[This is the final draft of a preprint, accepted for publication at *European Journal for Philosophy of Science*]

Title: Diagnosing errors in climate model intercomparisons

Author: Ryan O'Loughlin

ORCiD: 0000-0002-9106-1460

Institution: Queens College CUNY

Email: roloughlin@qc.cuny.edu

Diagnosing errors in climate model intercomparisons

Abstract. I examine error diagnosis (model-model disagreement) in climate model intercomparisons including its difficulties, fruitful examples, and prospects for streamlining error diagnosis. I suggest that features of climate model intercomparisons pose a more significant challenge for error diagnosis than do features of individual model construction and complexity. Such features of intercomparisons include, e.g., the number of models involved, how models from different institutions interrelate, and what scientists know about each model. By considering numerous examples in the climate modeling literature, I distill general strategies (e.g., employing physical reasoning and using dimension reduction techniques) used to diagnose model error. Based on these examples, I argue that an error repertoire could be beneficial for improving error diagnosis in climate modeling, although constructing one faces several difficulties. Finally, I suggest that the practice of error diagnosis demonstrates that scientists have a tacit-yet-working understanding of their models which has been under-appreciated by some philosophers.

1. Introduction

Scientists investigate Earth's climate via simulation models run on supercomputers. Sometimes these climate models give results that are at odds with each other. To climate modelers, such disagreements, as well as discrepancies between model results and other data sources, may suggest that there is something wrong in one or more models. I call these potential sources of disagreement "model errors." Clearly, diagnosing these errors and understanding how to fix them are important to climate modeling and to knowledge generation more generally. One endeavor to diagnose such errors is through the climate model intercomparison projects. In this paper, I address the following questions: how are model errors diagnosed? Why are diagnoses difficult? How can they be improved?

Climate model error diagnosis is either misunderstood or has been given little attention in philosophy of climate science. Many scholars have discussed the significance of model agreement (e.g., Parker 2011, 2018a; Lloyd 2015a; Winsberg 2018; Odenbaugh 2018; O'Loughlin 2021) and also interpretations and statistical evaluations of climate model ensembles (Annan and Hargreaves 2010, 2017; Jebeile and Barberousse 2021; Dethier 2022). Yet not many have discussed climate model error diagnosis. Lenhard and Winsberg (2010) are one major exception. They claim that it is impossible to say which part of a climate model is responsible for a particular error given the complexity of the model and how it was developed. However, given the prevalence of model error diagnosis in the scientific literature and practice, their skepticism is either unwarranted or its scope must be clarified and potentially revised.

My analysis is based on concrete examples from the scientific literature.[1] Scientists have diagnosed model errors by employing physical reasoning about model output and based on their knowledge of the climate system. Expectations about known behaviors of particular components of climate models are also drawn upon to explain model errors, and there are other strategies besides. These methods help scientists locate the source of errors and improve climate models as the models are further developed. In addition, since the 1970s, the infrastructure for intercomparing climate models has become larger and more diverse, and knowledge of individual models has become more dispersed across the growing number of experts helping build climate models. I suggest that the increasing complexity of model intercomparison practices is an alternative explanation for why model error diagnosis is difficult in practice, in contrast to Lenard and Winsberg's (2010) emphasis on individual model complexity and the historical legacy of code.

Further, to improve error diagnosis, I suggest that scientists should clearly state their expectations for likely model error and compile an "error repertoire" (inspired by and adopted from Mayo 1996) as reference and guidance for future model error analysis. Scientists' success in model error diagnostics, despite the complexity of models and the complexity of model intercomparisons, may suggest that scientists have a tacit-yet-working knowledge[2] about climate models' behavior—a kind of Duhemian "good sense"—that is worthy of future philosophical analysis.

In section 2, I review the current discussion of climate model error diagnosis by focusing on Lenhard and Winsberg (2010). In section 3, I describe the increasing complexity of climate model intercomparison practices that has occurred over time which makes error diagnosis more difficult. In

[1] The examples (and my emphasis in this paper) are focused on multi-model disagreement. For work centered on model-observation discrepancies, including examples of models being used to correct errors in observational and other data, see Lloyd 2012; Abraham et al. 2013; Mann 2018; Weart 2020; and Li (2022).

[2] By "tacit" I have in mind a sort of practice-based knowledge which scientists could perhaps explain to others if pressed but which they typically do not explain to others. Thanks to Matthew Mayernik for prompting me to clarify my use of this term and for pointing me to the work of Schmidt (2012) who discusses how, in many scientific and academic contexts, "tacit" is a "conceptual muddle that mystifies the very concept of practical knowledge" (163).

section 4, based on several examples of model error diagnosis, I distill general strategies behind error diagnostic practices. In section 5, I suggest an error repertoire as guidance for future error diagnosis.

2. Confirmation Holism and analytic understanding of climate models

The models we are concerned with are general circulation models (GCMs). GCMs simulate the atmospheric and oceanic circulatory patterns on earth and are used for applications in both weather and climate. GCMs are run on supercomputers and consist of computer code representing mathematical equations based on physical principles, such as classical physics (e.g., Navier-Stokes equations). These governing equations describe mass and energy transfer in the atmospheric, oceanic, ice, and land components of the climate system. For reasons of computational efficiency and due to the very small scales of certain physical phenomena, some processes (e.g., cloud physics, turbulence) are not explicitly represented in the model but are instead parameterized. Parameterizations—which we can think of as sub-models—are used to represent the effect of small-scale processes "at the grid scale of the model" (Gettelman and Rood 2016, 46). These sub-models come in varying degrees of complexity and may have empirical support or be derived from theory (Lloyd 2015a).

Lenhard and Winsberg (2010) claim that climate scientists do not have analytic understanding of their GCMs, meaning that scientists cannot "identify the extent to which each of the sub-models of a global model is contributing to its various successes and failures" (258). These "failures" include cases where a climate model's results are at odds with the results of other climate models, and so their account implies that error diagnosis in climate modeling is impossible. Their reasons for thinking this are fourfold, which I will explain in the following two subsections. The first three reasons concern what they claim are features of climate models and their development: fuzzy modularity, kludging, and generative entrenchment. Their fourth reason concerns examples from climate modeling wherein model error diagnoses were apparently either not possible or were severely limited.

2.1 Fuzzy modularity, kludging, and generative entrenchment

Let's begin with the notion of fuzzy modularity. "Modularity" refers to the fact that GCMs are composed of sub-models (the atmosphere module, the cloud parameterization, sub-parameterizations, the land module, etc.). Climate modelers typically differentiate between parameterizations, which represent specific processes at sub-grid scales, and modules, such as an atmosphere module, which themselves contain a host of parameterizations, but we can regard them all as different types of sub-models in that they are all *parts* of a *whole* GCM.[3] Lenhard and Winsberg use the term "fuzzy" to capture two different ideas about climate models. The first is that, as a GCM simulates climate, it is the *interaction* of the sub-models that jointly produce the model output. In their words,

> The overall dynamics of one global climate model is the complex result of the interaction of the modules—not the interaction of the results of the modules. For this reason, we like to modify the word "modularity" with the warning flag 'fuzzy': due to interactivity, modularity does not break down a complex system into separately manageable pieces (Lenhard and Winsberg 2010, 256).

This makes it difficult to isolate components of a GCM and infer exactly how they modify its overall behavior. For instance, if one is interested in diagnosing how a new cloud parameterization will change a GCM's response to aerosol forcing, it is not enough to examine both the GCM and the cloud parameterization independently—one also needs to examine how the model output changes after implementing the new parameterization. However, Lenhard and Winsberg emphasize that it is not possible to tell whether the behavior of the 'GCM + new cloud parameterization' is due to the interaction of the new cloud parameterization with the chemistry sub-model, with the vegetation sub-model, or some other component (or combination of components) in that GCM.

