Calibrating Statistical Tools: Improving the Measure of Humanity's Influence on the Climate

Corey Dethier

[[Penultimate version. Please see the final version at https://doi.org/10.1016/j.shpsa.2022.06.010.]]

Abstract

Over the last twenty-five years, climate scientists working on the attribution of climate change to humans have developed increasingly sophisticated statistical models in a process that can be understood as a kind of calibration: the gradual changes to the statistical models employed in attribution studies served as iterative revisions to a measurement(-like) procedure motivated primarily by the aim of neutralizing particularly troublesome sources of error or uncertainty. This practice is in keeping with recent work on the evaluation of models more generally that views models as tools for particular tasks: what drives the process is the desire for models that provide more reliable grounds for inference rather than accuracy to the underlying mechanisms of data-generation.

Keywords: statistics · calibration · attribution · measurement · climate change

0 Introduction

One of the major projects of climate science is what's called "attribution": determining how factors such as increasing greenhouse gas concentrations have contributed

to global warming. Attribution involves what looks like a kind of measurement.[1] Interventions on the environment using physical instruments—i.e., thermometers—generate instrumental readings, which are collected into data sets, and then measurement outcomes—the actual estimate for the contribution of a particular factor—are inferred from these data sets. This inference relies heavily on both background knowledge (in the form of climate models) and statistics, the latter of which plays a particularly important role in attribution studies due to the high levels of noise present in climate data.

Though the basic scaffolding of this measurement-like procedure has remained largely unchanged over the last thirty years, the details have not. Indeed, there have been changes to all of the elements outlined above: climate scientists have gathered more data, built more complex and accurate models, and developed new statistical techniques. The first two of these changes should be familiar—gathering more data and building more accurate theoretical constructs are (relatively) well understood parts of the scientific process. The third is less so. How should we understand changes in statistical techniques? What makes one statistical technique "better" than another in the context of climate science? And how does the general desire for better statistical techniques get translated into specific changes?

The present paper tackles these questions. I argue that the changes to statistical techniques found in attribution studies are akin to what's called "calibration" in philosophy of measurement, where an instrument or model employed in a measurement procedure is altered with the goal of producing measurement outcomes that are more precise, more accurate, or both.

In more detail, I argue for two conclusions. First, we should understand the changes to statistical techniques as alterations to an "instrument"—what statisticians call the "statistical model"—similar to paradigm cases of calibration found in discussions such as Bokulich (2020a) and Tal (2017). Though my examples involve changes to the inferential tools involved in measurement rather than the more familiar changes to physical instruments, I show that they largely have the same motivation and epistemological implications. Second, statistical models should be judged according to what I'll call an "artefactual" view (borrowing the term from Knuuttila 2011) according to which the new model is better than the old insofar as it is more reliable in licensing accurate inferences (rather than, say, insofar as it accurately represents any real process in the world). This artefactual view both aligns nicely with previous work on model evaluation generally and helps explain many of

[1]Following Parker (2017, 2020a), I'm going to put aside the question of whether attribution is "really" measurement in some deep sense; my interest here lies what's revealed by examining attribution studies in climate science through the lens of the philosophy of measurement.

the details of the calibration process.

In outline, the paper begins with some basic background on attribution and the technical difficulties that arise in the science (§1). I then detail how climate scientists have refined the statistical models employed in attribution, and how these refinements have both alleviated the original difficulties and introduced further complications of their own (§§2,3). The penultimate section argues that we should understand these cases in terms of calibration, or at least that the changes in question are *akin* to paradigm cases of calibration in important respects (§4). The final one argues that the proper way to evaluate statistical models—and thus the success of the calibration process—is through an "artefactual" lens (§5).

1 Before attribution: the statistics of detection

This section provides a brief conceptual and technical background for discussing the use of statistics in attribution. Throughout, I endeavor to present the relatively advanced statistical concepts in as close to plain English as possible. As we'll see, however, much of what is philosophically interesting about this case is to be found in the details, and so a certain amount of engagement with the technicalities is unavoidable.

It will be helpful to begin by saying a little bit about measurements. As is often the case in philosophy, there is no entirely uncontroversial definition of measurement (Tal 2020). In what follows, I'm going to adopt a relatively loose interpretation inspired by recent "model-based" accounts: say that a procedure is a measurement (or is "measurement-like") when it involves assigning values to parameters or quantities of interest on the basis of interactions between physical instruments and the system or target to be represented (compare Tal 2020, §7). What makes this view "model-based" is the recognition that the assignment of values cannot be conducted without a (sometimes tacit) model of the interaction(s) in question. That is, a scientist is only justified in drawing inferences from instrumental readings (e.g., the height of a column of fluid in a thermometer) to the value of a quantity of interest (e.g., the temperature) given background assumptions concerning the accuracy and precision of the thermometer.

In climate science, "attribution" refers to the science of determining the causes of various climate changes and events. Many attribution studies take on a measurement-like form.[2] Climate scientists gather instrumental readings of quantities such as tem-

[2]To date, attribution has received less attention than prediction in the philosophical literature, but see Parker (2010) for an overview or Dethier (2022) for a discussion of attribution as a case

perature, precipitation, humidity, etc. taken from around the globe.[3] A measurement outcome, usually an estimate for the contribution of a particular causal factor such as greenhouse gases or the combined anthropogenic influence, is then estimated on the basis of this data and information from climate models. In the cases that we'll be looking at below, for instance, the typical attribution study will conclude by providing an °C estimate for the total contribution of humans to global warming trends over a time period such as 1951 to 2010.

More technically, let $\mathbf{O}$ represent the observed climate data. We can think of $\mathbf{O}$ as a matrix in which each column is a climate variable $O_1(t)$ ... $O_n(t)$ representing (e.g.) the change in temperature at a particular measurement location or device over time. We suppose that to a first approximation this data can be decomposed linearly into a response due to "forcings"—that is, external causes such as an increase in greenhouse gas concentrations—represented by $\mathbf{O}_S$ (for "signal") and the "internal variation" of the system ($\mathbf{O}_N$, for "noise") like so:

$$\mathbf{O} = \mathbf{O}_S + \mathbf{O}_N \tag{1}$$

Intuitively, attribution studies aim to isolate $\mathbf{O}_S$ and to determine which of the various potential hypotheses concerning the origins of climate change are compatible with it; the more limited "detection" studies are limited to simply determining whether $\mathbf{O}_S$ is non-zero.

Equation (1) provides the bare-bones structure of what statisticians call a "statistical model." Statistical models aim to represent the relationship between the data (in this case, $\mathbf{O}$) and the target of the inference ($\mathbf{O}_S$).[4] A statistical model is essential to any inferential use of statistics: though the statistical models of Bayesian and classical statistics differ in various ways, neither can get off the ground without a statistical model of some sort. At minimum, we need some representation of the probabilistic relationship between the data and various hypotheses (i.e., the likelihood ratio).[5]

---

study in the use of computer simulations.

[3]Note that these readings are often already heavily processed before the study even begins; for discussion, see Edwards (2010), Lloyd (2012), and Parker (2016).

[4]In this respect, the category combines the traditional philosophical distinction between "models of data" and "models of experiment" (Suppes 1962). See also Mayo (1996, 128–41) and Spanos (2006).

