretative “paradigm.” I end by sketching how climate scientists might go about supporting these assumptions

0 Introduction

In order for an inference from a feature of a sample to a feature of a population to be justified, the sample must be representative of the population. In principle, representativeness can be secured in two distinct ways. First, by construction: the use of a particular sampling procedure can make it more likely that the sample is representative. Second, by empirical evaluation: the properties of a given sample can be checked against another sample from the same population (or some other proxy), and if it passes this check, then there’s a good chance that the original sample is representative.

The same reasoning applies to ensembles of models, particularly ensembles of climate models (see Parker 2011, 2018): in order to justify the inference from a feature of the ensemble to some conclusion about the world, the ensemble must be “representative” of the world in the appropriate way. As before, representativeness can be secured in two distinct ways. In the case of ensembles of climate models, justification by construction looks hopeless; it’s widely thought that extant ensembles are in no way analogous to a randomly-selected sample. They are, instead, “ensembles of opportunity,” constructed in a non-principled manner and in no way guaranteed to represent the full space of possible models (Knutti et al. 2010; Tebaldi and Knutti 2007). Things are murkier when it comes
to justifying inferences from ensembles by means of empirical evaluation, with the issue apparently turning on the choice of “paradigm” (Annan and Hargreaves 2010) or “framework” (Sanderson and Knutti 2012) for determining what it means for an ensemble to be representative. In particular, Annan and Hargreaves (2010, 2011b, 2017) argue that under a “statistically-indistinguishable” interpretation, extant ensembles do allow us to make inferences about the future climate.

In this talk, I argue that the apparent benefits of the statistically-indistinguishable view are illusory: the assumptions required by the view are strong enough to justify drawing inferences from an ensemble regardless of the choice of interpretative framework. I begin by briefly laying out the difference between the statistically-indistinguishable view and its primary competitor (§1). I then turn to drawing out the assumptions of the view and showing that they’re strong enough to justify drawing conclusions on either interpretation (§2). In spite of this conclusion, I think Annan and Hargreaves are on the right track, and I end by considering the implications for arguments involving ensembles of models and how climate scientists might go about supporting the assumptions that I’ve argued are necessary (§3).

1 Two frameworks for interpreting ensembles of climate models

Prior to the last decade, the standard practice for evaluating ensembles of models involved a method that the literature (now) characterizes as treating the ensemble as though it were “sampled from a distribution centred on the truth” (see, for example Glecker, Taylor, and Doutriaux 2008; Knutti et al. 2008). The appropriate analogy here is with a representative sample of a population: if we sample from a population where the biases or errors are uncorrelated, the expected outcome is a sample that is normally distributed around a mean value that will increasingly accurately represent the true value of the variable of interest as the size of the sample grows. While the scientists themselves call this a “paradigm” (Annan and Hargreaves 2010) or “framework” (Sanderson and Knutti 2012) for the “interpretation” of ensembles of models, it is essentially a claim about the nature of the relationship between the ensemble and the target system, namely that the former is “centered” on the latter. Call this the “truth-centered” view.

The alternative to the truth-centered view has emerged more recently, with some climate scientists proposing a “paradigm” in which the models are treated as “statistically indistinguishable
from the truth” (Annan and Hargreaves 2010, 2011b, 2017). This notably technical language is
drawn from Toth et al. (2003) and is glossed as the idea that the ensemble of models and the
target system are both drawn from the same population. Again, the view concerns the relationship
between the ensemble and the target system and the appropriate analogy is to a typical sampling
procedure. If the truth-centered view holds that the ensemble is centered on the target system in
the same way that a representative sample is centered on a population mean, then the “statistically-
indistinguishable” view holds that the relationship between an ensemble and the target system is
like (read: involves the same statistical or evidential relationship as) the relationship between prior
sample and the next sample.

The difference being flagged here may be unfamiliar; allow me to briefly spell it out with a
simpler example. Consider an ensemble of thermometers. If the errors in each thermometer are
independent, we can expect the readings given by the thermometers to be normally distributed
around the true temperature, meaning that (irrespective of the spread) the larger our sample size,
the better the mean reading, \( \hat{t} \), will be an as estimate of the true temperature, or \( T \).

Employing the ensemble in order to estimate \( T \) provides an exemplar of sampling from a distribution centered
on the truth. Now suppose we employ the same ensemble of thermometers for a different task,
around that of predicting the value recorded by the next thermometer. Notice that the prediction
will be the same as in the previous case: the expected value for the next thermometer is the mean
of the ensemble, \( \hat{t} \), just as the estimate for the true temperature was. In this case, however, what
matters for the accuracy of our estimate is not the number of samples but their spread. Even
if the population of thermometers (or readings) is stipulated to be identical, increasing the size
of the sample won’t be able to constrain our estimate beyond the spread observed in the actual
population. This example is a paradigm case in which the sample and the target system are
“statistically indistinguishable.”