[3] Lenhard and Winsberg seem to use "sub-model" and "module" interchangeably. In contrast, I adopt climate scientists' typical usage of these terms, except when directly quoting Lenhard and Winsberg. Effectively this means that sub-models are parameterizations or sub-parameterizations, and the term "modules" is (usually, but not always) reserved for larger pieces of a GCM such as the atmosphere module or ocean module.

The second notion of "fuzzy" relates to the development of sub-models (discussed further below). Lenhard and Winsberg claim that parameterizations are built and tested "on the basis of the parameterizations that are already part of the concrete model under construction" which means that later modeling "steps" are influenced by the "accumulated effects of previously implemented steps" (256).[4] This creates a "'fuzzy' kind of modularity: normally, [sub-models] are thought to stand on their own. In this way, modularity should have the virtue of reducing complexity. In our present case, however, the [sub-models] are interdependent and therefore lack this virtue" (256).

Another key idea Lenhard and Winsberg discuss is called "kludging," which was originally a slang term in the computer programming world. As philosopher Andy Clark describes it, a kludge is "an inelegant, 'botched together' piece of program; something functional but somehow messy and unsatisfying" (1987, 278). Moreover, a kludge may be poorly understood such that its limitations and range of applications are unknown. Kludges are relevant to GCMs, because GCMs are run on computers. As Lenhard and Winsberg say, "A kludge is built to optimize the performance of the overall model as it exists at that particular time, and with respect to the particular measures of performance that are in use right then. There is no guarantee that an implemented kludge is optimal in any general sense" (2010, 257).

Kludges also relate to Lenhard and Winsberg's claim that path-dependency and the historical character of climate model development can best be understood in terms of William Wimsatt's notion of "generative entrenchment" (Wimsatt 2007). The basic idea is that some components in climate models, including kludges and model components "that are not related to principled considerations," may have other model components functionally depending on them and may therefore constrain the ability of the GCMs' development at later stages (257).[5]

[4] Compare with Morrison (2021). Lenhard and Winsberg's description of model development appears reasonable but may not be accurate to practice.
[5] But see Morrison (2021) for a practice-informed study of how climate modelers prioritize, research, and implement updates to their model over the course of development. Also, large-scale rewrites of GCM code are sometimes done in practice, contrary to Lenhard and Winsberg's description of climate model development (e.g., see Neale et al. 2012).

Lenhard and Winsberg claim that the above-described features of GCMs—fuzzy modularity, kludging, and generative entrenchment—jointly result in a form of confirmation holism that imposes severe limitations for climate scientists who wish to isolate specific components of GCMs that are responsible for specific instances of the models' successes and failures.

The result, according to Lenhard and Winsberg, is a failure of analytic understanding, which is the level of understanding one has "when one is able to identify the extent to which each of the sub-models of a global model is contributing to its various successes and failures" (258). The problem Lenhard and Winsberg claim to identify is that, due to the complexity of interactions between sub-models, "it becomes impossible to independently assess the merits or shortcomings of each sub-model…The ideal of analytic understanding is profoundly impeded by what appears to be a particularly vicious form of confirmation holism" (258).

2.2 Examples of alleged failure to diagnose model error

Lenhard and Winsberg supplement their argument by discussing some empirical evidence, i.e., examples from the climate model intercomparison literature of a failure to identify model error by attributing it to specific sub-models. The examples they cite include the Atmospheric Model Intercomparison Project (AMIP) (Gates 1992), phase 1 of the Coupled Model Intercomparison Project (CMIP) (Meehl et al. 2000), and the Aqua-Planet Experiment Project (APE) (Neale and Hoskins 2000).

Lenhard and Winsberg note that one of the aspirations expressed early in the model intercomparison literature, especially AMIP, was to be able to "make inferences about the performances of the various sub-components of the models and to attribute the diagnosed strengths and weaknesses of the different models" (259). However, Lenhard and Winsberg note, "In their voluminous 1998 review of AMIP, Gates et al. conceded that there were still errors revealed—but not accounted for—by the intercomparison" (259). Lenhard and Winsberg say that in AMIP such diagnoses were achieved only to a limited degree and largely had to be postponed (259). Moreover, according to Lenhard and Winsberg, the situation did not improve all that much by the time the first two phases of CMIP were undertaken (around the year 2000). They go on to say that, following CMIP2, "One of the central original goals—deepened understanding of simulation mechanisms via attribution—was greatly downsized, indeed disappeared nearly entirely from the proposals of the [then-]recent CMIP3" (259).[6] Similarly, with APE, an intercomparison effort which imposed more boundary conditions and therefore simplified the GCMs, the scientists' goal to understand "the causes of differences in model performance…[was] postponed to a later stage (see APE, 2008)" (259). This brief description represents virtually all of the empirical evidence presented by Lenhard and Winsberg to show that climate scientists failed to diagnose model errors.

While Lenhard and Winsberg grant that the sources of *some* model errors were tracked down throughout these intercomparison efforts, they regard the attribution of model error as remaining largely out of reach and suggest that such limitations will persist going forward. From their perspective, such "failures seem to point to a systematic cause that pushes analytic understanding of these models out of reach…this failure is best understood as a form of confirmation holism arising from the need modelers face to adapt their efforts, often with kludges, to generatively entrenched features of GCMs" (259). In agreement with my analysis, Frigg et al. (2015, 967) read Lenhard and Winsberg as defending "the more radical claim that one will never be able to say where the successes and failures of climate models come from."

2.3 Inconsistency, obscurity, and mismatch

In sum, analytic understanding is argued to be unachievable due to fuzzy modularity, kludges, and generative entrenchment, which are all claimed to be features of GCMs and their development. This argument is supplemented with some examples from the climate model intercomparison literature. On Lenhard and Winsberg's view, then, scientists cannot diagnose model errors.[7]

[6] Here "attribution" refers to attributing the sources of success and failure in climate models to sub-components of those models. This should not be confused with detection and attribution work in climate science.
[7] Lenhard and Winsberg's account also implies that scientists cannot attribute sources of model success, however, that is the topic for another paper.

However, there are several problems facing Lenhard and Winsberg's account. I will highlight and explain three of them here.

The first problem is one of inconsistency. Lenhard and Winsberg themselves admit that some errors were tracked down, as mentioned above in section 2.2. This is obviously not consistent with the radical claim they seem to be defending, as articulated at the end of section 2.2 above, i.e., the "claim that one will never be able to say where the successes and failures of climate models come from" (Frigg, 967).[8]

A second problem is about obscurity. That is, it is unclear what counts as analytic understanding on Lenhard and Winsberg's view. According to Lenhard and Winsberg (2010), to have analytic understanding is to be able to "identify the extent to which each of the sub-models of a global model is contributing to its various success and failures" (258). However, this "extent to which" language is somewhat obscure and difficult to apply in practice, i.e., when looking at examples of error diagnosis in the climate science literature. To see this, let us briefly look at a recent high-profile example of error diagnosis. In a contemporary, single-model study, scientists at the National Center for Atmospheric Research iteratively ran their model nearly 300 times to determine why the model's surface temperature output was too high when initialized with new emissions input data (including greenhouse gas and aerosol emissions data).[9] They ran the model "with varying configurations and outputs" and ultimately arrived at a diagnosis: "the cloud production components of the model were the *primary cause* of output changes, as cloud generation is tied to the presence of aerosols within the atmosphere" (Mayernik 2021, emphasis added; see also Hoesly et al. 2018 and Gettelman et al. 2019). Gettelman et al. (2019) also detail how model behavior is impacted by changes to specific sub-models. These scientists are aware not only of the changes made to their model as it underwent development, but also the various sources of observational