[5]Formally, we can think of this statistical model as consisting of at least following three components: (1) a hypothesis space, (2) a sample space that characterizes the nature of the possible evidence, and (3) a probability density function that gives the likelihood of elements of the latter conditional on the former. See Sprenger (2019) for extended discussion of the third requirement.

Statisticians will be the first to tell you that statistical models are often idealized: the famous quip that "all models are wrong, but some are useful" (Box 1979) is originally owed to a statistician. The statistical models employed in early detection studies were *heavily* idealized. Essentially, we can think of the relevant statistical models as consisting of two assumptions about our data set, $\mathbf{O}$. First, the assumption that the distribution of a change in temperature throughout the world is effectively random, meaning that a change in temperature should show up primarily in the mean global temperature, which we can call $\bar{\mathbf{O}}$. Where $\mathbf{O}$ is an n-dimensional matrix, $\bar{\mathbf{O}}$ is a single column; for each time, we've averaged the O_i variables, yielding a much simpler statistic. Second, the assumption that the behavior of the global mean temperature can be decomposed into the effects of warming (if they exist) and random noise, or:

$$\bar{\mathbf{O}} = \mathbf{O}_S + v \tag{2}$$

where

$$v \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

which says that the distribution of the global mean temperature over time will behave like a normal distribution centered on the function $\mathbf{O}_S$ with standard deviation σ.[6]

While this statistical model is heavily idealized—we've known since before the first high-quality detection studies were conducted that changes in temperature cannot be expected to be randomly distributed—introducing these idealizations massively simplifies the problem of trying to detect climate change. The first assumption justifies reducing our extremely complicated data set to a single "detector" statistic, namely the global mean temperature $\bar{\mathbf{O}}$; the second warrants the use of standard χ^2 tests to determine whether $\bar{\mathbf{O}}$ is consistent with the null hypothesis that there has been no externally forced change in climate. And though the resulting tests rely on an idealized statistical model, they are arguably reliable enough for the purposes of detecting the existence of climate change.

Unfortunately, detection is not attribution: that there is *some* trend doesn't tell us whether that trend is attributable to (say) increases in greenhouse gas concentrations. The problem is that different potential explanations of warming have largely the same implications for mean surface temperatures, making it difficult to distinguish between different hypotheses using only this data. By contrast, these hypotheses differ more dramatically in their implications for how changes in temperature are to be distributed both across regions and through time. Roughly speaking, for instance,

[6] Of course, actual detection studies often employed various checks on these assumptions. See Santer et al. (1996) for an overview.

an increase in the sun's energy output will have a larger warming effect on the upper atmosphere, while an increase in CO_2 concentrations will trap heat lower down. To distinguish between these two hypotheses, therefore, the averaging approach surveyed in this section is a hindrance: it would be helpful to find a way of organizing the data that allowed for more detailed comparisons. In what follows, I'll outline how the statistical model used in detection studies has been successively refined in two major dimensions to allow for the proper attribution of climate change to humans.

2 From averaging to regularization

The first major innovation that allowed for the development of attribution studies is a method for generating alternative detector statistics that carry more information concerning the different potential explanations of climate change than the mean global temperature does. This method is commonly called "fingerprinting" and was introduced to climate science by a series of theoretical papers in the 80s and early 90s.[7]

It's helpful to begin by describing what we want. So call our eventual detector statistic $\mathbf{D}$, and suppose that it too can be decomposed into a signal component $\mathbf{D}_S$ and a noise component $\mathbf{D}_N$. What we're aiming for is a statistic with the maximal ratio of signal to noise—essentially, we want to isolate that part of the data where the signal is most likely to be visible if and only if it exists. Letting $\bar{\mathbf{D}}_N$ indicate the average value of $\mathbf{D}_N$, this intuitive idea can be expressed in terms of maximizing the ratio between $\mathbf{D}_S$ and $\bar{\mathbf{D}}_N$.

To find the appropriate detector statistic, we need information about the character of the expected signal and noise patterns; to identify where the signal is likely to be, we need to know what it looks like. This information is usually provided by climate models. More precisely, climate models provide two crucial pieces of information. First, the estimated signal, which I'll represent by $\mathbf{M}_S$ to indicate that it is estimated using climate models. This vector is used as a "weight" on the data: the more an element is in the expected direction of the signal, the more we want our detector statistic to focus on it. Second, the estimated internal variation of the system, given by a covariance matrix $\mathbf{\Sigma_M}$ (a covariance matrix is just the n-dimensional generalization of the variance term σ^2). Unlike what's true regarding the direction of signal, we want to focus as little as possible on areas with high (expected) variation, meaning that the data is weighted by the inverse of the covariance matrix ($\mathbf{\Sigma_M^{-1}}$) so

[7] In what follows I rely in particular on the Nobel prize-winning work of Hasselmann (1993, 1997). For an introduction to the relevant statistical principles, see von Storch and Zwiers (1999).

as to de-emphasize those areas in which variance is high.

If the climate models are accurate in their representations of the signal and the noise, applying these weights to the data will yield the best detector statistic for testing the existence and strength of the signal. More precisely, if we define the detector statistic $\mathbf{D}$ by applying the weights to the data like so

$$\mathbf{D} = \mathbf{\Sigma}_{\mathbf{M}}^{-1}\mathbf{M}_S\mathbf{O} \tag{4}$$

the result should maximize the signal-to-noise ratio.[8] We can then use $\mathbf{D}$ in much the same way that we employed the global mean temperature above, namely by applying relatively simple statistical tests to determine whether or not it is compatible with different hypotheses about the origins of climate change, including the null hypothesis that $\mathbf{D}$ is entirely accounted for by random variation.

For present purposes, $\mathbf{D}$ can be thought of as a replacement for $\bar{\mathbf{O}}$; essentially, the central move behind fingerprinting is a partial de-idealization of the statistical model employed in detection studies. There, the statistical assumption was that any change in temperature would be randomly distributed throughout the system and so the relevant "place" to look for changes in temperature was the global average. Climate scientists replaced this assumption by using climate models to determine where the change in temperature will show up (if it exists). The result is a detector statistic that should (and in fact does) allow for a more accurate picture of changes in the climate than the average itself does.

Of course, as is often true of de-idealizations, the move to this more powerful statistical model comes with its own complications. Most importantly, the detector statistic $\mathbf{D}$ is generated using the assumption that $\mathbf{M}_S$ and $\mathbf{\Sigma}_{\mathbf{M}}$ are perfectly accurate representations of the signal and noise. This is obviously unrealistic, as Hasselmann (1993) makes clear in his foundational paper laying out the technique. So while $\mathbf{D}$ represents a substantial improvement on the basic averaging technique, it's far from perfect, and the subsequent two decades have seen a number of attempts to develop practical means of generating improved detector statistics.

