Synthese (2022) 200:52
https://doi.org/10.1007/s11229-022-03477-5

ORIGINAL RESEARCH

When is an ensemble like a sample? "Model-based" inferences in climate modeling

Corey Dethier[1]

Received: 26 October 2020 / Accepted: 27 October 2021 / Published online: 28 February 2022

Abstract
Climate scientists often apply statistical tools to a set of different estimates generated by an "ensemble" of models. In this paper, I argue that the resulting inferences are justified in the same way as any other statistical inference: what must be demonstrated is that the statistical model that licenses the inferences accurately represents the probabilistic relationship between data and target. This view of statistical practice is appropriately termed "model-based," and I examine the use of statistics in climate fingerprinting to show how the difficulties that climate scientists encounter in applying statistics to ensemble-generated data are the practical difficulties of normal statistical practice. The upshot is that whether the application of statistics to ensemble-generated data yields trustworthy results should be expected to vary from case to case.

Keywords Climate models · Ensemble methods · Statistics · Model-based

1 Introduction

Much of contemporary science is driven by either models or computer simulations carried out using models. Climate science offers a particularly notable example; as we'll see, climate models are essential in estimating humanity's contribution to climate change. In climate science, it's common—arguably even standard practice—for the results provided by computer simulations using climate models to be treated as data. In particular, it's common for climate scientists to apply statistics to the different results generated by "ensembles" of climate models—to treat these results in the same way that they might treat data generated by sampling from a population. The resulting

This article belongs to the topical collection "Recent Issues in Philosophy of Statistics: Evidence, Testing, and Applications", edited by Molly Kao, Deborah Mayo, and Elay Shech.

✉ Corey Dethier
corey.dethier@gmail.com

[1] Leibniz Universität Hannover, Hannover, Germany

probability distributions are usually taken to be a more accurate representation of the climate than is provided by any one model.

From a traditional or "design-based" perspective on statistical inference, this practice might look odd and, indeed, it has been widely criticized by both climate scientists and philosophers of science.[1] After all, the models in an ensemble are not sampled randomly from a space of possible models (whatever that space looks like). Nor, as Winsberg, (2018, p. 99) points out, do we have any other good reason for thinking that our construction procedure will generate models that are normally distributed around the truth. So what, if anything, justifies the practice of applying statistics to the data produced by ensembles?

In this paper, I argue that statistical inferences from ensemble-generated data are justified in the same way as any other statistical inferences. The crucial step in both cases is specifying the probabilistic relationship between the target and the sample—that is, what's essential is that we can justify our choice what's called the "statistical model."[2] Or, more simply: we can treat the results generated by an ensemble of models like a sample when we know enough about the relationship between these results and the target to justify assigning specific likelihoods to the relevant results.

This view might be termed "model-based" for two different reasons. First, I'm endorsing a view of statistical inference according to which actual sampling procedures take a back seat to the choice of statistical model, and such views have often been termed "model-based" (Zhao, 2021). Second, my account of statistical inference in modeling contexts is consciously patterned on what are called "model-based" views of measurement (for an overview, see Tal, 2020, Sect. 7). While to my knowledge the terminology was independently developed, this coincidence is appropriate: both views stress that perfect experimental design is neither necessary nor sufficient for successful inference, and hold that what is instead essential is an accurate understanding of the probabilistic relationship between the target and the data that are generated by the experiment.

The upshot for the discussion of climate modeling is that general criticisms leveled at the very idea of applying statistics to the data generated by ensembles of models are misguided. While it is plausible that some of these applications are flawed, if they are it's for reasons familiar from paradigm examples in statistics, namely that the relevant samples are too small or not genuinely representative. And these familiar difficulties can be addressed so long as we understand how they affect the probabilistic relationship. Non-representative samples make statistical inference more difficult, but they don't make it impossible.

I begin the paper with a sketch of a model-based account of scientific inference, building largely on model-based accounts of measurement (Sect. 2). I then offer an abstract argument for the central conclusion, urging that the model-based picture

[1] See Betz (2015), Carrier and Lenhard (2019), Katzav (2014), Parker (2010b; 2010c; 2013), Parker and Risbey (2015), Stainforth et al. (2007) and Winsberg (2018).

[2] I'll be using "essential" in formulation of this thesis throughout this paper. I choose this term because "necessary" alone would imply not sufficient. On some views of statistics, however, an appropriately-warranted model of the probabilistic relationship is both necessary and sufficient. On other views, Bayesian ones for example, we need additional information (i.e., an assignment of priors for possible values of the quantity of interest).

given in the first section can be extended to cases in which the data are generated by an ensemble of models (Sect. 3). With the abstract picture on the table, I examine a case study–the use of "errors-in-variables" methods in climate change attribution—that illustrates that the application of statistics to ensemble-generated data can be successful in at least some cases (Sect. 4). The final section returns to the literature's criticisms of the application of statistics to ensemble–generated data and discusses whether and to what degree my arguments can be extrapolated (Sect. 5).

2 "Model-based" statistical inference

To derive knowledge from an instrumental reading, an agent must know how to interpret that reading. This interpretative step is often non-trivial. Consider the use of a pendulum to measure the radius of the earth.[3] Since the period of a pendulum is a function of the force acting on it, and the force is a function of distance from the center of the earth, the period of a pendulum can be used as a proxy for the distance from the surface to the center. The instrumental reading—in this case, the period of the pendulum—is distinct from the "measurement result"—the value for the radius that is *inferred* from the readings—and the inference from the former to the latter relies on substantive background knowledge about the workings of a pendulum, the effects of gravitation, and the off-setting centrifugal effect that acts on the pendulum due of the earth's rotation. That is, it relies on the existence of appropriately warranted "model" of the relationship between the period of the pendulum and the earth's radius. Accounts of measurement that place the epistemological focus on the model of the measurement process have come to be called "model-based" (for an overview, see Tal, 2020, Sect. 7). On model-based views, the crucial step in justifying the measurement outcome is justifying the choice of licensing model. My aim in this first section is to sketch a view of scientific inference—including statistical inference in particular—that is "model-based" in this same manner. This general discussion will set up the argument, given in the next section, that statistical inferences from ensemble-generated data should be understood in the same way.

It's helpful to begin by returning to the case of the pendulum. As Morrison (2015) argues in her treatment of this case, rarely is the pendulum so perfectly shielded, isolated, and controlled that there are *no* influences affecting the reading other than the radius of the earth. Our single reading may be affected, for example, by user error or by a systematic influence such as a non-uniformity in the density of the crust. Just as it is impractical to physically screen-off all of these other potential influences, it is equally impractical to explicitly account for all of the possible influences in the model of the relationship between the radius of the earth and a single reading. Rather than trying to either physically isolate the causal factor we're interested in or explicitly account for every possible influence on the instruments, therefore, scientists often account for these other influences by "black-boxing" them, where this means building a model in which random variables are used to represent some of the factors that affect the instrumental reading.