The thermometer example oversimplifies things (at least relative to the climate modelling case)
in at least one important respect: there’s a degree of freedom present in the climate modeling case

---

1The independence assumption is crucial; if it’s dropped, things can become substantially more complicated, as
distributions in which means are undefined and the law of large numbers fails to hold become real possibilities. Note
that we have no good reason to assume that the independence assumption holds in climate modeling (in general).

2The mathematical reason why is simple: in the truth-centered case, we’re asking about the probability of \( n \)
readings given the hypothesis about the true value; in the statistically-indistinguishable case, we’re asking about the
probability of one reading given the (presumably same) hypothesis about the true value.
that’s not present in the thermometer case. In the thermometer case, we know how the sample relates to the target because we know how the sample relates to the population—that’s secured by the sampling procedure—and we know how the population relates to the target. There’s enough causal (or, in the case of a population mean, stipulative) structure to determine whether we’re dealing with a truth-centered case or a statistically-indistinguishable one. That structure doesn’t exist in the climate modeling scenario; because there’s no well-defined population of models that we’re drawing the ensemble from, it’s more accurate to say that the relationship between an ensemble and the target is like that found in a truth-centered or statistically-indistinguishable sample. As we’ll see, this extra degree of freedom is important.

2 The strong assumptions of the statistically-indistinguishable view

For present purposes, the most important fact about the different interpretative frameworks is that they have different empirical implications, meaning that we should be able to tell which framework is the right one by examining how the ensemble of models does at predicting known data. Two empirical predictions are particularly noteworthy. First, if the ensemble is truth-centered, then the mean value for an estimated variable can be expected to be closer to the truth than that of any of the individual models. That expectation doesn’t hold in the indistinguishable case (Annan and Hargreaves 2011b, 4531)—though in both cases, the strength of the expectation depends on other factors, particularly the relationship between ensemble size and the dimensionality of the estimate. Second, if the ensemble is truth-centered, the variability of the target system should be low relative to the spread of the models. Specifically, we should expect observed values to be closest to the models at the center of the distribution much more frequently than they are closest to the outliers. Once again, this result doesn’t hold in the context of an indistinguishable ensemble, where we can expect that model spread and population variability are virtually identical, and thus that observed values should be closest to the models at the center of the distribution at roughly the same frequency that they are closest to the outliers.³

³Note, crucially, that there should still be more models near the center, and so the observations should be closer to the center of the distribution more frequently than they fall at the edges. The above condition is just an intuitive way of expressing the idea that the rank histogram should be flat in the statistically-indistinguishable case.
—have argued that climate model evaluations demonstrate that the empirical predictions of the statistically-indistinguishable view are born out—though they acknowledge that the data are somewhat messy and suggest that extant ensembles are neither perfectly truth-centered nor perfectly statistically-indistinguishable, but are merely closer to the latter (see, e.g., Annan and Hargreaves 2011b, 4536). Sanderson and Knutti (2012) explain this messiness by pointing out that an extant ensemble “may have elements of both interpretations that are not contradictory” and suggest that the explanation for the apparently mixed data is the fact that extant ensembles are tuned to contemporary data. The result of tuning is that ensembles will effectively act like a statistically-indistinguishable ensemble that oversamples around the present-day targets—giving them some truth-centered properties relative to this data, but also leading to the expectation that the behavior should be more like a pure statistically-indistinguishable ensemble as predictions move further into the future.

If right, that’s great news (epistemically speaking), because the statistically-indistinguishable view promises to allow us to make inferences that look illegitimate from the truth-centered view (Annan and Hargreaves 2010). In particular, if the truth-centered view were the correct one, then we wouldn’t be justified in thinking that extant ensembles surveyed the full range of possibilities—we should expect that the ensemble has too little variation relative to the population of possible scenarios that we want to capture. Historically, this has been one of the major arguments against making inferences from ensembles of climate models (see, e.g., Knutti et al. 2010). On the assumption that the statistically-indistinguishable view is the correct one, however, the amount of variation observed in the ensemble is much closer to what we would expect, meaning that we can much less worried that there are “extreme” cases that are both possible and not-represented by our ensemble (Annan and Hargreaves 2011a).

Unfortunately, I don’t think that this conclusion is justified. It’s helpful to return to the thermometer example one more time. Suppose that thermometers made by a single manufacturer tend to have less variation than thermometers in general—that is, if you learn that two thermometers share a manufacturer, you should expect them to be closer in reading than two thermometers selected randomly will be (on average). The situation we find ourselves in is one in which we’re evaluating a given sample against a set of thermometers that share a given manufacturer, and that this evaluation process indicates that the sample is statistically indistinguishable from the population
that we’re evaluating it against. If we have good reason to suppose that the target thermometer will also be made by the given manufacturer, then the statistically-indistinguishable framework offers us a good means of making predictions about this target. If, by contrast, we have good reason to doubt that the next thermometer will also be made by the given manufacturer, we have good reason to doubt that the statistically-indistinguishable framework offers us good inferences—because we have good reason to doubt that our sample is really representative of the one that the target thermometer will be drawn from.