[8] Thank you to an anonymous reviewer for prompting me both to think through these issues more carefully and to explicitly highlight this inconsistency.
[9] This episode has a fairly broad audience, as it was written up at the *Wall Street Journal* (Hotz 2022). Additionally, Castillo Brache (2022) uses this example to critique Lenhard and Winsberg's (2010) account.

and theoretical evidential support for the sub-models (e.g., see Bogenschutz et al. 2013; Gettelman & Morrison 2015; Gettelman et al., 2015). It is unclear, however, whether Lenhard and Winsberg would regard this example as demonstrating analytic understanding. For example, they could claim that the modelers only identified model components (e.g., the cloud sub-model) that produced certain results *in conjunction with* the rest of the model, and that we can't say for sure whether the cloud sub-model itself is truly to blame and, if so, whether it is 100% to blame, 50% to blame, etc.[10] In other words, Lenhard and Winsberg could argue that, while this case exemplifies some sort of helpful analysis, it does not amount to showing the "extent to which" certain sub-models contributed to model error. If this is the right way to understand Lenhard and Winsberg, then this response seems available to refute any alleged example of error diagnosis. This would imply that climate model error diagnoses which appeal to specific model components are impossible in principle because one could always respond along holist lines and one could always question whether an identified error source is the primary culprit, a secondary (lesser) cause of error, and so on. There would be no need to even look at the scientific literature or to attempt to acquire empirical evidence of error diagnoses in practice. However, since Lenhard and Winsberg (2010) themselves consider empirical evidence by looking at the climate model intercomparison literature (see Section 2.2 above), they clearly *do not* want to rule out the possibility of error diagnosis in this way.

In light of the above analysis, and because they do not offer any detailed positive examples of error diagnosis, Lenhard and Winsberg's notion of analytic understand remains obscure.[11] I suggest that philosophers of science instead focus on the strategies scientists use to diagnosis (or ostensibly use to diagnose) model errors, the associated explanations scientists offer (if any), and determine what type(s) of understanding this practice amounts to in climate modeling.

The third problem is one of mismatch. That is, the examples Lenhard and Winsberg (2010) discuss all come from climate model intercomparison projects which involve *dozens of distinct models*

[10] Thanks to an anonymous reviewer for prompting me to think more critically about this.
[11] They also do not offer any detailed positive examples of attributing sources of model success.

and yet their version of confirmation holism is a skeptical claim about scientists being unable to achieve analytic understanding of *individual* climate models. This is because their argument is rooted in alleged features of individual GCMs, such as fuzzy modularity, kludges, and generative entrenchment. However, in the context of climate model intercomparisons, the failure to diagnose model error could also be explained by features of the intercomparison effort itself. Thus, I claim that there is a social epistemology element to the problem of error diagnosis—it's not just about simulations governed by complex intermingled computer code. Let's explore this idea further.

3. Model intercomparisons old and new

Here I show that features of model intercomparison practices, rather than the features of climate models that Lenhard and Winsberg focus on, may better explain difficulties in diagnosing model error. Recognizing this allows us to give a more fine-grained account of how error diagnosis should be approached in future analyses of climate models.

I contrast the early and informal model intercomparisons (section 3.1) with those which began circa 1989 with AMIP (section 3.2).[12]

3.1 Early and informal climate model intercomparisons

Climate model intercomparisons were informally conducted at least as early as the 1970s, during which time computationally simpler and more understandable models were compared to GCMs. While agreement between the more understandable simpler models and the more complex GCMs was taken to be epistemically significant (e.g., see Schneider and Dickinson 1974, 456), diagnosis of model differences also sometimes figured into climate scientists' analysis, e.g., differences in representation of both radiative processes and atmospheric stratification at the poles figured into an analysis of why 1-D models diverged from a GCM in their estimate of climate sensitivity (see Schneider 1975).

[12] For further historical reading, see Gates 1979; Arakawa 2000; Washington 2006; Edwards 2010, 2011; Randall et al. 2018; Weart 2020.

Further climate model intercomparisons were made in 1978, at the Global Atmospheric Research Programme conference in Washington, DC where scientists met to discuss, present, and compare climate models and modeling results. This was "the first of many 'intercomparison' meetings" (Weart 2020, 21), and included 81 scientists from 10 countries. Comparisons between a single GCM and one or two simpler models were presented, and further model-model discrepancies figured into many presentations (Gates 1979). Additionally, at this conference, climate scientist Stephen Schneider suggested a possible "first law of climate modeling" to ensure that only one change at a time be made when constructing hierarchies of climate models, so that cause and effect relationships would be understandable (Schneider 1979). As Schneider put it:

> …[T]he field of climate modeling needs to "fill in the blanks" at each level in the hierarchy of climate models. For only when the effect of adding one change at a time in models of different complexity can be studied, will we have any real hope of understanding cause and effect in the climatic system. The comparison, both across the hierarchy of models and with [independent] data…can provide improved confidence in the sensitivity performance of a model. In essence, we can conclude by stating what could be called a "first law of climate modeling." That is: To use climatic models to understand cause and effect linkages in the climatic system, it is necessary to make no more than one change at a time in a model, be it a boundary condition, numerical scheme, or physical parameterization. (1979, 748, original emphasis)

This "first law" was implicitly followed (and still is) in some cases of model development and in perturbed physics ensembles (in which a single parameter is varied across a range of plausible values) but is not true of the multi-model intercomparisons such as AMIP, where GCMs differ from one another in a multitude of ways.[13] I will return to this point in section 3.2 below.

In the 1979 Charney Report, which compared results from two structurally different GCMs (and some simpler models) there weren't any in-depth model error diagnoses. However, the authors did highlight model differences at a coarse level and, regarding global-scale changes under projections of increasing CO_2, they noted that "CO_2-induced climate changes made with the various models examined

[13] For more on climate model hierarchies, see Held (2005) and Jeevanjee et al. (2017).

are basically consistent and mutually supporting… [and] differences in model results are relatively small and may be accounted for by differences in model characteristics and simplifying assumptions" (National Academy of Science 1979, 17). These two GCMs came from two research groups, one model was developed by Syukuro Manabe and colleagues at the National Oceanic and Atmospheric Administration and the other was developed by James Hansen and colleagues at NASA Goddard Institute for Space Studies.[14]

The GCM used by Hansen and colleagues was also the subject of an intergenerational model intercomparison a few years later, in 1983. By "intergenerational intercomparison" I mean the evaluation of a GCM during and after model development—the comparison between an earlier and later version of a model. Hansen et al. very explicitly evaluate the changes in model output as a function of singular changes to the model physics, i.e., to the model's parameterizations, as they developed their "model II" from "model I" (see Figure 1 below). Note that such intergenerational intercomparisons of a single GCM with its predecessor is a common practice in climate modeling for model developers today (e.g., see Neale et al. 2012; Danabasoglu et al. 2020).[15]

[Insert Figure 1 here – for pre-print version, see end of document]

Thus, a defining feature of these early model intercomparisons is that they were between a relatively small number of models. Moreover, in these intercomparisons some diagnoses of model error (and model behavior more generally) were in fact possible. Finally, these model intercomparisons were not coordinated, in contrast to AMIP.

3.2 Coordinated Model Intercomparisons

[14] These two GCMs were configured in a total of five different ways (e.g., varying in terms of how snow and ice were represented, whether a deep ocean was used, and whether seasonal change was represented) to make five distinct projections.
[15] These exploratory activities fall under what Wilson (2021) refers to as "Model dynamic exploration."