[3] For further discussion of this example, see Dethier (2021) and Morrison (2015).

A simple example is as follows. Suppose we have a data set of calculated distances for the length to the center of the earth based on our instrumental readings; call this data set Y. If we suppose that each element of Y was measured at the same location and let r represent the true distance between the center of the earth and that location, our statistical model might characterized with a simple equation, $Y = r + f_X$, where r is a constant and f_X is a random variable representing the random error in our measurements. Of course, this model will be inappropriate—it will be a *mis*-specification—if the elements of Y were measured at different locations rather than at a single location. In that case, we'll need a different model. On a first pass, for instance, we might expect that $Y = r(\theta, \psi) + f_X$ where $r(\theta, \psi)$ is a function representing the true distance from the center of the earth to different points on the surface.

A statistical model represents the probabilistic relationship between the target of the inference and the data. For the purposes of this paper, we can understand "probabilistic relationship" here in terms of likelihoods.[4] So our data Y has a certain likelihood on different hypotheses about the nature of r. These likelihoods are specified by the statistical model, and the quality of the statistical model is a function of how accurately it represents these likelihoods. For the conclusions of statistical inferences to be justified, the statistical model must generate accurate (enough) likelihoods; if it doesn't—if the model is "mis-specified"—the conclusion is not warranted.

Statisticians take two different perspectives to statistical models. On one view, often called "design-based," the importance of statistical models is downplayed relative to the importance of certain sorts of sampling procedures—randomization, blinding, etc.—by which an experimenter can directly control how the data set is generated. On the other view, traditionally call "model-based" (Zhao, 2021), these physical procedures are important only insofar as they serve to fix the probabilistic relationship between the data and the target. On this latter view, what's essential for successful statistical inference is just we're able to justify the assumption that the statistical model accurately represents this relationship. Deviations from a perfectly designed experiment—non-random sampling, to use the traditional example—are acceptable so long as the model accounts for these deviations.

There may well be good reasons to prefer a design-based view in practice; from a philosophical perspective, however, the model based-view is the superior one. To see why, recall the problem that we began with, namely, a situation where we're working with a single data point that is influenced by a number of different factors. Fundamentally, the problem in this case is *not* that the sampling procedure is faulty or the sample size too small. It's that we don't know enough about the relationship between this single data point and the target. Were we in a position to know how every factor other than the one we were interested in measuring affects our single data point, we could draw successful inferences even from this single-element sample. Since it's impractical to come by this knowledge, however, we design experiments that don't require it. Nevertheless, on the model-based picture, there's nothing wrong with non-

[4] Cashing "probabilistic relationship" out just in terms of likelihoods should be understood as something of an idealization, as often we'll need more detailed probabilistic data than the simple likelihood relationship between actual data and a given hypothesis space. So, for instance, in doing classical statistics we need to know the likelihood of any data "at least as extreme as" the observed data. In what follows, I'm going to forego these complications for sake of generality and ease of exposition.

standard or non-ideal sampling procedures in principle. On the contrary, the advantage of standardarized procedures is merely practical: they make the choice of statistical model—what is usually called "model specification"—*easier*. As the statistician Fred Smith puts it: "The advantage of randomization is that if a randomized design has been employed no further justification is needed; the whole scientific community will accept the sample that has been selected. With other forms of sampling users would need to be convinced in each case" (Smith, 1983, p. 399).

Again from a philosophical perspective, a crucial advantage of the model-based view is that it explains why we can use statistics in cases—like our running example—where no actual sampling takes place. With our imagined pendulum, there is no real population that is actually being sampled. Instead, the statistical reasoning employed is essentially *analogical*; the idea here is we choose or "specify" a model such that the processes that determine the character of the data behave *like* the probabilistic sampling procedure in the model.[5] As such, the statistical model isn't required or expected to represent the underlying mechanisms that actually serve to generate the data. Instead, it is successful if and only if it accurately represents the probabilistic relationship between the target and the data. In applying statistics to this case, therefore, we're not supposing that the actual mechanisms of data generation are anything like a sampling procedure. Much more minimally, we're supposing that the these mechanisms—whatever they are—have similar probabilistic implications on the data as a sampling procedure would, where again we can understand this in terms of the likelihood values given by the model being near (enough) to the real likelihoods.

The foregoing offers a rough picture of a "model-based" account of scientific inference, with a particular focus on statistical inferences. The two important takeaways are the following. First, statistical inferences rely on justifying the choice of a particular statistical model that represents the relationship between the target and the data in the sense that it delivers likelihood values for the data on different hypotheses about the target. Second, the actual mechanisms of data generation are not relevant to the justification of statistical inferences. That is, varying these conditions makes no difference to justification if we hold fixed the accuracy of the likelihood values delivered by the statistical model. Instead, a statistical inference is justified if and only if and insofar as the choice of statistical model that licenses the relevant inference is itself justified; enacting physical controls on experiments and ensuring (e.g.) randomization makes it (much!) easier to justify the choice of statistical model but is strictly speaking not necessary.

3 Statistical inferences from ensembles

The last section put forward a "model-based" view of statistical inference, according to which the actual mechanisms of data generation are relevant to the justification of statistical inferences only insofar as they serve to fix the probabilistic relationship between the data and the target: if we fix the assumption that our statistical model

[5] One can find statements to essentially this effect throughout the statistical literature. See, e.g, Cox and Wermuth (1996, p. 13), Kass (2011, p. 2), Royall (1992, p. 229).

accurately represents this relationship, then variation in the actual mechanisms of data generation makes no difference to the justification of the inference. In the present section, I'm going to assume that this basic picture is correct, and argue that on this assumption, there are no in-principle barriers to the application of statistics to the results generated by running simulations on the different models that make up an ensemble. What's essential in both the modeling and experimental case is that we can justify the choice of statistical model; while the actual mechanisms of data generation can make this more or less difficult, they don't change the fundamental picture.