One of the proposed improvements has been widely adopted in the subsequent literature. Recall: $\mathbf{\Sigma}_{\mathbf{M}}$ is an estimate of the true covariance matrix, and plays a crucial role in picking out the detector statistic—more precisely, its *inverse* ($\mathbf{\Sigma}_{\mathbf{M}}^{-1}$) is used to decrease the weight on elements of the data where variance is expected to be high. The problem here is that the true covariance matrix is extremely large, meaning that to generate an accurate estimate of it, large quantities of data are

[8] Usually, the relevant matrics are defined in such a way that they have to be transposed before being used as weights. This complication makes little difference to the discussion, however, and so I've forgone it in favor of terminological and conceptual simplicity.

necessary. Without large quantities of data, the extreme elements of the covariance matrix—the areas of highest and lowest variance—are liable to be misrepresented by $\mathbf{\Sigma_M}$. This misrepresentation of the extreme elements leads the covariance matrix to be "ill-conditioned," meaning essentially that taking the inverse dramatically multiplies the degree of misrepresentation.[9] The problem here is quite significant: Ribes, Planton, and Terray (2013, 2821) estimate that at minimum 1 to 2 *orders of magnitude* more data points than are normally available would be required to generate a "well-conditioned" estimate of even a reduced version of the matrix.

In early approaches, this problem was addressed by truncation: rather than attempt to use the entire covariance matrix, the k most significant elements of it were chosen instead. There are two main problems with employing truncation in this manner. First, it weakens the test: truncation requires throwing away part of the signal because it may be misleading when inverted. It would be preferable to strip away or control for the effects of inversion in a manner that didn't require throwing away the data, or at least required throwing away much less of it. Second, the results are often highly sensitive to the choice of k, which is at least somewhat arbitrary. In principle, the bigger k is the better; truncating as little as possible means throwing away as little of the data as possible. On the other hand, the bigger k is the more likely the resulting matrix is to introduce serious and misleading errors when inverted. Climate scientists developed various techniques for estimating the ideal truncation level in a given data set, but these were considered imperfect and many of the bigger studies ran tests at multiple truncation levels (see, e.g., Gillett et al. 2013).

To resolve these problems, Ribes, Azaïs, and Planton (2009) introduced a method that they term "regularization." Technically, their method can be understood in either classical or Bayesian terms, but the conceptual insight is clearer from a Bayesian perspective. Effectively, rather than truncating $\mathbf{\Sigma_M}$, they use $\mathbf{\Sigma_M}$ as evidence for the character of the true matrix; more precisely, they introduce a "prior" covariance matrix—in practice, the prior that is used is the objective / (relatively) uninformative identity matrix $\mathbf{I}$—and then generate a posterior estimate for the true covariance matrix via a weighted updating procedure on $\mathbf{\Sigma_M}$. The result is usually expressed as

$$\mathbf{\Sigma_I} = \gamma\mathbf{\Sigma_M} + \rho\mathbf{I} \tag{5}$$

where γ and ρ are weights that are set to minimize the expected error in $\mathbf{\Sigma_I}^{-1}$. $\mathbf{\Sigma_I}^{-1}$ then replaces $\mathbf{\Sigma_M}^{-1}$ in the calculation of $\mathbf{D}$.

Note that the result is still not a perfect representation of the true internal variation, let alone a perfect representation of the process by which $\mathbf{O}$ is actually gen-

[9] Though in fact there's often not enough data to render it invertible at all, particularly without substantial pre-processing.

erated. The regularization procedure still requires us to treat the estimate—now $\mathbf{\Sigma_I}$ rather than $\mathbf{\Sigma_M}$—as though it were the real covariance matrix rather than an estimate (that is, uncertainty about the accuracy of $\mathbf{\Sigma_I}$ is not built into the statistical test), and the new method does nothing to eliminate any actual sampling errors. But regularization does allow climate scientists to make use of the information provided by $\mathbf{\Sigma_M}$ while worrying *less* about misestimating various statistical properties due the mathematical instability of the inversion: because $\mathbf{I}$ is (nearly) uninformative but well-behaved under inversion, using it as a prior and updating on $\mathbf{\Sigma_M}$ yields a result that contains (most of) the information included in $\mathbf{\Sigma_M}$ without the features that amplify inaccuracies when inverted.

So, to summarize: the major innovation that allowed climate scientists to not just detect climate change but to test more complex hypotheses about its origins was the introduction of fingerprinting techniques to generate a better model of the relationship between climate change and the observed data—that is, to generate a better statistical model. Further, as just emphasized, this central innovation did not fix the procedure of estimating human contribution to climate change once and for all. On the contrary, climate scientists continued to introduce refinements to the fingerprinting technique with the aim of developing a statistical model that could be used in more reliable and informative tests.

3 From consistency to regression

Where the first major innovation concerns the generation of the detector statistics, the second concerns the relationship between the detector statistic and various hypotheses about the cause of climate change. Recall the basic statistical model used in detection, in which the average global temperature is assumed to result from the combination of the true change in the climate and random noise. This model only allows for consistency checks: all that we can test is whether the observed data is compatible with different hypotheses concerning the true change in temperature. In practice, the most such studies were able to offer were qualitative determinations that humans are responsible for climate change to some degree; they didn't allow for anything like reliable quantitative estimates of how responsible we are.

Soon after the introduction of fingerprinting, it was recognized that the increase in information provided by the change in the detector statistic allowed for more sophisticated and thus more informative tests. In particular, as emphasized by Allen and Tett (1999) and Levine and Berliner (1999), we can use $\mathbf{D}$ not just to check the consistency of the data with this or that hypothesis about the origins of climate change, but also to run regressions to find the coefficients of best fit. That is, having

selected a detector statistic $\mathbf{D}$, we set that detector statistic equal to the contribution of n causal factors like so:

$$\mathbf{D} = \sum_{i}^{n} \beta_i \mathbf{X}_i + v_{\mathbf{D}} \tag{6}$$

where each $\mathbf{X}$ term represents the signal of a particular causal factor (again estimated using climate models), each β term the contribution of that factor, and $v_{\mathbf{D}}$ is the random noise term. Basically, we can think of the $\mathbf{X}$ terms like $\mathbf{X}_{GHG}$ as representing what would happen if $\mathbf{D}$ was entirely attributable to the increase in greenhouse gas concentrations.[10] The statistical test then concerns finding the β terms that maximize the fit between the two sides of the equation using least-squares methods. In standard tests, if $\beta_{GHG} = 1$ in the equation of best fit, then the test says that greenhouse gases account for 100% of the observed trends in the data. The change to a regression model thus provides climate scientists with a means of testing not just whether humans are responsible for climate change, but estimating how of much of it we're responsible for.

As with the fingerprinting technique described in the last section, this change in statistical model does not yield anything like a perfect model. Recall that the $\mathbf{X}$ terms are estimated using climate models, which we know can be inaccurate, a situation described in the statistical literature using the term "measurement error" (see Carroll et al. 2006). When there's measurement error in a regression, there's liable to be errors in the estimates of the β terms: there will either be aspects of $\mathbf{D}$ that we should attribute to a causal factor but that instead end up being registered as noise or aspects of $\mathbf{D}$ that we should attribute to noise but are instead attributed to a causal factor—and indeed, it's likely that that there will be both. These sorts of errors should be expected to crop up in attribution studies, even if we think that the climate models that we use in estimating the signals are themselves perfectly accurate representations of the laws of the climate system. After all, climate models—like the climate itself—contain internal variation or noise, meaning that we should expect that the signals generated by climate models will be noisy in much the same way that the data are.