With the Atmospheric Model Intercomparison Project (AMIP), which began in 1989, intercomparison practices changed dramatically. AMIP was "coordinated" in the sense that: (i) each participating modeling group was required to run its model according to certain boundary conditions, in this case, sea surface temperatures and sea ice extent were prescribed from observational data; (ii) each modeling group had to submit their model output data in a specified gridded format to facilitate model-model and model-observation comparisons; and (iii) each modeling group had to submit data for specified variables over the prescribed time period (e.g., monthly averages at each grid point for sea-level pressure for the years 1977-1988) (Gates 1992).

Despite this coordination, differences between models (i.e., concerning how they were developed, what their resolutions were, what parameterizations they used, etc.) were *not* systematic or prescribed. The different modeling groups didn't coordinate with the other modeling groups about how to build their respective models in systematically different ways to explore structural model uncertainty in a principled fashion. For these reasons and others, the multi-model ensembles that began with AMIP and now continue to today in various forms, are often referred to as "ensembles of opportunity" (Tebaldi and Knutti 2007). Moreover, with AMIP, 31 modeling groups participated in total, "representing virtually the entire international atmospheric modeling community" at the time (Gates et al. 1999, 29). Thus, instead of comparing one or two GCMs to each other and to simpler models, the coordinated model intercomparison projects involve dozens of models (and now, around 100 models) hailing from a growing number of institutions.

These realities of scientific practice are important for understanding why model error diagnosis was more difficult to achieve than anticipated in the examples described in section 2.2 above. These realities include the increasing number of participating models, the messy relationships between these models, and the increasing number of model developers and developing centers.

First, AMIP, and the many other coordinated model intercomparison projects that followed involved more models than previous intercomparisons (31 atmospheric GCMs being jointly analyzed in

AMIP vs. a handful of GCMs being analyzed one at a time at the 1978 conference). Second, the relationships across the AMIP models were neither hierarchical nor systematic—they have diverged from the prescriptions of Schneider's "first law." One clear example of this is in chapter 9 of the Intergovernmental Panel and Climate Change's fourth assessment report, where different treatments of aerosols are described (Hegerl et al. 2007, see especially their figure 9.5). Instead of a hierarchy of models which differ from one another only with respect to aerosol representations, these models also exhibit structural differences (e.g., in terms of which processes are omitted vs. parameterized), differences in resolution, and others. Third, individual model development knowledge is epistemically dispersed across multiple teams because models consist of multiple modules and dozens or more process representations (sub-models) requiring experts from a diverse range of fields (e.g., see National Research Council 2012).

More generally, the conceptualization, implementation, tuning, and testing that goes into building a particular state-of-the-art GCM is not fully known by any individual scientist on the development team, let alone scientists working at other modeling institutions. In other words, the facts of model development (e.g., concerning which parameterizations were used for various processes and how they, or other parts of the model, were tuned, measured, and empirically or theoretically supported) were more widely epistemically dispersed than previous model intercomparisons, largely as a consequence of there being more GCMs and more scientists working to develop them.

Until fairly recently, climate model tuning (also known as model calibration) was a fairly opaque and under-discussed practice.[16] Tuning involves adjusting parameters or individual model components in order to improve the fit with observational data of interest. Model tuning is sometimes discussed as a hindrance to determining model skill—the worry is that a model which performs well is doing so for the wrong reasons, i.e., that a models parameters/components were adjusted *without sufficient justification*

[16] For examples of candid discussions of model tuning by climate scientists, see Mauritsen et al. 2012; Schmidt and Sherwood 2015; Schmidt et al. 2017; Hourdin et al. 2017.

and only in order to fit observations.[17] As Parker (2018b, section 4.2, par. 6) notes, matters become more complicated when one considers that more generally (i.e., aside from actually tuning the model), "modelers can be familiar with [certain observational] data and may well make choices in model development—choices which could reasonably have been somewhat different—with the expectation that they will improve the model's performance with respect to those already-seen data." In the context of difficulties facing error diagnosis, the main issues are that each modeling group tunes their GCM at least somewhat differently, the way a model is tuned may impact its biases, and the knowledge of how a given GCM was tuned largely remains local to that model's home institution.

It's also worth noting that the climate modeling community was fairly small in the early days (e.g., see Edwards 2010; 2011), such that individual scientists could claim to know all the ins and outs of their GCM and potentially compare it with their colleague's model by discussing it one-on-one. The fact that GCMs continued to increase in complexity (i.e., increasing the number of physical processes represented by adding more and more sub-models) while the climate modeling community also grew, means that the expertise required for diagnosing errors because more and more dispersed and diagnosing model errors likely became much more challenging.[18]

These features of scientific practice shed some additional light on why diagnosing model errors may have been so difficult in the examples Lenhard and Winsberg (2010) discuss. Imagine trying to tease apart every single difference between each GCM. Even if the models individually were fully understood by the scientists who developed them, we would expect difficulties in diagnosing model-model discrepancies during intercomparison because inter-model differences were so numerous. Moreover, the iterative re-running of a GCM hundreds of times (recall the example from section 2.3 above) to conduct a sensitivity test is not an option in the multiple model context, or at least it is not at all clear how to

[17] See Steel and Werndl (2013), Frisch (2015), and Schmidt and Sherwood (2015) for a philosophical discussion.
[18] The analysis in Cess et al. (1989) serves as a sort of midpoint between the uncoordinated model intercomparison and the coordinated ones. This intercomparison included some closely related models (i.e., from the same institutions) as well as more distinct models and analyses of the former were more fine-grained than those of the latter (e.g., see their discussion of GFDL I and II on their page 515). Moreover, many of the scientists involved helped develop the models being analyzed.

conduct one given numerous and nonsystematic inter-model differences. Failing to diagnose model disagreement in AMIP was thus underdetermined—perhaps the failure was due to individual model complexity, but it also may have been due to the dispersal of facts across hundreds of practitioners concerning how the different models were developed, tested, etc.

There are additional factors that could explain the failure to diagnose errors in AMIP, making the issue even more underdetermined. E.g., there was a limitation of available observational data to compare model simulations against (e.g., see Gleckler et al. 1995, 793). This could have hampered error diagnosis efforts: e.g., if scientists thought a particular model-observation discrepancy was caused by X, and X is thought to impact the simulation of Y, then a lack of observational data to compare Y against is a major problem. Perhaps another relevant factor was the comparative ease of compiling output data from the models (which was then becoming available in a uniform format) and analyzing the statistical features of the whole model ensemble. The thinking could be: "why diagnose the causes of model disagreement when we can easily aggregate and statistically analyze the model results?"

Climate scientists and philosopher of science Touzé-Peiffer et al. (2020) reinforce the point I am making. They analyze the history of the coupled model intercomparison project (CMIP) and its structural effects on climate research. In their analysis, Touzé-Peiffer et al. characterize a climate model as "not just the sum of the code" and associated assumptions, but as a "dynamical entity with which it is possible to interact" (9). By this, Touzé-Peiffer et al. mean that through the trial-and-error use of a climate model (initialize it, run it, compare it to observations and other model output, make tweaks to the model, repeat) "climate scientists can acquire …knowledge about the behaviour of a climate model, what it is doing and why" (9). This knowledge is *collective*, resulting from collaborative efforts of scientists working within a single modeling institution who focus on "separate but complementary aspects of the same climate model" (9).