Over the last two decades, a number of philosophers of science have argued that the results of calculations or simulations carried out using models can serve as evidence for hypotheses about the world.[6] To illustrate the point, it will be helpful to consider a schematized picture of a climate model.[7] Our imagined model represents the earth's atmosphere: it consists of a number of gridded shells, with each shell representing a layer of the atmosphere and each grid box a location in that layer (individual grid boxes can then be picked out using the triple $\langle \theta, \psi, r \rangle$). 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 final component of the model is a series of equations that determine how a change in the climate variables of one box affects its neighbors. At the simplest level, quantities like heat will simply defuse through the system, but of course there are more complicated effects as well. So, just to take the most basic but also the most important example, the wavelength of the energy going down through the atmosphere is different from the wavelength of that going up, and the latter is affected by greenhouse gas concentrations in a way that the former isn't. The upshot is that the movement of energy up through the atmosphere works differently from the movement of energy down, and our model must account for these differences.

As outlined, the model encodes our assumptions and knowledge (and more besides; see below) about the climate system. To use this model to generate evidence, climate scientists run simulations of the evolution of the system under different conditions. So, for instance, a common simulation involves doubling the CO_2 concentration in the model and then repeatedly solving the various equations to determine what effects this doubling has on the rest of the climate variables that the model represents (see Eyring et al., 2016). The results of this kind of simulation is that the variables in the model take on new states. If $T(r, \theta, \psi, t)$ represents the average temperature in the grid box picked out by the spherical coordinates $\langle r, \theta, \psi \rangle$ at time t, then running the simulation would yield a series of values $T(r, \theta, \psi, 0) = x$, $T(r, \theta, \psi, 1) = x'$, etc. I'm going to call these values and any basic calculations that can be made using them (e.g., $\Delta T(r, \theta, \psi)$ or $\bar{T}(t)$) "model reports." This term should be understood as analogous to "instrumental readings" in the sense that where the latter are "directly" observable features of an instrument such as the location of a pointer on a scale, the former are directly "observable" features of the model. And we can learn from these features,

[6] See, e.g., Lusk (2016), Maki (2005), Morgan (2002), Morrison (2015), Parker (2009; in press) and Winsberg (2010).

[7] The picture I'll be presenting is, of course, greatly simplified in a number of ways. For a deeper discussion, see one of the extant climate model primers such as Gettelman and Rood (2016) and McGuffie and Henderson-Sellers (2014). For a philosophical introduction, see Winsberg (2018, pp. 27–54).

at least in the right circumstances. Given that we are not logically omniscient agents and climate models (in particular) are extremely complicated, we won't know what our assumptions indicate the climate would be like if CO_2 concentrations doubled. Observing the model reports thus serves as evidence both for what our assumptions entail and thereby (to the degree that those assumptions are reliable) for what the world is in fact like.[8]

As is true of instrumental readings, background knowledge is required to interpret model reports. Keeping with the earlier example, in early simulations in which additional CO_2 was introduced into climate models, it was typically introduced all at once: the model would be set into a stable state, the CO_2 concentration would be doubled, and then the simulation run until a new stable state was reached. Obviously, this method for introducing CO_2 is idealized, and this idealization is going to affect different elements of the model in different ways. So, for instance, even if the results of the simulation accurately represent how much a given change in CO_2 will eventually increase average global temperatures, they might misrepresent how quickly that change will come about. In order for an agent to reliably use the resulting model reports as evidence for hypotheses about the world, in other words, they must have sufficient background knowledge of how the model works and how it relates to the true climate system. As in the instrumental case, we can think of this background knowledge being encoded in a model, though now it is a model of the relationship between the model report generated by simulations run on the climate model and the target rather than between the instrumental reading and the target.[9]

Qua data, therefore, model reports and instrumental readings are alike in at least insofar as (a) they are capable of serving as evidence, (b) they require interpretation, and (c) this interpretation is supplied or justified by an understanding of how the data relates to the target of interest. The claim I'm defending in this section goes further, however: it requires not just that a model report is *like* an instrumental reading, but in addition that sets of model reports generated by different models aggregate (or can be aggregated?) in the same way that sets of instrumental readings do. In shorthand, my claim is not just that model-generated data are like instrumentally generated data, but moreover that *ensemble*-generated data—the model reports generated by running simulations on the different models that make up an ensemble—are also like instrumentally generated data, at least insofar as the application of statistics is concerned.

I think a straightforward case for the analogy can be made at the fully general level, though a full defense will require looking more closely at the details. In the context of measurement, we motivated the use of a set of instrumental readings—and accordingly, the incorporation of probability theory into the licensing model—on the grounds that individual readings are influenced by a wide variety of factors that it would be impractical to represent explicitly. Rather than representing them explicitly, scientists use a sample of different readings and "black-box" the influences other than the target of interest.

[8] See Parker (in press) for a more in-depth picture of learning from simulations along these same lines.

[9] Frigg and Nguyen (2016) make a similar point to this in their discussion of interpretive "keys," though our analyses diverge insofar as they are concerned solely with representational questions rather than epistemic ones.

The use of ensembles of models has the same effect.[10] Just as is true of instrumental readings, model reports often track influences other than the true value of the target quantity. So, for example, computational constraints or a desire for generality may lead us to introduce idealizations or approximations into a model, and these idealizations can cause it to misrepresent the target in a variety of ways. Similarly, there may be features of the system that we don't understand very well, and a failure to perfectly represent these aspects of the system can have the same effect. Finally, as Parker and Winsberg (2018) and Schmidt and Sherwood (2015) have stressed, climate modeling cannot be dissociated from particular value choices; to borrow a mundane example from Schmidt and Sherwood (2015), modelers in Australia and England are likely to prioritize accuracy in different parts of the world. In all of these cases, there's some feature of the model that's determined not by the underlying system or our knowledge thereof, but by external factors.

Ideally, we would account for these factors when drawing conclusions from model reports. As in the instrumental case, one means of doing so would be explicitly building them into the model of the relationship between the reports given by the climate model and the target, but this is rarely practical: we're rarely in a position to know how a particular idealization affects the quantities that we're interested in, making it impossible to directly account for its presence in our reasoning. As with instrumentally generated data, therefore, the best way to account for these other influences may be to black-box them: to build multiple models that have different influences (Katzav and Parker, 2015). In principle, building an "ensemble" in this manner allows us to put aside the effects of the idealizations or assumptions that are unique to a particular model and concern ourselves only with those that are shared between all of the members of the ensemble.

This is in effect what climate scientists actually do: they build ensembles of models with different "structural" assumptions and idealizations and then apply statistics to draw inferences from the set of reports given by the ensemble as a whole.[11] It's easier to understand this practice with an example. Recall above that through simulation, a single model will deliver data in the form of values for a variable $T(r, \theta, \psi, t)$, the average temperature in each grid box over time. If we're interested in the change in temperature over a given time period, we can think of the model as providing a three-dimensional picture—a vector field, basically—of the change in temperature in every grid box. When dealing with an ensemble of models, rather than a single vector field, climate scientists work with a set of around ten of these vector fields. The simplest of means of applying statistics in this case involves treating each of these vector fields as though it were sampled from a population of possible representations of the true climate and then using this sample to determine what the true climate is likely to be like. As I've stressed, and as I'll illustrate in the next section, the details of how this works will depend on the choice of statistical model—that is, on the assumptions about

[10] For further discussion of the motivation for the use of ensembles along these same lines, see Parker (2010b; 2010c).