The solution that climate scientists developed is to move to a more complex form of regression model originally introduced to attribution studies by Allen and Stott (2003).[11] So, keeping with our convention and letting $\mathbf{M}_{\mathbf{X}_i}$ and $v_{\mathbf{X}_i}$ be the model-

[10] On the subjunctive character of this variable, compare Mayo (1996, 135–38).

[11] This form of regression model is often known as an "errors-in-variables" (EIV) model in the statistical literature, but in attribution studies that terminology is typically used to refer to the more complex models introduced by Huntingford et al. (2006).

generated estimate of and internal noise term for the i^{th} $\mathbf{X}$ respectively, the standard presentation of the more complex regression problem is given by

$$\mathbf{D} = \sum_{i}^{n} \beta_i (\mathbf{M}_{\mathbf{X}_i} - v_{\mathbf{X}_i}) + v_{\mathbf{D}} \tag{7}$$

Essentially: to account for the fact that the estimates of the signals are possibly inaccurate, an additional statistical model—here of the relationship between $\mathbf{M}_{\mathbf{X}_i}$ and $\mathbf{X}_i$—is necessary. This additional statistical model is embedded in the first, and more complicated statistical tests are required to account for the additional source of variation (specifically, you need a "total" least squares algorithm rather than an "ordinary" least squares algorithm).

As was true of fingerprinting, the move to a regression model was originally motivated by the desire to answer questions that were previously outside of our reach. Also as was true of fingerprinting, this initial innovation was quickly complicated by refinements to the resulting statistical model aimed at improving the reliability of the resulting tests. This process is ongoing, and is in no sense finished by the innovations outlined here. For instance, it's usually assumed that $v_{\mathbf{X}_i}$ and $v_{\mathbf{D}}$ have the same "structure," meaning that they differ by only by a scalar factor. This is explicitly understood to be an idealization, in that it requires assuming that natural variation is the "dominant" source of noise in *both* the estimate of $\mathbf{X}_i$ and $\mathbf{D}$ itself. To date, there seems to be no consensus on the best method for improving the regression model in this respect; Huntingford et al. (2006) introduce further variation terms into the model determined using the distribution of simulation results given by ensembles of climate models, while more recent studies have tended to follow Schurer et al. (2018) in addressing this potential source of error by running multiple regressions using the estimates from different climate models. Just as is true of fingerprinting methods, in other words, climate scientists continue to study how to change the regression model in a way that will result in more accurate and/or precise estimates of the relevant quantities.

4 Calibrating statistical models

Before turning to the philosophical analysis, allow me to briefly review where we are. The measurement(-like) procedure used in contemporary attribution studies consists of the following. First, climate models are used to estimate various quantities of interest. Two quantities in particular—the expected signal ($\mathbf{M}_S$) and a corrected form of the inverse of the internal covariation matrix ($\mathbf{\Sigma}_{\mathbf{I}}^{-1}$)—are used to generate

a detector statistic ($\mathbf{D}$) from the "raw" data ($\mathbf{O}$). This statistic, estimates for the signal of causal factors such as the change in greenhouse gas concentrations ($\mathbf{M}_{\mathbf{X}_i}$), and estimates for the remaining variation ($v_{\mathbf{X}_i}$ and $v_{\mathbf{D}}$) are then used to estimate the contribution (β_i) of each causal factor. We can represent the resulting statistical model as a single equation:

$$\mathbf{\Sigma}_{\mathbf{I}}^{-1}\mathbf{M}_S\mathbf{O} = \sum_i^n \beta_i(\mathbf{M}_{\mathbf{X}_i} - v_{\mathbf{X}_i}) + v_{\mathbf{D}} \tag{8}$$

The presentation of this procedure and model are schematic, and there are various important differences between studies that fit this general pattern.[12] Nevertheless, the schematic description above captures the most common approach used in attribution studies in the late 2010s and—as we've seen—this approach relies on a statistical model that is substantially more complex than the simple model that was common in the early 1990s.

How should we understand the changes that are responsible for this increase in complexity? My suggestion is that the evolution of the statistical model can be understood as a kind of *calibration*. Of course, "calibration" can be used to refer to many different kinds of procedures. So, for example, zero-ing out a scale before using it to weigh something is a kind of calibration; it's adjusting a tool so that the resulting measurement outcome is more accurate. Similarly, climate scientists often talk about tuning or calibrating their models, where they mean adjusting particular parameters to generate the best fit with empirical data (Frisch 2015; Steele and Werndl 2013).

In the recent literature on philosophy of measurement, however, "calibration" is taken to mean one of two things. For Tal (2017), the outcome of a calibration process are functions that map instrumental readings (and potentially auxiliary parameters) to outcomes and vice-versa, like so:

$$\text{outcomes} \;=\; f(\text{readings},\; \text{aux. parameters}) \tag{9}$$

The goal of calibration is to identify the right functions (Tal 2017, 33). Calibration, in other words, is a matter of modeling a measurement process (Tal 2017, 34)

Bokulich (2020a), by contrast, explicitly departs from Tal's conception in taking "calibration" to be a matter not of modeling but of altering the measurement process. That is, where Tal understand calibration to be a matter of identifying the function (and auxiliary parameters) that relates readings to outcomes, for Bokulich it's a matter of fixing the measurement process to account for that information (Bokulich

[12] There are also quite a few studies that have suggested alterations to it; see, e.g., Hannart (2016), Katzfuss, Hammerling, and Smith (2017), Ribes, Qasmi, and Gillett (2021), and Ribes et al. (2017).

2020a, 430). To make the contrast concrete, consider an adjustment like zero-ing out a scale. Neither Bokulich nor Tal considers this kind of adjustment to be an instance of calibration in their preferred sense. For Tal, calibration is determining how the readings on the scale relate to the target quantity (weight). In particular, calibration is determining that the reading on the scale is equal to the weight of the object only when the auxiliary parameters are fixed in the right way—i.e., when the starting reading is equal to zero. For Bokulich, gathering this information is an importantly different step from putting it into practice; she uses "calibration" to refer to the step of introducing changes to the measurement process of the basis of this information. On her approach, therefore, calibration would be introducing the step "zero out the scale" into the measurement procedure.

In what follows, I'm going to adopt Bokulich's understanding of "calibration"—to my eyes, what's interesting about the attribution case outlined above are the changes that have been made to the process, meaning that Bokulich's approach is more likely to give us insight into attribution. As such, I'll understand calibration to be any change to either a physical process or the means of inferring outcomes from the indications that are generated by the physical process that aims making the procedure as a whole deliver more reliable results (compare Bokulich 2020a, 431). Note that in this sense, calibration can either operate at the individual level—changing the way that an individual calculates a particular outcome in particular setting—or a more social one, as in many of examples cited by Bokulich and Tal where the calibration process results in setting new measurement standards for the entire field. It is also (partly as a consequence) often temporally extended and iterative: frequently used measurement procedures are being consistently updated and refined to improve their accuracy and precision.[13]

In arguing for the thesis that the changes detailed above can be considered a kind of calibration, it will be helpful to have a paradigm case on the table for comparison. So consider Bokulich's discussion of calibration in carbon dating. As Bokulich (2020a, 437) recounts, initial carbon dating projects were based on the assumption of a constant atmospheric concentration of the carbon isotope ^{14}C. Of course, this assumption is false, but difficult to do away with: in order to determine how ^{14}C concentrations differ with time, one needs an independent means of measuring the age of various objects.[14] Estimates for how ^{14}C concentrations have changed are provided

[13]I'm assuming that the terms "accuracy" and "precision" are well-understood; see Bokulich (2020a) for a discussion of how these terms are typically used in philosophy of measurement. Following Bokulich, I'll also sometimes adopt the convention of using "reliable" to refer to procedures that generate results that are both accurate and precise.