Touzé-Peiffer et al. further claim that if knowledge about a given climate model is collective, it typically stays at the level of one research team working on one model. Indeed, as they note, "due to the complexity of the models involved in CMIP, acquiring knowledge about the behavior of a climate model takes time and scientists generally focus their efforts on one particular model" (9). Under these circumstances, it would be unsurprising for model error diagnosis in a case such as AMIP to be severely limited, as such diagnosis would require the synthesis of several dispersed sets of collective knowledge about each GCM under consideration.

However, I think it is fair to ask: *was there really* such a failure to diagnose model error as Lenhard and Winsberg suggest? In fact, AMIP spawned 26 diagnostic subprojects aimed at analyzing the various sources of model error and model differences, and several of these subprojects *were* successful in identifying some sources of model error.[19] In the next section we consider two examples from these diagnostic subprojects, and then we look at two contemporary examples of model error diagnosis.[20]

Before proceeding, I should note that several philosophers and other scholars of climate modeling (e.g., Frigg et al. 2015; Baumberger et al. 2017; Carrier and Lenhard 2019; Touzé-Peiffer et al. 2020) have also responded to Lenhard and Winsberg (2010) by pointing out clear examples of error diagnosis in the climate modeling literature. I will not merely be adding to these examples: I will also explore the different *strategies* scientists use when making these diagnoses and I will explore the possibility of an error repertoire for climate modeling (Section 5 below).

4 AMIP-era and contemporary examples of successful model error diagnosis

4.1 Isolating cloud radiative effects using observational data

[19] A list of publications from these diagnostic subprojects can be found here: https://pcmdi.llnl.gov/mips/amip/abstracts/abhme.html

[20] Touzé-Peiffer et al. (2020) also give examples of successful model error diagnosis, saying "In fact, in the literature, we can find many studies investigating the link between the results of a model and its parameterizations (e.g., Hourdin et al., 2013; Notz et al. 2013)." They also mention "studies comparing radiation codes in different climate models, such as Oreopoulos et al. (2012) and Pincus et al. (2015), where the authors analyze not only the model results, but also the corresponding parameterizations and the assumptions they make" (9).

First, there is Gleckler et al.'s (1995) study, in which scientists attribute differences in derived ocean heat transport across 15 GCMs to differences in how these models represent cloud radiative feedbacks.

These scientists use results from model simulations of radiative fluxes at the surface of the ocean to calculate what ocean heat transport (from the tropics to the poles) would look like in each of the models if ocean surface temperatures weren't prescribed.[21] They find that calculated ocean heat transport in some of the GCMs is in the wrong direction for some latitudes—i.e., Northward in much of the Southern Hemisphere. They suspect that cloud feedbacks were relevant to this discrepancy based on previous modeling results (i.e., Cess et al. 1990).

To investigate whether cloud feedbacks *really were* the culprit for this discrepancy, Gleckler et al. calculate cloud radiative forcing both in the models and in observations. Cloud radiative forcing is defined as the difference between net top-of-the-atmosphere [TOA] radiation and a "clear sky" (i.e., without clouds) TOA radiation (Ramanathan et al. 1989). They find important differences in observation-derived and model-derived cloud radiative forcing, as well as differences across the models. Moreover, they find that the strength of cloud radiative forcing correlates with ocean surface radiative fluxes both in models and in observations (they explain why this is to be expected based on certain TOA and surface energy budget equations; see Gleckler et al. 1995, 791-792). From this they suggest that the GCMs' "inadequate simulations" of cloud radiative forcing are to blame for the discrepancies between calculated ocean heat transport in the models and in observations (794). To informally test this, they recalculate ocean heat transport using a combination of model data and cloud forcing "corrections" from observational data. The resultant ocean heat transport is no longer in the wrong direction in the southern hemisphere, which these scientists take as a positive sign that their error diagnosis was correct.

[21] Recall: in AMIP, sea surface temperatures *were* prescribed. But these scientists still wanted to know what this heat transport would look like because future applications of these models would include coupling them to ocean models.

While the analysis does not go model by model and look at how each individual GCM represents cloud radiative forcing, their analysis does diagnosis a cause for why models disagreed with known data. They began with a certain expectation about the source of model error and then used physical reasoning (using energy budget equations and finding a correlation between cloud radiative forcing strength and ocean transport), and finally they tested their diagnosis.

4.2 Using dimension reduction techniques

Second, there is Sengupta and Boyle's (1997) analysis which employs a dimension reduction technique to compare GCMs both with observations and with one another. This technique, common principal component analysis, allows scientists to reduce the dimensionality of data while preserving as much variance as possible. Scientists compute a few of the largest orthogonal (i.e., independent) components that maximally preserve the original variance of the data. These components are assumed to be statistically representative characteristics of the original data. In this way, they can compare the identified components of different data sources and show whether and how model output and observational data are similar, as defined with the components. In one part of this study, Sengupta and Boyle look at the differences in 200-hpa (atmospheric pressure) output from four GCMs compared to observations. This subset of models "a priori were expected to have some common type of error patterns," because the models all started from the same code (1997, 826). Of the four models, all but one used the same convective parameterization (a sub-model which calculates the effects of convective clouds, which form through vertical motion of humid air parcels). The authors note that one may expect that "the convective parameterization might play an overwhelming role in determining the model characteristics," (826) and thus that the models which shared this parameterization would be grouped together (i.e., have the same principal components "explaining" their variance). However, this turned out not to be the case and other model differences (i.e., two of the models represented land-processes and radiation differently) apparently were more important reasons for why those models differed from the observational data. In this way, they were able to identify specific sources potentially responsible for model-model and model-observation discrepancies.

These two examples show that in AMIP climate scientists did point to specific aspects of models as the source of model error. In the Gleckler et al. example this involved physical reasoning about the effect of clouds on Earth's energy budget, and in the Sengupta and Boyle example dimension reduction techniques were utilized. The next two examples are more contemporary.

4.3 Utilizing background knowledge and assessing dynamic simulations in regional climate models

Our third example concerns a study of regional climate models (RCMs) and comes from Bukovsky et al. (2017). These scientists look at RCM mean model output of projected changes in spring and summer precipitation in the southern great plains in the United States. These RCMs are driven by (i.e., fed input data from) four different GCMs at their boundaries. The RCM results are compared and differences in the driving GCMs and some GCM projections were also analyzed.

Regarding the GCM comparison, Bukovsky et al. draw from past modeling studies to suggest that for two of the GCMs, "it is likely that the projected increase" in precipitation by these GCMs is due to the type of convective parameterization scheme used by both GCMs (8281). While this diagnosis makes physical sense based on the process of convective precipitation, Bukovsky et al. also note that *a characteristic response* of this convective parameterization scheme is to "convect too easily to allow CAPE [convective available potential energy] to build up in the environment (as illustrated by consistently low CAPE values in [the Community Climate System Model] CCSM in Marsh et al. (2007))" (8283). They further note that similar problems have been discovered in previous analyses (e.g., Zhang and McFarlane 1995; Zhang 2002). Thus, a known behavior of a specific sub-model (the convective parameterization) is identified as likely to be causally relevant to the GCM's too-high projection of precipitation. Here the diagnosis is tentative, but the authors explicitly make a connection between the behavior of a parameterization and the consequences of that parameterization's behavior for the climate model projection, i.e., certain precipitation patterns.

In the same study, Bukovsky et al. also look at RCM projections and tie the differences to their respective driving GCMs. There is a discussion of an outlier: the RCM projections driven by one of the GCMs (namely, HADCM) give a very different picture concerning changes in the upper-level jet stream compared to the RCM projections driven by the other GCMs.