[11] As we'll see, it's often been objected that this model-building strategy is ineffective. My discussion in this paragraph should not be taken to imply that these different model construction processes generate probabilistically independent models.

the relationship between the sample and the true climate—but this broad description will hold generally.

At least at this level of abstraction, both the motivation and the methodology employed here are no different from that of a standard example of statistical inference from data sets composed of instrumental readings: the scientists are essentially taking the individual data points to be the result of a combination of random and non-random processes, with the aim of extracting info about one of the non-random influences. Or, better, they're taking the data points to behave *as though* they were the result of such processes. What's essential to this methodology is justifying the choice of statistical model, where this can once again be thought of as justifying the claim that likelihoods calculated on the basis of the chosen model are close (enough) to the real likelihoods.

To summarize the argument just given, I've claimed that inferences from single model reports are like inferences from single instrumental readings: in both cases, the crucial move in justifying the inference lies in showing that the statistical model accurately represents the relevant probabilistic relationship. How the data points are actually generated is irrelevant once we've fixed the accuracy of the representation. Further, I've argued that the relationship between a single model report and a set of model reports generated by an ensemble is like that between a single instrumental reading and a set of them. In both cases, the data set as a whole introduces variation into the picture, and this variation can be used—at least in principle—to "black-box" various factors that we don't want to model explicitly. Given these two facts, we should expect that drawing inferences from sets of model reports will be like drawing inferences from sets of instrumental readings. The difference in these two cases is *merely* a difference in how the data are generated, which (as we saw in the last section) is not a difference that's relevant to the application of statistics, because statistics is only about capturing the probabilistic relationships that the data-generation process introduces.

The above arguments are very general; nevertheless, if they are right, they show that there's nothing *in principle* objectionable about the practice of applying statistics to ensemble-generated data. There may well be serious practical problems of the same sort that we regularly encounter in everyday statistical practice: choosing the right statistical model is often quite hard. Indeed, a number of philosophers and climate scientists have documented problems with the application of statistics to ensemble-generated data; Carrier and Lenhard nicely summarize these difficulties as follows: "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, pp. 3–4). All of these problems are symptoms of the same fundamental problem, namely that the methods of model construction are very different from random sampling procedures and thus that the data generated by running simulations on extant ensembles do not comprise a representative sample of the relevant space of possibilities.

On the model-based view that I've defended here, the fact that the sample is non-representative does not *alone* impugn our ability to apply statistics to it. As stressed in Sect. 2, all that's needed for the successful application of statistics in this kind

of case is that the statistical model accounts for the non-random character of the sampling procedure. What makes randomized sampling so useful is that it makes model specification easier; when dealing with cases like the present one in which the sample is known to be non-representative, it's harder to justify the choice of statistical model. Indeed, it may be that for various practical reasons, we simply don't have enough information to discriminate between different statistical models in this case. Crucially, however, in this scenario it would not be the non-representative character of the ensemble, but rather our inability to account for the non-representative character, that undermines the application of statistics to ensemble-generated data. As such, even if we're not able to successfully justify statistical inferences from ensemble-generated data in realistic cases, that doesn't threaten my thesis: the model-based view has a clear explanation for how that can happen.

On the model-based view, the problems with extant ensembles are practical problems in that there are (at least in principle) ways of addressing the non-representative character of the sample. From a more design-based perspective, by contrast, the non-representative character of extant ensembles looks much more problematic. Recall that on a design-based view, what justifies statistical inferences are the actual sampling procedures employed. And thus the fact that the actual procedures of model construction are not like random sampling procedures and don't produce a truly representative sample does seem like a plausible reason for thinking that there are a principled reasons why we cannot apply statistics in this case. A number of philosophers have endorsed positions that at least seem to take this view towards ensemble-generated data. Winsberg (2018, p. 98), for example, describes the application of statistics to ensemble-generated data as "conceptually troubled" for essentially the reasons outlined above. And both Betz (2015) and Katzav (2014) have argued for "possibilist" interpretations of ensembles according to which climate models don't represent the world in the right way to justify the application of statistics. If this design-based view is correct, the problem lies not in choosing the statistical model, but in the character of the ensemble itself. And thus, unlike what's true on the model-based view, there's nothing that we can do—even in principle—to account for the non-representative character of the sample.

I've argued that we should prefer the model-based view of statistical inference. As such, we should prefer the former of these two views towards the problems with extant ensembles. Given the abstract presentation of the arguments to this point, however, it's open to those skeptical of the application of statistics to ensemble-generated data to draw the opposite conclusion—namely, to argue that the obvious problems with applying statistics to the data generated by extant ensembles undermines my general arguments regarding the model-based view.

In the next section, therefore, I'll examine a case in which the application of statistics to ensemble-generated data has been at least somewhat successful in generating trustworthy results. The idea here is not that this example is representative in the sense that we should extrapolate the success in this case to other cases.[12] Instead, the case study serves as a kind of proof-of-concept: it illustrates the point that the non-random

[12] Though, as I'll emphasize, the case below is representative in the sense that it is typical for contemporary attribution studies to incorporate the application of statistics to ensemble-generated data.

character of extant ensembles can be addressed in at least some cases by adopting the right statistical model. It thus provides evidence for my central thesis by showing that while the non-representative character of the sample generated by extant ensembles makes successful statistical inference harder, it doesn't render it impossible.

4 Statistics and climate fingerprinting

As of 2013, the Intergovernmental Panel on Climate Change (IPCC) estimated that human-driven changes to greenhouse gas (GHG) concentrations had caused a 0.5 to 1.3°C increase in temperatures since 1951 (IPCC, 2013, p. 869).[13] It is not straightforward to estimate this quantity. The main technique that climate scientists use is called "fingerprinting" (Hegerl and Zweirs, 2011; Parker, 2010a). In this section, I'm going to examine one particular strand of fingerprinting studies: those that involve what are called "errors-in-variables" (EIV) methods and that rely a non-trivial application of statistics to ensemble-generated data. I take this case study to illustrate two things. First, it reinforces the earlier argument by showing how the actual practice of applying statistics to ensemble-generated data fits within a broader model-based understanding of statistical practice. Second, it illustrates that there are, at minimum, some applications of statistics to ensemble-generated data that are more epistemically trustworthy than methods that don't make use of such data.