[14]Chang (2004) contains discussion of a virtually identical problem in the context of temperature

by the combined evidence from tree rings (used to estimate more recent changes) and lake fossils (used to estimate more ancient changes). In other words, contemporary carbon dating procedures are (partially) de-idealized relative to the original carbon dating methods in the sense that they employ a more complex representation of atmospheric ^{14}C concentrations. As Bokulich (2020a, 440) stresses, however, this de-idealization comes with a cost in that the new method trades increased accuracy for additional potential sources of error: the accuracy of carbon dating procedures is now reliant on the accuracy of tree ring and fossil dating procedures. Because the de-idealization is only partial—and because the other dating procedures are themselves imperfect—the calibration of carbon dating remains an ongoing process.

Both of the two major examples outlined above share deep parallels with the carbon dating case. I'll discuss each in turn.

Fingerprinting. As we saw in section 1, early attempts to measure the human contribution to climate change began with a simple but false assumption—namely that temperature changes are randomly distributed—comparable to the assumption that ^{14}C is constant over time. This assumption was then replaced with a representation of how temperature changes are distributed derived from "independent" sources of evidence, namely climate models. As in the case of carbon dating, this partial de-idealization improved the reliability of the measurement procedure but did not remove all potential sources of error—indeed, it introduced additional potential sources of error insofar as the measurement procedure now relies on the accuracy of climate models. These new sources of error invited iterative improvements such as the regularization procedure outlined in section 2. As in the original move to fingerprinting, regularization involves a change to the measurement procedure that is explicitly understood as aimed at a limited improvement to the reliability of the measurement outcomes. That is, it's explicitly understood that the resulting procedure, while better, is still not perfectly reliable.

Regression. Our second example is similar. As before, the story begins with an extremely simple and idealized model in which there is (a) the warming signal and (b) noise. This model was complicated by the shift to a linear representation of the the various factors involved in warming. This shift improved the overall reliability of the results in that it allowed scientists to answer quantitative questions about the origins of climate change, but the resulting regression model was explicitly recognized as containing additional remaining idealizations and thus additional potential sources of error. And, as in the fingerprinting case, we saw scientists continuing to work on improving the measurement procedure by iteratively introducing more complications into the statistical model. In more detail, though employing a representation

measurement.

of the effects of individual causal factors (represented by the $\mathbf{X}$ terms) allowed climate scientists to make more reliable estimates of humanity's overall contribution to climate change, these representations are themselves uncertain, opening the door for more complex treatments that better account for the potential of error in the final measurement outcome due to a misrepresentation at this stage.

Of course, while the two cases outlined above are extremely similar to the case of carbon dating discussed by Bokulich, there are a number of features of the climate modeling case that don't make it into the above summaries and deserve further discussion.

First, and most notably, the changes discussed in this paper are changes to specifically statistical elements of the measurement procedure. In particular, they're changes to the statistical model that's used to distinguish between signal and noise. That's not to say that attribution studies have not benefited from the calibration of other, more familiar, elements; they have. Rather, it's to point out that the statistical models employed in measurement *can* undergo changes similar to what we call calibrations in the context of physical instruments and theoretical models—and, further, that these changes can be extremely valuable in improving the overall reliability of the procedure.

The next section will discuss the implications of this first point in more detail, but it is worth emphasizing one here: what's going on in these cases is not improving the statistical model qua representation—at least not primarily—but rather improving it qua component of the larger measurement process. So, for example, the move from $\mathbf{\Sigma_M}$ to $\mathbf{\Sigma_I}$ is *not* driven by the thought that $\mathbf{\Sigma_I}$ is a more accurate representation of the "true" internal variation than $\mathbf{\Sigma_M}$. If anything, $\mathbf{\Sigma_I}$ is a less accurate representation. Instead, what drives this change are worries about the mathematical properties of $\mathbf{\Sigma_M}$ and the errors that they can introduce into the final estimates. In other words, while paradigmatic examples such as Bokulich's carbon-dating case involve theoretical corrections or alterations to physical instrumentation rather then changes to the statistical model, the motivation for the present changes is the same as we find there.

Second, and related, unlike what is the case in most paradigm cases of calibration, the changes found in the present examples were not driven by empirical work.[15] Instead, the examples outlined above were motivated first and foremost by either (a) a desire for quantitative answers where only qualitative answers were previously possible or (b) mathematical / statistical difficulties arising from specific idealizations. Both of these motivations deserve a comment. On the first front, it's an interesting

[15]That's not to say that Bokulich and Tal don't recognize a role for theoretical considerations in calibration; see, e.g., Tal (2018, 642).

question where calibration ends and simply measuring a different quantity begins. It's not obvious that scientists aiming to quantitatively estimate human contributions to climate change are using a "calibrated" version of the procedure employed by scientists aiming for mere detection—there's an argument to be made that these are simply two different procedures. Of course, as Chang (2004) brings out, the same questions arise in the context of thermometer development and, regardless, the answer doesn't really matter for my main point. Even if the moves to fingerprinting and regression themselves are best understood as involving new measurement procedures rather than changes to the old one, the subsequent move to regularized fingerprints and more complex regression equations are much more in line with the paradigm case. These refinements represent changes to a single measurement procedure—the same quantity is being estimated using the same data (source) both before and after the refinement—that serves to make the resulting estimate more reliable. In this respect, these refinements are exactly like Bokulich's carbon-dating case.

The second motivation involves addressing specific mathematical or statistical difficulties arising due to particular idealizations. As I'll stress even more in the final section, climate scientists seem to have never been deeply concerned with trying to achieve a statistical model that perfectly represents what the statisticians gloss as the process by which the data is in fact generated. At every juncture, the statistical models have been understood to be heavily idealized. Only some of these idealizations can be expected to seriously undermine the reliability of the measurement practice, and it's these specific idealizations that scientists have devoted their time to addressing. Notably, the aim here has not necessarily been to remove the problematic idealization—that's not what regularization does, for instance—but to neutralize it as a potential source of error.

A final point worth emphasizing is that in the case of attribution studies, there's no centralized authority who fixes *the* way to measure human contribution to climate change.[16] On the contrary, individual climate scientists decide which methods to adopt in any particular study.[17] The propagation of a technique like regularization through the community can be slow, and sometimes new statistical models that seem to show empirical promise are never universally adopted. So, for instance, Huntingford et al. (2006) introduced a variation of the regression model that is less idealized

[16]On my understanding, this fact distinguishes the present case from some of those discussed by Tal, but not the carbon dating case discussed by Bokulich (2020a).