This situation just outlined is essentially the one we find ourselves in in climate science: in order to justify employing the statistically-indistinguishable interpretation, we have to assume that the future climate will be like the present climate in the (very) specific sense that models that do well at predicting present climate data will also do well at predicting future climate data. This isn’t simply a Humean induction problem. The climate is extraordinarily complex, and it’s well-known that ability to capture one aspect does not always track ability to capture other aspects (see Parker 2018). It may well be that imperfect models that nevertheless do a good job with the natural and direct forcing conditions exhibited in present climate data don’t do a good job at handling the feedback loops that we expect to become more important in the future. In other words, the assumption is substantive: given the limits of the models, it would not be at all surprising to find that the models that are the best at capturing present trends are not the best at predicting future ones (compare Frisch 2015).

For now, assume that we can make this assumption: the models that do best at capturing present trends are also those with the most accurate predictions. On this assumption, the fact that the ensemble has too little variation for a good truth-centered sample becomes unproblematic. As I noted above, the issue on the truth-centered interpretation is that there might be possibilities that our ensemble doesn’t represent because it’s too narrow—or, to put it slightly differently, the worry is that the ensemble as a whole might be over-fitted to present conditions. So if we can guarantee that over-fitting to present conditions is not a problem—by assuming that models that do well in present conditions will also do well under future conditions—then we’re just as free to use the truth-centered interpretation as we are to use the statistically-indistinguishable one.

The above argument does oversimplify in some ways; it’s certainly not the case that two interpretations will always be equally well-justified or rely on identical assumptions. My claim is just
that given the specific worries that undermine drawing inferences on the truth-centered view, the statistically-indistinguishable view is of no help. The worry is that there’s a lack of variation in extant ensembles (see Knutti et al. 2010; Tebaldi and Knutti 2007), and switching interpretations doesn’t buy us more variation.

3 The implications for climate science and ensembles of models

Where does the foregoing leave us? Essentially where we started: it’s difficult to justify arguments from an ensemble of climate models having some feature to the world being likely to have that feature (for recent overviews, see Baumberger, Knutti, and Hadorn 2017; Parker 2018). I suspect that this conclusion strikes many as pessimistic. I’m not inclined to see it that way. I’ll end this talk by showing that the failure of ensemble-based arguments is much less consequential than it may initially appear and argue that—as Annan and Hargreaves’ work indicates—what’s needed to better support these arguments is empirical evidence.

The worst-case scenario for arguments from ensembles of climate models would be an ensemble where every model makes accurate predictions if and only if each of the others do. In a situation like that one, having an ensemble of models would be no different from having a single model. In other words, even if we gain nothing from having ensembles of models, we can’t be epistemically worse off than we would be if we only had a single model—at least to the extent that the various models are in agreement. Given this fact, how should we view the epistemic benefit of ensembles of models? My suggestion (in keeping, I think, with recent work on robustness more generally; see Schupbach 2018; Staley forthcoming) is that ensembles of models serve to account for particular sources of error. An ensemble should not be viewed as providing a certain amount of confidence in a result that’s separate from the confidence provided by the individual models. Instead, each additional member of the ensemble should be seen as providing some additional confidence due to the fact that it approaches the problem in a slightly different way than the other models do. Rather than the ensemble as a whole being representative of some population of models that has some sort of special connection with the truth, we should see the models as tools that we have confidence in and the ensemble as a whole as a way to correct for possible errors that might arise in the use of a single tool.
As such, it would be a mistake to see work that undermines arguments from ensembles as (directly) undermining the evidence for the conclusions or results of climate science. The proper way to view this work, instead, is as showing that building ensembles of models is not sufficient to address all of the possible sources of error. Other methods need to be employed as well. What I’ve shown in the foregoing is that ensembles of models—or extant versions thereof—are not particularly useful for accounting for the sorts of errors that might arise from the fact that the future climate will be heavily affect by factors that don’t make a substantial difference to the present climate. In order to account for that source of error, we need something different, namely the ability to evaluate how well our models do at capturing or predicting rapid feedback-driven changes to climate. If it’s true that extant models are quite good at representing the climate under such conditions, then we have good reason to think that this error has been accounted for.

My suggestion, in other words, is that the assumption that I indicated was necessary for the justifiable use of the statistically-indistinguishable interpretation is one that is justifiable. We can evaluate whether or not it holds. In doing so, we’ll address the particular worry that motivated the move the statistically-indistinguishable interpretation in the first place—not by switching “paradigms” but instead by increasing our knowledge of how the models work and when they’ll be successful. Our foray into Annan and Hargreaves’ suggestion that we switch to a statistically-indistinguishable interpretation of ensembles of climate models, in other words, hasn’t left us entirely where we started, as it makes it clearer precisely what information we would need to address the worries that we were responding to in the first place.

4 Conclusion

In this talk, I’ve argued that the apparent benefits of the statistically-indistinguishable view are illusory: the assumptions required by the view are strong enough to justify drawing inferences from an ensemble regardless of the choice of interpretative framework. Along the way, I’ve tried to show why the failure of arguments from ensembles of models shouldn’t be seen as devastating and indicated how empirical research can better support the conclusions that we’re interested in.
References