The central idea behind all fingerprinting studies is that different potential drivers of climate change have different effects on the distribution of temperature changes across the system. Think again of a representation of the global temperature as consisting of a vector in every atmospheric grid box; call the vector field that represents the observations recorded in this way Y.[14] The aim of fingerprinting is to regress Y onto different causal factors that *might* be responsible for a change in temperature. Each of these factors has its own signature or fingerprint, which we'll represent with X terms (e.g., the arbitrary X_i or the specific X_{GHG}). These signatures or fingerprints are themselves vector fields that provide a picture of a change in temperature at each point on the globe—essentially, we can think of X_{GHG} as representing the three-dimensional distribution of the of temperature change that *would* be observed *were* greenhouse gases the only factor causing the temperature to change. Estimating how much these different factors have affected the global mean temperature is then a matter of constructing a regression line:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots$$

Each β term in this equation represents the percent of temperature change due to a specific factor. This is the quantity that scientists use fingerprinting studies to esti-

[13] Since this paper was initially submitted, the IPCC has revised both their estimates and their scale for expressing humanity's contribution to warming; for discussion, see IPCC (in press, chapter 3).

[14] In this context, "observations" should be understood broadly; substantial corrections, interpolations, and transformations are required to render instrumental readings into a form that can actually be used in fingerprint study. For discussion, see Edwards (2010), Lloyd (2012) and Parker (2016), 2020).

mate. In effect, they're decomposing the observed fingerprint into weighted partial fingerprints, with the weights indicating how impactful each factor has actually been.

The description just given is idealized in a number of ways; the actual practice of fingerprinting is much more complicated than this. One simplification is particularly important in the present context. Standard regression techniques rely crucially on the accuracy of the X terms; that is, they don't take into account the possibility of errors in these terms, or what are called "measurement errors" in the statistical literature.[15] *If* one applies standard statistical techniques in a case where there's measurement error—note, that is to say, if one mis-specifies the statistical model in a given way—the resulting estimates for the β terms are liable to be inaccurate. The most well-known problem in this context is what's called regression "dilution" or "attentuation"—essentially, an underestimation of the relevant β terms due to there being "more" variation than the model expects—but, as Carroll et al. (2006, p. 41) note, other kinds of biases can crop up as well. The possibility of measurement errors is important in the context of fingerprinting because the signatures of different causal factors are not given directly by either theory or observation. Instead, they are estimated by means of computer simulations using climate models, and there is often non-trivial amount of uncertainty about the details of different signatures.

One way around around this problem is to adopt what are called "errors-in-variables" (EIV) statistical methods rather than the traditional linear regression method.[16] Roughly speaking, the difference between a standard regression and an EIV method is that where the former treats X terms as known, the latter substitutes a distribution over possible values of X, where this distribution is calculated by applying statistical tools to ensemble-generated data.[17] The technical details of EIV methods can and do differ in important ways from study to study; fortunately, many of these technical details are unnecessary for philosophical purposes. What's important is that EIV methods require the application of statistics to ensemble-generated data.

A more detailed, but still schematic sketch of EIV methodology follows. Recall the earlier characterization of X variables: effectively, each represents a scenario in which the relevant causal factor (e.g., greenhouse gases) accounts for 100% of observed warming. A probability distribution for the true value of a given X variable can be estimated using simulations run on climate models. To carry out this estimation procedure, of course, we need a statistical model. Traditionally, statisticians have employed two different statistical models in the context of measurement errors. In the first, which is used when the errors are "classical," each individual data point behaves as though it is sampled from a distribution centered on X, which means that the mean of the distribution should approximate X. In the second, which is used when the errors are

[15] The classic surveys on this topic in the statistical literature are Carroll (2006) and Fuller (1987); I'll be relying on Carroll (2006) in particular in the coming discussion.

[16] EIV methods were introduced to fingerprinting by Huntingford et al. (2006) and the EIV methods employed in Gillett et al. (2013) served as a basis for the IPCC's estimate of humanity's contribution to climate change in the Fifth Assessment Report (IPCC, 2013, p. 883).

[17] Note that there's a broader sense in which most of methods used in fingerprinting since the work of Allen and Stott (2003) are EIV methods in that they're multi-level or hierarchical models that assume that the X terms are uncertain quantities that must be estimated. Climate scientists usually (but not always) use "EIV" to refer to the more restricted case in which the degree uncertainty is estimated using an ensemble of models, which is the case that we're interested in here.

"Berkson errors" (after Berkson, 1950), by contrast, X behaves as though it were sampled from the same population as the data points. Following statistical tradition and using $\bar{W}$ to indicate the mean of the data points, in a Berkson scenario the probability density function for X is given by the following equation:

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-\bar{W}}{\sigma}\right)^2} dz$$

Standard examples of Berkson errors include cases in which we can't measure the degree to which a number of individuals were exposed to some chemical or drug, and so we use a proxy that gives us the average exposure. It's not the case that everyone was in fact exposed to the average exposure, however, and so the random (from the point of the view of the model) deviations from average exposure must be accounted for.

Intuitively, we might expect that the ensemble-generated data used in attribution studies would behave more like the classical scenario than the Berkson one: since every model aims to capture the truth, we would expect them to be clustered around it. As was discussed above, however, the actual construction procedures of models are not anything like random sampling procedures. As a consequence, empirical validation studies indicate that ensemble-generated data behaves much more like the Berkson scenario (Annan and Hargreaves, 2010; 2011; Sanderson and Knutti, 2012). Climate scientists usually describe this fact by either saying that the truth behaves as though it were sampled from the same population as the models or that the ensemble is "statistically indistinguishable" from the truth, but both of these formulations mean essentially the same thing from a statistical perspective. They're merely different ways of describing the statistical model that yields the most accurate estimates of the likelihood of the model reports on different hypotheses about the true value of X. Once this model is selected, it can be used along with ensemble-generated data to generate a distribution over possible values of X.

Once climate scientists have picked a statistical model to capture the relationship between ensemble-generated data and the X terms, this model is then embedded in the traditional regression model. The resulting resulting EIV model now looks like this:

$$Y = \beta_1(\bar{W}_1 + f_{Z_1}) + \beta_2(\bar{W}_2 + f_{Z_2}) + \ldots$$

where the other variables are the same as before and each f_Z is a (normally distributed) random variable. Fingerprinting is then a matter of adapting standard algorithms for solving regressions (e.g., a least-squares analysis) to account for the additional source of variation—an extremely difficult and important task, but one that won't concern us here.