[17]Note that the same authors often contribute to studies that employ different methods. So, for instance, Gillett et al. (2021) runs a fairly standard attribution analysis in the mold discussed here; two of its authors also contributed to Ribes, Qasmi, and Gillett (2021), which adopts a markedly different Bayesian approach.

than the one outlined above. It suffers, however, from being more difficult to carry out, and climate scientists seem to have largely decided that it's better—because both less labor-intensive *and* less reliant on questionable auxiliary assumptions—to achieve the same effect by carrying out multiple regressions and effectively averaging over the results (see Hegerl et al. 2019; Schurer et al. 2018). But the decision-making process here is extremely decentralized.

It seems to me that these differences between the changes to the statistical models used in attribution studies and the paradigm cases of calibration are not significant enough to prevent us from extending the concept of calibration to the former. At its most basic, calibration (in Bokulich's sense) involves altering how we estimate some quantity in such a way that the refined procedure is understood to generate more accurate or precise estimates of the same quantity. This is precisely how climate scientists understand the alterations surveyed in the prior sections: the refinements involved in fingerprinting, regularization, and regression are taken to allow for tests that generate more reliable estimates for humanity's contribution to climate change than were previously possible. What's changed is simply which element of the measurement process is being adjusted: the examples discussed above show that calibration—or at least something like it—can also involve changes to the inferential tools employed in the measurement process. Even if we want to restrict "calibration" to the more limited phenomenon, therefore, the cases discussed in this paper share a kinship with paradigm cases of calibration.

This kinship matters. If the statistical tools employed in measurement procedures undergo changes like those found in traditional cases of calibration and if, further, we want to be able to evaluate whether these changes are successful, then we need some account of what makes one statistical model better than another. I turn to this problem now.

5 Evaluating statistical models

The prior four sections have outlined how the statistical models employed in attribution studies have been continually refined over the last three decades in a process that should be understood as a kind of calibration. This final section turns to the evaluation of statistical models and shows that recent accounts of model evaluation—accounts that focus on the adequacy of the model for a particular purpose or task—can be extended to neatly account for the calibration of statistical models. On the view that I'll defend, what makes for a good statistical model is whether (and to what degree) it allows for reliable inferences from (actual) data to the quantity of interest. A statistical model is capable of fulfilling this function to the extent that

it accurately represents the relationship between hypotheses and data. But accuracy in other, deeper, senses is largely irrelevant: if the statistical model gets the probabilities right, it doesn't matter if it misrepresents the actual mechanisms of data generation (say). What makes calibration successful, then, is if the changes to the statistical model improves the representation of the probabilities.

In many domains, good statistical practice means designing experiments to fit a given statistical model; this is the idea behind randomized control trials, for instance.[18] In attribution studies, however, the data is produced by a natural experiment that we have little ability to physically control. The upshot is that climate scientists have to mold their statistical models to fit the structure of the experiment that actually exists rather than the other way around. As we've seen, this a continuing and iterative process: over time, the standard or typical statistical model employed in attribution studies has been improved by the refinements outlined above.

If we begin from a perspective on statistics motivated by the first type of case—where the goal is physically controlling the experiment so that it best approximates a true random distribution—a natural question arises regarding the second type of case: how should we judge whether (or to what degree) the calibration process discussed in the prior sections is successful? Or, more simply, what makes for a good statistical model of a non- (or only partially) randomized process? The natural answer to this question would be that a good statistical model is just one that accurately represents the relationship between the data and the target quantity. Calibration would then be a matter of building a more accurate model. But by this metric, every one of the statistical models surveyed above is a failure: all of them are heavily idealized. Worse, at least some of the alterations surveyed above don't seem to actually make the models more realistic—or at least, they do so only incidentally. Regularization, for example, has been widely adopted even though $\mathbf{\Sigma_I}$ is no more accurate than $\mathbf{\Sigma_M}$ qua representation of the "true" covariance matrix. And if the statisticians are to be believed, the same is true to a greater or lesser extent of statistical models generally: they always involve at least some degree of idealization.

As Mayo has pointed out, the refrain that all models are false—while true so far as it goes—isn't really informative; it doesn't tell us "how statistical models may be used to infer true claims about problems of interest" (Mayo 2018, 296). On Mayo's view, a statistical model is "adequate for a problem of statistical inference" (Mayo 2018, 297) if it allows scientists to thoroughly test or "probe" the hypotheses in question. For Mayo, who prefers a classical approach, this means that a well-designed statistical model is one that makes it suitably difficult for a false hypothesis

[18] See Worrall (2002) and Zhao (2021) for discussion and criticism of this view in the social and life sciences.

to pass the test (Mayo 1996, 178). She suggests that such models "approximate" the real world in some way (Mayo 2018, 296–97), but there's no guarantee that better approximations will lead to models that are better for statistical testing—indeed, given that "no useful models are true" (Mayo 2018, 297), it can't be that usefulness always increases with accuracy.

Mayo's view fits nicely with the recent literature on model evaluation, which stresses that models should not *in general* be evaluated in terms of completely accurate representation of the target system.[19] As Bokulich (2020b) emphasizes, models are used for a variety of different purposes: sometimes, the role of a model is to represent the target as accurately as possible; in other contexts, however, the model might be used (e.g.) as a control in an experiment testing the effects of some intervention. In these cases, as I put it in Dethier (2021), the model is less like a truth-apt proposition and more like a reliability-apt tool: the appropriate question is whether it reliably or adequately performs the role that it has been tasked with. (Or, better: whether it performs its role in a way that renders the relevant procedure as a whole reliable.) Since, on this picture, models are treated as tools or instruments rather than representations, I'm going to borrow a term from Knuuttila (2011) and call the family of views of model evaluation defended in this literature "artefactual." More precisely, a view of model evaluation is "artefactual" iff it says that the quality of a model is determined by how well it performs a particular function in a particular context.

What properties do the statistical models involved in attribution studies have to have in order to fulfil their function of licensing inferences from data to hypotheses about the causes of climate change? Or, in Mayo's terminology, what properties must they have to facilitate the severe or thorough testing of hypotheses about the past? Ultimately, and regardless of whether we're employing a Bayesian or classical approach, whether or not a hypothesis passes a statistical test depends on the likelihood relationship between said hypothesis and the observed data.[20] In order to fulfil their function, therefore, the statistical models employed in attribution studies must accurately represent these likelihoods: a model that accurately represents

[19] See the work of Bokulich and Parker in particular (e.g. Bokulich 2021; Bokulich and Parker 2021; Parker 2009, 2020b) as well as Currie (2017) and Dethier (2021).