Bukovsky et al. identify the cause of this discrepancy as the simulation of the jet stream in HADCM and HADCM-driven RCMs. They note that the jet stream "is not realistically simulated to start with over North America, so the changes do not represent changes to a realistically simulated phenomenon. It is too weak, positioned incorrectly, and does not evolve properly through the summer" (8286). In other words, the poor performance of HADCM in simulating jet streams in a control scenario was used to explain (and was thought to be causally relevant to) the poor performance of the HADCM-driven RCMs in the climate change scenario. In this case, the error diagnosis involved pointing to the incorrect or inaccurate dynamic representation of a process and its consequences.[22]

4.4 Focusing on singular model differences in a small geoengineering modeling intercomparison

A fourth example is found in the Geoengineering Model Intercomparison Project (GeoMIP), in which GCMs simulate climate scenarios with decreased incoming solar radiation to offset warming from continued increases in CO_2 concentrations. Pitari et al. (2014) evaluate GCMs simulating stratospheric aerosol injections (i.e., spraying SO_2 into the stratosphere) as specified under two different GeoMIP experiments, paying particular attention to model projections of ozone. What is striking about their analysis is that they only focus on four models, and they give an in-depth characterization of the features of each model, as well as the differences between the models (see Pitari et al. 2014, 2631). Recall the explanation in section 3 above of why error diagnosis was so difficult in AMIP: there were too many models which differed from one another non-systemically and knowledge of individual model behavior and development was widely dispersed. One way to address this is to intercompare smaller numbers of

[22] For philosophical discussions of dynamical sufficiency in modeling (which concerns the representation of how a system changes over time) see Lloyd et al. (2008) and Kawamleh (2022).

models and to include relevant model developers in the intercomparison analysis. In the below example, we can see the payoff of evaluating models in this way.

A key uncertainty in modeling stratospheric aerosol injections concerns representations of aerosol chemistry and aerosol microphysics due in part to insufficient observational data (Kravitz and MacMartin 2020). Thus, it is important for Pitari et al.'s analysis to highlight the differences in aerosol microphysics representations across models. For example, they note that only one model "includes a module for aerosol microphysics for the explicit prediction of the aerosol size distribution" while the "other models prescribe fixed aerosol size distributions" (Pitari et al. 2014, 2631). Further details about aerosol characteristics in the models are then given. As we'll see in more detail below, crucial to their analysis is that only one of the four models omits the representation of heterogeneous chemical reactions on the surface of sulfate aerosols.

Pitari et al. also describe model diagnostics from previous modeling studies on projections of ozone depletion and ozone mixing ratios compared to observational data. They note several strengths and limitations of the models related to ozone, e.g., how "all models agree well" with the satellite observational data concerning ozone levels in the tropical lower stratosphere between 100 and 30 hPa, as well as limitations, e.g., how at "altitudes above 7 hPa [two of the models] slightly overestimate the observations" (2635). Pitari et al. conclude their description of model diagnostics:

> A full set of diagnostics covering radiation, stratospheric dynamics, transport and chemistry, upper troposphere and lower stratosphere features, natural variability and long-term projections of stratospheric ozone, and stratosphere-troposphere interactions, have been used in previous intercomparison projects developed in the context of WMO [World Meteorological Organization] activities. These diagnostics enabled the use of the participating models as tools to predict the future evolution of stratospheric ozone and for future sensitivity studies and climate change scenarios… (2636)

The above alludes to how much background knowledge about the models being evaluated was seriously considered by these scientists. This background knowledge includes not only facts about model components such as aerosol chemistry representations etc., but also about past model performance. The importance of expert background knowledge in understanding climate model evaluations has been noted elsewhere in the philosophical literature (e.g., Winsberg 2018; Jebeile and Crucifix 2020) and is also evident in some of the other examples of model error diagnosis discussed above.

This expert background knowledge was brought to bear in a very detailed example of model error diagnosis, which relates to how atmospheric chemistry is represented by the each of the models. More specifically, Pitari et al. note that all three “models with heterogeneous chemistry simulate a significant increase in ozone depletion in the Antarctic region” and they attribute this to “a combination of increasing sulfate aerosol [surface area density] …and enhanced formation of [polar stratospheric clouds] produced in turn by local adiabatic and nonadiabatic cooling…the latter due to the feedback of photochemical ozone losses” (2645). In contrast, one of the models “does not include heterogeneous chemistry on the surfaces of the aerosols,” and, so the “missing heterogeneous chemical reduction” of nitrogen oxides on aerosol surface area density “does not allow in this model a limitation of the ozone loss above 50 hPa” (2645). They continue by explaining that this ozone parameterization difference leads to polar temperature decreases that exceed that of the other models (at least above 50 hPa).

We thus have yet another example of model-model discrepancy being diagnosed. Here the interesting features include a small number of models, a sophisticated level of physical reasoning which relates model components to model output which is likely only possible because of the expert background knowledge about the models in question, as well as knowledge of their past performance.

5. Forward: An error repertoire for climate modeling

From section 4 above it should be clear that model error diagnosis is not only possible, but also practiced. Based on the scientific literature reviewed above, error diagnosis is conducted with varying degrees of both precision and confidence, and the explanations that result may sometimes only be comprehensible to other experts (e.g., the diagnosis in Pitari et al. 2014). Recall that Lenhard and Winsberg argue that error diagnosis is not possible due to the characteristics they take to be part and parcel of climate models: generative entrenchment, fuzzy modularity, and kludges. Yet, a more grounded argument runs in the opposite direction: we begin with successful examples of error diagnosis, such as those described above, and see what we can learn from them. With the above examples of error diagnosis in mind, let's take a step back for a moment and think about error diagnosis in broader terms.

In the introduction to her 1996 book, *Error and the Growth of Experimental Knowledge*, philosopher Deborah Mayo discusses everyday strategies that humans use to detect errors in the world around us. Summarizing and slightly modifying the terminology used in Mayo (1996, 4-7), two that stand out as relevant to our discussion are:

> **(i) Building and consulting a list of errors that are expected or commonly encountered.** E.g., the last time the coffee maker didn't work, it was because I forgot to fill it with water. Perhaps that's the case this time, too.
>
> **(ii) Recognizing errors based on their plausible effects and identifying instances of those effects.** E.g., if my car's tire pressure is too low, one likely effect is that my gas mileage will be worse. Given my bad gas mileage on yesterday's trip, I should check the tire pressure.

Both of these strategies are part of what we can call an "error repertoire". While Mayo (1996) restricts the specific notion of an "error repertoire" to (i), we can broaden the notion to include (ii), and we can also include other specific strategies that scientists use to diagnosis model error, such as those documented above.

Both (i) and (ii) are exemplified in section 4 above. In the Gleckler et al. example, it was anticipated that differences in cloud parameterizations would be a source of error. As they note, atmospheric GCMs "are known to disagree considerably in their simulations of the effects of clouds on the Earth's radiation budget (Cess et al. 1990), and hence the effects of simulated cloud-radiation interactions on the implied meridional energy transports are immediately suspect" (Gleckler et al. 1995, 793). Similarly, in Bukovsky et al. (2017), *previously known* behaviors of different convective parameterizations are identified. These expectations, combined with physical reasoning about convective precipitation, allowed Bukovsky et al. to identify a source of anomalous model behavior. They also had reasons to expect regional models driven by HADCM to perform poorly when it came to simulating changes to the upper-level jet stream: those regional models did a poor job of simulating that process (when driven by HADCM) in the first place!