A schematic picture of the role of ensembles in fingerprinting is sketched in Fig. 1. To summarize: in estimating the contribution of various different factors to warming, climate scientists employ a "errors-in-variables" statistical method to account for the possibility that they've mis-estimated the effect of any one factor. This part of the application is "standard" statistical practice: we're simply complicating our regression techniques to account for measurement error. Ensemble-generated data enter the

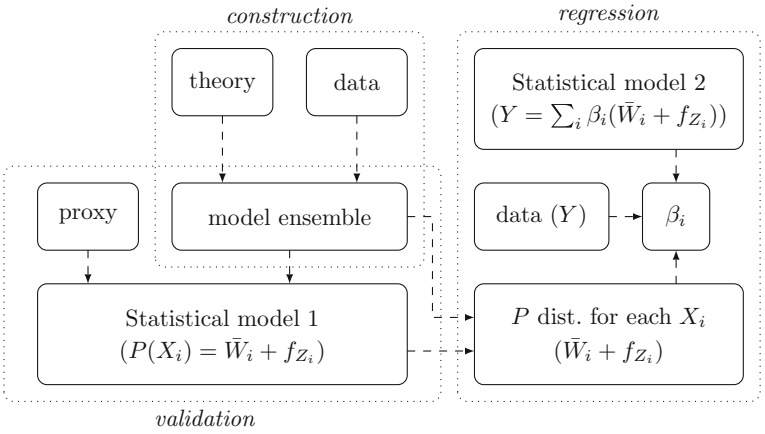

Fig. 1 A schematic diagram of the role of ensembles of climate models in EIV methodology. On this picture, climate models play two important roles. First, the comparison between ensemble-generated data and proxy data is used to build a statistical model of the relationship between the ensemble-generated data and the true signals X_i (this step is labeled "validation"). Second, this statistical model and the ensemble-generated data are used to generate a distribution over possible values of each X_i variable that allows for the calculation of the weights (β_i) in the step labeled "regression"

picture because they are used to estimate a distribution of possible values for the independent variables (the Xs) that the observations are regressed onto. Of course, because estimating this distribution involves the use of statistics, the results are only justified insofar as we can motivate the the choice of statistical model. To identify the right statistical model in this context, climate scientists use proxy data to examine the relationship between ensemble-generated data and X; this proxy data reveals that the truth behaves roughly as though it were sampled from the same distribution as the ensemble-generated data. EIV methods thus depend crucially on the ability of climate scientists to treat ensemble-generated data as like a sample; as we've seen, *part* of this process involves the construction of a statistical model (what's labeled "statistical model 1" in Fig. 1) that relates the ensemble-generated data to the true fingerprint that the date are used to estimate.

The above is intended to illustrate that the application of statistics to the data generated by ensembles is, at the level of methodology, just like the application of statistics to experimental data. What's crucial, in both cases, is a model of the probabilistic relationship between the data and the target; this model is chosen of the basis of our background knowledge, which in this case (as is often true) is informed by empirical comparisons with proxies.

There's another question here, which is whether this practice is in fact successful. It's hard to answer this question directly, of course: to know whether the estimates generated by EIV methods are more accurate than other methods, we would need to know the true contribution of humans to climate change.[18] The best that can be achieved are comparisons against proxy data that has known properties. This comparison has been

[18] This problem is simply the realization of a common problem in measurement; see Bokulich (2020) for a discussion of the same problem in the context of carbon dating, for instance.

carried out by Hannart et al. (2014), indicating that their version of the EIV method is generally more accurate than methods that don't employ ensemble-generated data and that in some scenarios it is substantially more accurate. It isn't surprising that this would be the case. To carry out *any* fingerprinting study, climate scientists need *some* estimate of the value of the X terms. The basic regression approach simply assumes that the best estimate of X is perfectly accurate. The EIV method, by contrast, uses ensemble-generated data to construct a distribution over possible values of X. Even if we can't expect the EIV methods to be perfectly reliable, therefore, we should expect them to *more* reliable than the standard regression method: insofar as the distribution used in the EIV method is closer to the appropriate or true distribution than one that assigns all of the probability to the mean of the distribution, the EIV method should generate more accurate results.

In this section, I've examined one application of statistics to ensemble-generated data. In the next section, I'll further discuss the implications of this case for the general argument of the paper, but let me end this section by reiterating the role of this example in the argument. The point here is not that the use of ensemble-generated data in EIV methods is particularly representative of all applications of ensemble-generated data in climate science. Instead, the example is a proof-of-concept: it shows that climate scientists can productively apply statistics to ensemble-generated data so long as they are sufficiently careful to choose the right statistical models for the job. I note that while the particular EIV methodology may not be representative of the use of ensemble-generated data in attribution studies let alone in climate science broadly speaking, the reliance of EIV methods on the application of statistics to ensemble-generated data is certainly not unique. More recent studies—such as Gillett et al. (2021) and Ribes et al. (2021), both of which play an important role in the Sixth Assessment Report (IPCC, in press)—often forego EIV methods but employ ensemble-generated data in other, philosophically similar, ways.

5 Climate fingerprinting and statistical inference

In Sect. 3, I noted that there are a number of problems with the application of statistics to ensemble-generated data and that one might read these problems as undermining the paper's thesis, namely that the application of statistics to ensemble-generated data is just like any other application of statistics. If these problems are unique to the modeling context or somehow insuperable, then the application of statistics to ensemble-generated data is not at all like any other application of statistics. If, however, these problems are normal statistical problems that can be resolved in some cases, then there's nothing conceptually objectionable about the application of statistics to ensemble-generated data. The case study of the last section is meant to serve as a kind of proof-of-concept of my view: while there are (serious) problems with the use of statistics in the context of EIV methods, these problems are (a) normal and (b) don't ultimately render the EIV method useless or misleading. This final section returns to the criticisms leveled against the application of statistics to ensemble-generated data and discusses the limits of the case study and the arguments offered here.

Beginning with the criticisms. As I've stressed throughout this paper, the proper application of statistics requires choosing a statistical model that adequately represents the relationship between the data and the target of the inference. This is true regardless of how the data are generated. In easy cases, we're able to assume that the data-generation process approximates a genuinely random sampling procedure. As we saw above, that isn't the case in the context of ensemble-generated data. Not only does the actual practice of model construction give us no reason for thinking that model reports will be normally distributed around the truth (Winsberg, 2018, p. 97), empirical validation studies indicate the results of simulations carried out on different models don't behave like independent draws and reports are more narrowly distributed around the target values than an ideal sample would be (see Annan and Hargreaves, 2011; Knutti et al., 2010).