[20] There are sometimes other factors in both the classical and Bayesian context. For classical statistics, for instance, we also need to capture the probability of more "extreme" data that could have been observed; in the Bayesian case, we want to know the prior probability distribution. My point here is only that accurate representation of the mechanisms that generate these probabilities are unimportant. Spanos makes essentially the same point in arguing that "the only thing that matters ... is whether the data $\mathbf{x}$ can be realistically be viewed" as a typical outcome of the probabilistic process given by the model (Spanos 2006, 100). See also Dethier (2022) and Sprenger (2019).

the likelihood of the data on various hypotheses will license the right inferences regardless of whether or not it accurately represents how the data are generated in any deeper sense. And, thus, thoroughly testing a hypothesis in (roughly) Mayo's sense involves showing that the statistical model employed accurately represents the relevant likelihoods.[21]

As stressed by the literature on model evaluation more generally (see, e.g., Dethier 2021; Parker 2020b), the quality of a model usually depends on contextual factors; the same is true here. Because attribution studies rely on both actual data and information from climate models, the relevant context for attribution studies consists of (at least) the character of the data sets and climate models available. For instance, a statistical model that is perfectly suited for cases in which our climate models are infinitely accurate is not one that is well-suited for the actual conditions that we find ourselves in. As we saw above, much of motivation for changes to statistical models arises from the imperfect nature of the climate models employed in estimating various quantities.

Though the view just sketched is relatively abstract—see Parker (2020b) for one detailed way of precisifying the general artefactual picture—it neatly accounts for much of the behavior of climate scientists in the calibration of statistical models. Artefactual views of model evaluation require that statistical models are appropriately catered to the particularities of the context in which they are being used. Statistics is not a one-size-fits-all science: we can't use the same statistical model in every case, because which statistical model (or models) are appropriate depends on (among other things) the actual relationship between the data and the target of inference (compare Mayo 1996, 447–49). The statistical model that is appropriate when our data are genuinely randomly distributed around the target quantity is very different from the one we need in attribution contexts, where the data have a much more complicated relationship with the target we want to estimate. As a result, in ongoing projects such as the attribution of climate change, it's unsurprising that we would find both (a) that the early statistical models were rather simple first approximations based on general principles and poorly tailored to the specifics of the case and (b) that there exists a process of calibration in which these models are rendered increasingly well-suited for the particular task.

[21] I suspect Mayo would object to this gloss—and rightly so insofar as it implies either that a good model yields an "automatic" pass or reject rule or that there's no room for the kind of local questions that go "beyond statistical methodology, as traditionally conceived" (Mayo 1996, 449). I don't intend either implication, however: though this hasn't been the focus in this essay, Mayo is right to stress how much non-statistical work goes into showing that a given model is adequate. See also Spanos (2006).

Artefactual views also entail that idealizations are problematic only insofar as they undermine the ability of the model to perform its assigned task. As we saw in the discussion of regularization, one way that the project of attribution can go wrong is if the estimate for internal variation ($\mathbf{\Sigma_M}$) is ill-conditioned in the sense that it can't be inverted without introducing substantial uncertainty. Notice that this problem isn't purely representational; the problem isn't really that treating $\mathbf{\Sigma_M}$ as though it were an accurate representation of the true internal variation is an idealization, but rather that this *specific* idealization introduces technical problems for the calculation of the detector statistic. That is: this idealization is considered worrying precisely because it inhibits the ability of the statistical model to generate accurate likelihood values (compare Mayo 2018, 296).

It's worth reiterating that the different solutions that climate scientists suggested to resolve this problem did not involve eliminating the idealization, but rather neutralizing it as a source of error. The first of these methods was truncating $\mathbf{\Sigma_M}$. The truncated form of $\mathbf{\Sigma_M}$ is not a more accurate representation of the true internal variation than the non-truncated form; if anything, it is less accurate. Similarly, the move from truncation to regularization was motivated primarily by two desires: (a) to better solve the technical problem that motivated truncation in the first place and (b) to reduce the imprecision introduced by the (semi-)arbitrary choice of truncation level k. It's telling, in this respect, that (to my knowledge) no attempts have been made to either truncate or regularize representations of internal variation elsewhere in the statistical model. The simple explanation is that covariance matrices involved in fixing the distribution of random variables like $v_{\mathbf{D}}$ are not inverted, and so the technical problem does not arise.[22] The misrepresentation of the internal variation in these cases doesn't matter (as much) for the purposes that the model is put to.[23]

Finally, recall again the contextual character of model evaluation on artefactual views. From a purely conceptual perspective, we can distinguish between three distinct components of a measurement(-like) procedure: the physical instruments that deliver the data, the theoretical understanding of the target that makes the data

[22] As Mayo emphasizes, errors—and thus the strategies for neutralizing them—should be individuated at a relatively fine grain when evaluating methodology (see, e.g., Mayo 1996, 19); an error owed to matrix inversion is different from an error owed to bad data because they have different implications and can be rectified in different ways.

[23] How can we justify our conviction that calibration efforts improve inferential accuracy without appealing to representational accuracy? Both of the cases surveyed here seem to have been initially motivated on theoretical grounds—there are reasons to expect an increase in inferential accuracy by making the change—which were then bolstered by comparing the performance of the different methods against data with known properties (see, e.g., Hannart, Ribes, and Naveau 2014; Ribes, Planton, and Terray 2013).

meaningful, and the statistical model that connects the two. In measurement, the goal is to use the data to fill in or update the theoretical understanding; a successful measurement procedure relies on having good (enough) instruments and data, good (enough) prior theoretical understanding to determine what the data are telling us, and good (enough) statistical models to carry out the inference. On artefactual views, the quality of each of these elements cannot properly be judged separately from the others: whether the statistical model is "adequate" for the particular measurement procedure depends on the nature of the instruments and theory that are actually in place (compare Parker 2020b, 463; Dethier 2021, 1462–63). So while there's a conceptual distinction between these three elements, they are deeply interconnected both causally—our theoretical understanding partly determines which statistical models we use—and for evaluative purposes.

The problem that motivates this paper is how to evaluate the changing character of statistical models: what makes one model better than another? On the artefactual view sketched here, the answer to this question in the context of climate change attribution is that a model is better insofar as it allows for more accurate and precise estimates of human contributions to climate change within the context of the actual data and background knowledge—that is, the climate models—available. In effect, it's a matter of how accurately the model captures the probabilistic relationship between *this* data and hypothesis given *that* background information. Taking this artefactual view helps us understand the calibration (or calibration-like) process that the statistical models employed in climate change attribution have undergone over the last three decades: what climate scientists are doing is effectively tailoring one of their instruments to allow for increasingly precise and accurate estimates of the quantity they want to measure. What separates this case from more familiar instances of calibration is simply that the instrument in question is a statistical model instead of a thermometer.

Adopting this view also makes it clear how climate scientists can justify using statistical models that they nevertheless think can be improved. On the artefactual view, a statistical model can be improved just in case it's possible to increase either the precision or the accuracy of the measurement outcomes that it is used to generate. Climate scientists are therefore justified in using sweeping idealizations such as those surveyed in the first section so long as they're careful to account for the imprecision or inaccuracy that this use introduces. From this perspective, a model may both be "adequate" in that it warrants the conclusion given and simultaneously in need of improvement insofar as the conclusion given is not as precise or accurate as we would like it to be. The context-sensitivity of the artefactual view thus not only captures the calibration of statistical models, it explains why climate science doesn't

need to wait for that calibration process to complete before delivering at least some conclusions.

6 Conclusion

In this paper, I've discussed the development of attribution studies in climate science, and particularly how the statistical models employed in attribution studies have changed over the last three decades. The primary upshot of this discussion is that these changes should be viewed as a kind of calibration process: they're iterative revisions to a measurement(-like) procedure motivated primarily by the aim of neutralizing particularly troublesome sources of error or uncertainty.