The point here is that scientists expect certain broad types of errors even before they occur, and the effects of errors can provide clues to their source(s). In some instances, scientists' expectations may be based on tacit expert knowledge, e.g., concerning the idiosyncratic behavior of a particular convective parameterization based on its construction or past uses. This convective parameterization may have a known impact on modeling results (e.g., a telltale bias in precipitation trends), thus providing a further clue.[23] In other instances, expectations may be informed primarily by climatic knowledge, e.g., background about the impact of clouds on the earth's energy balance (based on observations and theory) which may lead scientists to anticipate certain types of errors related to cloud parameterizations.[24]

Lenhard and Winsberg may respond by saying that these examples are too speculative to count as error diagnoses that demonstrate analytic understanding (setting aside, for a moment, the obscurity of this notion highlighted in section 2.3 above). Indeed, Lenhard and Winsberg may say "sure, scientists have hunches and arguments to support them, but this is not the same as definitively saying exactly why a model erred by pointing to a specific model component." Note that this is stronger than Lenhard and Winsberg's original skeptical claim about error diagnosis, but I believe the weaker skeptical claim—that model errors simply cannot be diagnosed because scientists are unable to say where the sources of model failure come from—has been debunked by the examples given in section 4 above. One reply to this stronger skeptical claim is to note that there are no guarantees in science, so "definitive" is an inappropriate standard. It is also worth noting, however, that other cases of error diagnosis *do* seem definitive, at least based on the language used by scientists, especially the descriptions used by Pitari et al. (2014) in describing one model's ozone parameterization and in Bukovsky et al.'s description of the upper-level jet stream simulation. Moreover, in the Sengupta and Boyle example, the influence of

[23] E.g., see Sun et al. (2006); Birch et al. (2015).
[24] Examples of early work on clouds in relation to the Earth's radiation budget include theoretical work (e.g., Schneider 1972) and observational work (e.g., Hartmann and Short 1980).

identified common principal components are quantified, which, while perhaps not "definitive," is nonetheless very specific information about divergences in model behavior.[25] Of course, this doesn't automatically mean that these diagnoses *are* definitive (quantitative or not), but they have at least passed muster as required by peer review and they are clear examples of *scientists* expressing the view that a given model's error(s) are, at least in part, attributable to a particular model subcomponent.

One may still insist that scientists are often too loose with their diagnoses, e.g., saying that a particular model error results from poor representations of clouds (an admittedly common "diagnosis") doesn't provide us with details explaining the exact extent to which, or way in which, a specific cloud parameterization leads to such an error. While such information may be difficult to acquire, scientists do have some methods at their disposal that are superior to the loose diagnosis that "the clouds are to blame." More specifically, in some cases, error diagnoses can be tested by postulating that, e.g., "if X is the cause of this discrepancy, then we expect to also find A." We see something like this in the example from Gleckler et al. in section 4.1 above. The cause of the discrepancy was thought to be GCMs' poor simulations of cloud radiative forcing, and one expectation of this was that substituting observation-based cloud radiative forcings would correct for the discrepancy (i.e., would result in agreement across models and between models and observations for inferred ocean heat transport). They found that the substitution did result in a correction, thereby providing additional evidence that their diagnosis was correct.

Based on the discussion so far, we may be able to make some recommendations for how error diagnosis can be fruitfully applied in climate modeling intercomparisons. Some strategies may be relatively straightforward to apply, and indeed, are likely commonly applied in practice.[26] These include, for example, employing reasoning about known physical relationships, making use of tacit expert

---

[25] See Kuo et al. (2020) for a recent statistical analysis of models which differed in their deep convective parameterizations. So-called "process-level" analyses which use statistical methods as well as physical arguments also becoming more common (e.g., see Maloney et al. 2019).

[26] Indeed, the practice of tinkering with a single model over the course of model development and iteratively making changes may also involve error diagnosis (e.g., see Hansen et al. 1983; Danabasoglu et al. 2020; Mayernik (2021), although such a strategy may only work for single-model evaluations.

knowledge concerning previous model behavior, and using dimension reduction analysis to identify explained variance.

However, it may be worth considering whether scientists can construct an "error repertoire," as mentioned above, to guide error diagnosis in climate modeling. The idea would be to combine (i) and (ii) from above, along with several of the specific strategies scientists already use to diagnose model errors, to help diagnose model error more systematically.

The error repertoire I have in mind would consist of something like the following:

(a) A list of previously encountered model errors and the source(s) of those errors, with an explanation of how the error was detected (including which model output variables were used), how it was dealt with, and how localizable it was.

(b) A set of guidelines for doing error diagnostics in various contexts (e.g., single model, global multi-model ensemble, high resolution regional model ensemble, etc.). This might involve combining several of the strategies identified in section 4 above. E.g., a dimension reduction technique could first give a quantitative picture of which model components are (apparently) most responsible for model error. Then a physical explanation could be offered after analyzing the dynamical simulation of specified variables and whether they are sufficiently realistic or have telltale biases. Finally, a test could be done, to see of the suspected error source is indeed the culprit.[27]

(c) A deliberate effort to hypothesize about model errors prior to analyzing the model output. E.g., "we expect vegetation sub-model X to cause bias Y, which we should be able to detect by comparing several GCMs (some which have X, some which don't) to observations Z." If hypotheses about model errors are made prior to analyzing the results from model ensembles, error diagnosis can be conducted in a less post-hoc fashion.[28] Ideally, then, this would be completed before (b), directly above.

The above, I submit, would provide further opportunities for scientists to demonstrate an understanding of specific pieces of their models and how those pieces relate to model performance, akin to the "analytic" type of understanding that Lenhard and Winsberg claim is out of reach.

Granted, given the multitude of obstacles that make error diagnosis difficult (see sections 2 and 3 above), one may think it is not worthwhile (or even possible) to construct such a repertoire.[29] That is, given the complexity of current individual models, the idea that knowledge about a model is collective,

[27] A "crucial test" would be superior, i.e., a test which distinguishes between the primary suspected error source in question and the other suspected error sources.
[28] Thanks to Ben Kravitz for inspiring this suggestion.
[29] Thanks to an anonymous reviewer for emphasizing this point.

the increasing number of models, and the highly non-systemic relationships between these models, etc., we ought to be very skeptical that an error repertoire could be constructed in the first place. For this reason, I think an error repertoire could begin with just a few models and could begin by drawing from strategies *scientists already use* to diagnose model errors. Thus, analyzing model differences across a smaller number of models, as Pitari et al. (2014) did, may have multiple payoffs: it can allow for more in-depth analyses (as we saw in section 4.4 above), and it can provide a testbed for a climate model "error repertoire." The idea would be to intercompare three or four distinct models (from different institutions) using (a) – (c) above, with relevant climate model developers also weighing in to highlight important inter-model differences. I suspect this endeavor would yield benefits with respect to both the quality of error diagnoses, and to the understanding scientists' gain regarding their respective models. Unfortunately, this testbed strategy also comes with downsides: by focusing only on a small number of models, model structural error would be poorly sampled (i.e., it would be a very small "ensemble of opportunity" (Tebaldi and Knutti 2007)). Further, the direct benefits of this error repertoire would likely be limited to the specific models that are part of the testbed, and there are other challenges besides.[30]

However, there are also reasons to expect that an error repertoire (of some form – perhaps not the exact one I outlined) would be of genuine scientific interest. First, there is much interest in the recent "hot model" problem (Gettelman et al. 2019; Voosen 2021; Hausfather et al. 2022; see also section 2.3 above), which involves figuring out why some models which are more *realistic* are, at the same time, too sensitive to greenhouse gases (far more so than many other models). This research shows both that scientists really do care why their models give incorrect results and that there is currently no agreed upon framework to assess model error. Perhaps an error repertoire could be beneficial here. Second, there has

[30] A big challenge concerns resource availability. When presenting some of these ideas at [omitted for review], a climate modeler asked whether error diagnosis efforts should be focused on errors that have clear solutions vs. errors that are significant but difficult to understand or fix. Even if it is agreed that an error repertoire would be valuable, this doesn't mean that the resources are available to construct or implement one.

been a push to conduct "process-level" or "process-oriented" diagnoses of model biases (e.g., see Bukovsky et al. 2017; Maloney et al. 2019; Eyring et al. 2019).[31] In particular, Maloney et al. describe:

> [P]rocess-oriented diagnostics (PODs) that are designed to inform parameterization improvements to address…long-standing model biases (e.g., Eyring et al. 2019). A POD characterizes a specific physical process or emergent behavior that is hypothesized to be related to the ability to simulate an observed phenomenon (2019, 1665).