The upshot is that samples comprised of ensemble-generated data are bad in the technical sense that they are non-representative. Two points, however. First, samples that are non-representative are commonplace throughout the sciences—non-representative samples are a very normal problem for scientists to encounter in cases where it is difficult (or unethical) to exert substantial control over the data-generation process. The reasons why it's difficult to control the data-generation process are different in the case of climate model ensembles than they are in (say) macroeconomics, but the effect is the same in the sense that scientists are faced with samples that make statistical inference more difficult.

Second, that a sample is non-representative does not guarantee that statistics cannot be applied to it in a productive manner. Again, what's essential is that we're able to justify the choice of the statistical model; so long as we can capture the non-representative character of the sample in our model, we can justify the relevant statistical inferences. It's simply not the case that samples like those generated by extant ensembles cannot license statistical conclusions *in principle*. Indeed, as we saw above, samples with similar statistical properties to those generated by extant ensembles have been recognized in the measurement error literature since the 1950s, and there are well-known techniques for drawing reliable statistical inferences from them. The non-representative character of ensemble-generated data makes justifying the choice of statistical model harder, but this is hardly unique to ensemble-generated data. In practice, model specification is often a difficult and complicated process that relies on a combination of background knowledge, curve-fitting techniques, empirical calibration, and validation checks (Spanos, 2006); it's only in the easiest cases that the data comes out naturally in a normal distribution.

As the case study indicates, the non-representative character of ensemble-generated data can be addressed in at least some cases; the EIV method discussed above is ultimately more accurate than a standard regression method *because* it incorporates the application of statistics to ensemble-generated data. It's another question entirely whether the success of this case can be extrapolated to other applications of statistics to ensemble-generated data. There are at least two ways in which the case study discussed above may not be representative. First, it's widely recognized that any given climate model is better at answering some questions than others, and the same is true for ensembles. It's plausible that one of the questions that ensembles might be relatively good at answering is the one posed in EIV studies—that is, climate ensembles might

be particularly good at estimating what climate change to date would look like if it was driven entirely by (e.g.) CO_2. Or it may be that the particular role played by the probability distributions in this case makes it so that the final results are relatively insensitive to small errors in the characterization of these distributions. In either case, there would be some reason why ensemble-generated data are better suited to be used in EIV methods than other contexts. We thus can't treat the success of the application of statistics to ensemble-generated data in this case as a guarantee that the similar applications will be equally successful in helping us answer other questions.

The second way in which the EIV case may be non-representative concerns the relationship between the target being estimated and the proxies that are used to justify the choice of statistical model. As Parker has emphasized in her discussion of the use of ensemble methods in predicting the future climate (see, in particular, Parker, 2010b, p. 269), comparison with present-day proxies is in no way guaranteed to lead us to the right statistical model for the estimation of *future* climate variables. For this kind of empirical validation to be successful, we would need to be able to extrapolate the statistical model from the context of the present-day climate to the future, an extrapolation that is particularly problematic in the context of the climate science, where we expect that the future will be importantly different from the present. As the climate scientists involved in validating the choice of statistical model explicitly recognize, future temperatures may well have a very different probabilistic relationship to model reports than present-day temperatures do (Annan and Hargreaves, 2011, p. 4). As a consequence, it may be that the application of statistics to ensemble-generated data concerning future climate change is quite distinct from the case study discussed in the last section. At the very least, it is epistemically riskier in that the target in the EIV case is more similar to the proxy targets than the future climate variables are expected to be.

The above are reasons why we should not assume, on the basis of the case study outlined above, that it will be easy or even feasible to productively apply statistics to ensemble-generated data in any climate context. Once again, however, it's important to stress that the difficulties that would prevent the productive application of statistics in these cases are not unique to either climate science or the application of statistics to ensemble-generated data. Indeed, the problems with extrapolating the success of a model from one domain to another are well-appreciated in philosophy of science, and particularly well-appreciated in domains such as economics and psychology where there are persistent problems with extrapolating results found in laboratory settings to the "real world" (see, e.g., Steel, 2008). Further, the particular extrapolation problem found in this case is essentially no different from the extrapolation problem found in projecting the results of any single climate model into the future. In both cases, the problem is simply that we expect the future climate to be different from the past climate. Regardless of whether we're employing a single model or applying statistics to ensemble-generated data, therefore, the possibility of differences between the past climate and the future climate introduces an additional source of uncertainty.

In this section, I've argued that that while the application of the statistics to ensemble-generated data is non-trivial, the difficulties faced in this application are familiar from other applications of statistics and there are—at least in principle—methods for addressing these problems. How successful these methods are will vary

from case to case, but as the last section illustrated they are at least in some cases successful in rendering the application of statistics to ensemble-generated data productive and reliable—or at least more productive and reliable than the alternatives.

The upshot of this discussion has been to reinforce the central thesis of the paper: to justify statistical inferences—whether from instrumentally generated or ensemble-generated data—what's essential is justifying the choice of statistical model. Close attention to actual applications of statistics to ensemble-generated data reveals only normal difficulties associated with model specification in cases where we can't exert substantial control over the data-generation process. The resulting problems are real, and may well undermine certain applications of statistics to ensemble-generated data, but don't constitute anything like the kind of in-principle problem that would pose trouble for my thesis.

6 Conclusion

The title of this paper is a question: when is an ensemble like a sample? The answer to this question that I've defended in this paper is that an ensemble is like a sample—or, better, the data generated by an ensemble compose a sample—when we know enough about the relationship between the data and the target of interest to treat it as one. On the picture of scientific inference that I've offered, what justifies scientific inferences generally speaking is our understanding of the probabilistic relationship between the data that the inference is based on and the target of the inferential process. This understanding is captured in a statistical model. All that the application of statistics to ensemble-generated data requires, therefore, is that we're able to chose the right model of the probabilistic relationship between the results generated by the ensemble and the target of the inference. This is often a difficult task, particularly in contexts in which there are no great proxy targets to validate the statistical model against. As a result, the application of statistics to ensemble-generated data often faces a number of practical problems. But the practical problems are the familiar difficulties found in any application of statistics.

Acknowledgements I would like to thank Anjan Chakravartty, Wendy Parker, Ryan O'Loughlin, and three anonymous reviewers, all of whom gave extremely helpful comments on an earlier version of this paper; this version is better, and likely unrecognizable, thanks to their contributions. Thanks also to Nathan Gillett, Auélion Ribes, and the audience at the 2020 Central APA.

Funding Open Access funding enabled and organized by Projekt DEAL. Funding for this paper was provided by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project 254954344/GRK2073

References

Allen, M. R., & Stott, P. A. (2003). Estimating signal amplitudes in optimal fingerprinting, part i: Theory. *Climate Dynamics, 21*, 477–491.