To make sense of calibration in the context of statistics, it's important to recognize that statistics is flexible: because statistics can be applied in a wide variety of contexts, our statistical models must be tailored to the context that we actually find ourselves in. To this end, I've argued for the extension of what I've called "artefactual" views of model evaluation to the statistical case: what makes a statistical model good, on this view, is that it is well-tailored to the particular purpose to which it is put. Once we recognize that for the purposes of evaluation statistical models are more akin to tools than to propositions it becomes clearer why climate scientists have focused their calibration efforts on particular elements of the statistical models and why they've consistently used these models while recognizing that they are idealized and imperfect.

References

Allen, Myles R. and Peter A. Stott (2003). Estimating Signal Amplitudes in Optimal Fingerprinting, Part I: Theory. *Climate Dynamics* 21.5-6: 477–91.

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

Bokulich, Alisa (2020a). Calibration, Coherence, and Consilience in Radiometric Measures of Geologic Time. *Philosophy of Science* 87.3: 425–56.

— (2020b). Towards a Taxonomy of the Model-Ladeness of Data. *Philosophy of Science* 87.5: 793–806.

— (2021). Using Models to Correct Data: Paleodiversity and the Fossil Record. *Synthese* 198.S24: 5919–40.

Bokulich, Alisa and Wendy S. Parker (2021). Data Models, Representation, and Adequacy-for-Purpose. *European Journal for Philosophy of Science* 11.31: 1–36.

Box, George E. P. (1979). Robustness in the Strategy of Scientific Model Building. In: *Robustness in Statistics*. Ed. by Robert L. Launer and Graham N. Wilkinson. New York: Academic Press: 201–36.

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

Chang, Hasok (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Currie, Adrian (2017). From Models-as-Fictions to Models-as-Tools. *Ergo* 4.27: 759–81.

Dethier, Corey (2021). How to Do Things with Theory: The Instrumental Role of Auxiliary Hypotheses in Testing. *Erkenntnis* 86.6: 1453–68.

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

Edwards, Paul (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge: MIT Press.

Frisch, Mathias (2015). Predictivism and Old Evidence: A Critical Look at Climate Model Tuning. *European Journal for Philosophy of Science* 5: 171–90.

Gillett, Nathan P. et al. (2013). Constraining the Ratio of Global Warming to Cumulative CO_2 Emissions Using CMIP5 Simulations. *Journal of Climate* 26.18: 6844–58.

Gillett, Nathan P. et al. (2021). Constraining Human Contributions to Observed Warming since the Pre-industrial Period. *Nature Climate Change* 11: 207–12.

Hannart, Alexis (2016). Integrated Optimal Fingerprinting: Method Description and Illustration. *Journal of Climate* 29.6: 1977–98.

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

Hasselmann, Klaus (1993). Optimal Fingerprints for the Detection of Time-dependent Climate Change. *Journal of Climate* 6.10: 1957–71.

— (1997). Multi-pattern Fingerprint Method for Detection and Attribution of Climate change. *Climate Dynamics* 13: 601–11.

Hegerl, Gabriele C. et al. (2019). Causes of Climate Change over the Historical Record. *Environmental Research Letters* 14.12: 123006.

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

Katzfuss, Matthias, Dorit Hammerling, and Richard L. Smith (2017). A Bayesian Hierarchical Model for Climate Change Detection and Attribution. *Geophysical Research Letters* 44.11: 5720–28.

Knuuttila, Tarja (2011). Modelling and Representing: An Artefactual Approach to Model-based Representation. *Studies in History and Philosophy of Science Part A* 42.2: 262–71.

Levine, Richard A. and L. Mark Berliner (1999). Statistical Principles for Climate Change Studies. *Journal of Climate* 12.2: 564–74.

Lloyd, Elisabeth (2012). The Role of "Complex" Empiricism in the Debates about Satellite Data and Climate Models. *Studies in History and Philosophy of Science Part A* 43.2: 390–401.

Mayo, Deborah G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago, IL: University of Chicago Press.

— (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Parker, Wendy S. (2009). Confirmation and Adequacy-for-Purpose in Climate Modelling. *Aristotelian Society Supplementary Volume* 83.1: 233–49.

— (2010). Comparative Process Tracing and Climate Change Fingerprints. *Philosophy of Science* 77.5: 1083–95.

— (2016). Reanalyses and Observations: What's the Difference? *Bulletin of the American Meteorological Society* 97.9: 1565–72.

— (2017). Computer Simulation, Measurement, and Data Assimilation. *The British Journal for the Philosophy of Science* 68.1: 273–304.

— (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.

Ribes, Aurélien, Jean-Marc Azaïs, and Serge Planton (2009). Adaptation of the optimal Fingerprint Method for Climate Change Detection using a Well-conditioned Covariance Matrix Estimate. *Climate Dynamics* 33.5: 707–22.

Ribes, Aurélien, Serge Planton, and Laurent Terray (2013). Application of Regularised Optimal Fingerprinting to Attribution. Part I: Method, Properties and Idealised Analysis. *Climate Dynamics* 41.11-12: 2817–36.

Ribes, Aurélien, Saïd Qasmi, and Nathan P. Gillett (2021). Making Climate Projections Conditional on Historical Observations. *Science Advances* 7.4: 1–9.

Ribes, Aurélien et al. (2017). A new Statistical Approach to Climate Change Detection and Attribution. *Climate Dynamics* 48.1: 367–86.

Santer, Benjamin D. et al. (1996). Detection of Climate Change and Attribution of Causes. In: *Climate Change 1995: The Science of Climate Change*. Ed. by John T. Houghton et al. Cambridge: Cambridge University Press.

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

Spanos, Aris (2006). Where Do Statistical Models Come from? Revisiting the Problem of Specification. *Lecture Notes-Monograph Series* 49: 98–119.

Sprenger, Jan (2019). Conditional Degree of Belief and Bayesian Inference. *Philosophy of Science* 87.2: 319–35.

Steele, Katie and Charlotte Werndl (2013). Climate Models, Calibration, and Confirmation. *The British Journal for the Philosophy of Science* 64.3: 609–35.

Suppes, Patrick (1962). Models of Data. In: *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*. Ed. by Ernest Nagel, Patrick Suppes, and Alfred Tarski. Stanford: Stanford University Press: 252–61.

Tal, Eran (2017). Calibration: Modelling the Measurement Process. *Studies in the History and Philosophy Part A* 65-66: 33–45.

— (2018). Naturalness and Convention in the International System of Units. *Measurement* 116: 631–43.

— (2020). Measurement in Science. In: *Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. URL: https://plato.stanford.edu/entries/measurement-science/.

von Storch, Hans and Francis W. Zwiers (1999). *Statistical Analysis in Climate Research*. Cambridge: Cambridge University Press.

Worrall, John (2002). *What* "Evidence" in Evidence-Based Medicine? *Philosophy of Science* 69: S316–30.

Zhao, Kino (2021). Sample Representation in the Social Sciences. *Synthese* 198.10: 9097–115.