Annan, J. D., & Hargreaves, J. C. (2010). Reliability of the CMIP3 ensemble. *Geophysical Research Letters, 37*, 1–5.

Annan, J. D., & Hargreaves, J. C. (2011). Understanding the CMIP3 model ensemble. *Journal of Climate, 24*, 4529–4538.

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association, 45*, 164–180.

Betz, G. (2015). Are climate models credible worlds? Prospects and limitations of possibilistic climate prediction. *European Journal for Philosophy of Science, 5*, 191–215.

Bokulich, A. (2020). Calibration, coherence, and consilience in radiometric measures of geologic time. *Philosophy of Science, 87*, 425–456.

Carrier, M., & Lenhard, J. (2019). Climate models: How to assess their reliability. *International Studies in the Philosophy of Science, 32*, 81–100.

Carroll, R. J., et al. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Cox, D. R., & Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation*. London: Chapman & Hall.

Dethier, C. (2021). How to do things with theory: The instrumental role of auxiliary hypotheses in testing. *Erkenntnis*, *81*, 1453–1468.

Edwards, P. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge: MIT Press.

Eyring, V., et al. (2016). Overview of the Coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. *Geoscience Model Development, 9*, 1937–1958.

Frigg, R., & Nguyen, J. (2016). The fiction view of models reloaded. *The Monist, 99*, 225–248.

Fuller, W. A. (1987). *Measurement error models*. New York: John Wiley & Sons.

Gettelman, A., & Rood, R. B. (2016). *Demystifying climate models: A users guide to earth system models*. Berlin: Springer.

Gillett, N. P., et al. (2013). Constraining the ratio of global warming to cumulative CO2 emissions using CMIP5 simulations. *Journal of Climate, 26*, 6844–6858.

Gillett, N. P., et al. (2021). Constraining human contributions to observed warming since the pre-industrial period. *Nature Climate Change, 11*, 207–212.

Hannart, A., Ribes, A., & Naveau, P. (2014). Optimal fingerprinting under multiple sources of uncertainty. *Geophysical Research Letters, 41*, 1261–1268.

Hegerl, G. C., & Zweirs, F. (2011). Use of models in detection and attribution of climate change. *Wiley Interdisciplinary Reviews: Climate Change, 2*, 570–591.

Huntingford, C., et al. (2006). Incorporating model uncertainty into attribution of observed temperature change. *Geophysical Research Letters, 33*, 1–4.

IPCC (2013). Climate change 2013: The physical science basis. In T. F. Stocker et al. (Ed.), *Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge: Cambridge University Press.

IPCC (in press). Climate change 2021: The physical science basis. In V. Masson-Delmotte et al. (Ed.), *Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge: Cambridge University Press.

Kass, R. E. (2011). Statistical inference: The big picture. *Statistical Science, 26*, 1–9.

Katzav, J. (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*, 228–238.

Katzav, J., & Parker, W. S. (2015). The future of climate modeling. *Climatic Change, 132*, 475–487.

Knutti, R., et al. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate, 25*, 2739–2758.

Lloyd, E. (2012). The role of complex empiricism in the debates about satellite data and climate models. *Studies in History and Philosophy of Science Part A, 43*, 390–401.

Lusk, G. (2016). Computer simulation and the features of novel empirical data. *Studies in History and Philosophy of Science Part A, 56*, 145–152.

Mäki, U. (2005). Models are experiments, experiments are models. *Journal of Economic Methodology, 12*, 303–315.

McGuffie, K., & Henderson-Sellers, A. (2014). *The climate modeling primer* (4th ed.). Chichester: Wiley Blackwell.
Morgan, M. S. (2002). Model experiments and models in experiments. In L. Magnani & N. J. Nersessian (Eds.), *Model-based reasoning: Science, technology, values* (pp. 41–58). Dordrecht: Kluwer.
Morrison, M. (2015). *Reconstructing reality: Models, mathematics, and simulations*. Oxford: Oxford University Press.
Parker, W. S. (2009). Does matter really matter? Computer simulations, experiments, and materiality. *Synthese, 169*, 483–496.
Parker, W. S. (2010). Comparative process tracing and climate change fingerprints. *Philosophy of Science, 77*, 1083–1095.
Parker, W. S. (2010). Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in the History and Philosophy of Modern Physics, 41*, 263–272.
Parker, W. S. (2010). Whose probabilities? Predicting climate change with ensembles of models. *Philosophy of Science, 77*, 985–997.
Parker, W. S. (2013). Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change, 4*, 213–223.
Parker, W. S. (2016). Reanalyses and observations: What's the difference? *Bulletin of the American Meteorological Society, 97*, 1565–1572.
Parker, W. S. (2020). Evaluating data journeys: Climategate, synthetic data and the benchmarking of methods for climate data processing. In S. Leonelli & N. Tempini (Eds.), *Data journeys in the sciences* (pp. 191–206). Cham: Springer.
Parker, W. S. (in press). Evidence and knowledge from computer simulation. *Erkenntnis*.
Parker, W. S., & Risbey, J. S. (2015). False precision, surprise and improved uncertainty assessment. *Philosophical Transactions of the Royal Society Part A, 373*, 20140453.
Parker, W. S., & Winsberg, E. (2018). Values and evidence: How models make a difference. *European Journal for Philosophy of Science, 8*, 125–142.
Ribes, A., Qasmi, S., & Gillett, N. P. (2021). Making climate projections conditional on historical observations. *Science Advances, 7*, 1–9.
Royall, R. (1992). The model based (prediction) approach to finite population sampling theory. *Lecture Notes-Monograph Series, 17*, 225–240.
Sanderson, B. M., & Knutti, R. (2012). On the interpretation of constrained climate model ensembles. *Geophysical Research Letters, 39*, 1–6.
Schmidt, G. A., & Sherwood, S. C. (2015). A practical philosophy of complex climate modelling. *European Journal for Philosophy of Science, 5*, 149–169.
Smith, T. M. F. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society Series A, 146*, 394–403.
Spanos, A. (2006). Where do statistical models come from? Revisiting the problem of specification. *Lecture Notes-Monograph Series, 49*, 98–119.
Stainforth, D. A., et al. (2007). Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transactions of the Royal Society Series A, 365*, 2145–2161.
Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.
Tal, E. (2020). Measurement in science. In E. N. Zalta (Ed.) *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/measurement-science/.
Winsberg, E. (2010). *Science in the age of computer simulation*. Chicago, IL: University of Chicago Press.
Winsberg, E. (2018). *Philosophy and climate science*. Cambridge: Cambridge University Press.
Zhao, K. (2021). Sample representation in the social sciences. *Synthese*, *198*, 9097–9115.