Their emphasis is on quantifying model biases systematically and ranking models across different metrics (i.e., across different variables related to processes of interest). This goes some way towards the error repertoire I described above, and my specific recommendations of looking at a small number of models, hypothesizing about model errors prior to analyzing model results, and testing suspected sources of model errors, can all complement (and potentially improve) the process diagnostics that Maloney et al. discuss.

In sum, based on the empirical evidence from model comparisons I've considered, I suggest that when we think about model error diagnosis in climate modeling, we should ask not *whether* model error diagnosis is possible, because it obviously is. In place of this black and white question[32], I have suggested questions such as: why is model error diagnosis so difficult? What methods do scientists use to diagnose model errors? How might error diagnosis be improved? Further, what does the practice of error diagnosis tell us about how (or whether) scientists understand their models?

From my analysis in this paper, we can take a significant step towards answering these questions. First, features of model intercomparisons are important for understanding why model error diagnosis is so difficult. Models inter-relate to one another in a highly non-systemic way and the number of experts required to understand a single model—never mind the 100+ GCMs now being used for research—means that knowledge of different sub-models, facts of model development, testing history, etc. is highly dispersed. Second, the methods scientists use to diagnose model errors include physical reasoning, iteratively running simulations making only small changes each time, employing dimension reduction

[31] This emphasis on process representations in climate models has also inspired some philosophical accounts, e.g., Lloyd et al. 2021; Kawamleh 2022.
[32] My suggestion here is influenced by Lloyd's logic of research questions (Lloyd 2015b) as well as van Fraassen's pragmatic theory of explanation (van Fraassen 1980).

techniques, forming expectations about model error based on past studies, utilizing expert knowledge of a specific model or sub-model's behavior, and testing error diagnosis by examining the consequences of correcting for the diagnosed error. Third, error diagnosis can be improved by constructing an error repertoire as outlined above and by intercomparing a few models at a time rather than dozens or more.

Finally, the practice of error diagnosis in climate model intercomparisons tells us that scientists do have some understanding of their models: they anticipate certain problems (e.g., related to convective parameterizations and to cloud representations) and they provide explanations as to why these problems occur. Some of these explanations and diagnoses may seem so esoteric as to not be worth philosophers' time. Indeed, in his recent book, Winsberg says,

> I think that when we look on the work of those who are in the business of modeling highly complex non-linear systems, the best we are ever going to be able to do is to arrive at a situation where "a simulation modeler could explain to his peers why it was legitimate and rational to use a certain approximation technique to solve a particular problem" by appealing to "very context specific reasons and particular features."[33]

However, there may be philosophical benefit in paying further attention to the working knowledge that climate modelers have about the behaviors of their models and trying to characterize what they are doing in broader terms. One way to do this is by examining how scientists diagnose, communicate, explain, and (hopefully) correct for errors in complex modeling.

6. References

Abraham, J. P., M. Baringer, N. L. Bindoff, T. Boyer, L. J. Cheng, J. A. Church, J. L. Conroy, et al. 2013. "A Review of Global Ocean Temperature Observations: Implications for Ocean Heat Content Estimates and Climate Change." *Reviews of Geophysics* 51 (3): 450–83. https://doi.org/10.1002/rog.20022.

Annan, J. D., and J. C. Hargreaves. 2010. "Reliability of the CMIP3 Ensemble." *Geophysical Research Letters* 37 (2). https://doi.org/10.1029/2009GL041994.

———. 2017. "On the Meaning of Independence in Climate Science." *Earth System Dynamics* 8 (1): 211–24. http://dx.doi.org/10.5194/esd-8-211-2017.

Arakawa, Akio. 2000. "A Personal Perspective on the Early Years of General Circulation Modeling at UCLA." In *General Circulation Model Development: Past, Present, Future*, edited by David A. Randall, 1–65. New York: Academic Press.

[33] Quotation marks pick out quotes from Goodwin (2015), pp. 342-343.

Baumberger, Christoph, Reto Knutti, and Gertrude Hirsch Hadorn. 2017. “Building Confidence in Climate Model Projections: An Analysis of Inferences from Fit.” *WIREs Climate Change* 8 (3): e454. https://doi.org/10.1002/wcc.454.

Birch, Cathryn E., Malcolm J. Roberts, Luis Garcia-Carreras, Duncan Ackerley, Michael J. Reeder, Adrian P. Lock, and Reinhard Schiemann. 2015. “Sea-Breeze Dynamics and Convection Initiation: The Influence of Convective Parameterization in Weather and Climate Model Biases.” *Journal of Climate* 28 (20): 8093–8108. https://doi.org/10.1175/JCLI-D-14-00850.1.

Bogenschutz, Peter A., Andrew Gettelman, Hugh Morrison, Vincent E. Larson, Cheryl Craig, and David P. Schanen. 2013. “Higher-Order Turbulence Closure and Its Impact on Climate Simulations in the Community Atmosphere Model.” *Journal of Climate* 26 (23): 9655–76. https://doi.org/10.1175/JCLI-D-13-00075.1.

Bukovsky, Melissa S., Rachel R. McCrary, Anji Seth, and Linda O. Mearns. 2017. “A Mechanistically Credible, Poleward Shift in Warm-Season Precipitation Projected for the U.S. Southern Great Plains?” *Journal of Climate* 30 (20): 8275–98. https://doi.org/10.1175/JCLI-D-16-0316.1.

Carrier, Martin, and Johannes Lenhard. 2019. “Climate Models: How to Assess Their Reliability.” *International Studies in the Philosophy of Science* 32 (2): 81–100. https://doi.org/10.1080/02698595.2019.1644722.

Castillo Brache, L. A. 2022. “Fixing High-ECS Models: The Problem of Holism Revisited.” In *Climate Sensitivity, Paleoclimate Data, & the End of Model Democracy* [Symposium]. PSA 28th Biennial Meeting, Nov. 10-13, Pittsburgh, PA, U.S.

Cess, R. D., G. L. Potter, J. P. Blanchet, G. J. Boer, A. D. Del Genio, M. Déqué, V. Dymnikov, et al. 1990. “Intercomparison and Interpretation of Climate Feedback Processes in 19 Atmospheric General Circulation Models.” *Journal of Geophysical Research: Atmospheres* 95 (D10): 16601–15. https://doi.org/10.1029/JD095iD10p16601.

Cess, R. D., G. L. Potter, J. P. Blanchet, G. J. Boer, S. J. Ghan, J. T. Kiehl, H. Le Treut, et al. 1989. “Interpretation of Cloud-Climate Feedback as Produced by 14 Atmospheric General Circulation Models.” *Science* 245 (4917): 513–16. https://doi.org/10.1126/science.245.4917.513.

Clark, Andy. 1987. “The Kludge in the Machine*.” *Mind & Language* 2 (4): 277–300. https://doi.org/10.1111/j.1468-0017.1987.tb00123.x.

Council, National Research. 2012. *A National Strategy for Advancing Climate Modeling*. https://doi.org/10.17226/13430.

Danabasoglu, G., J.-F. Lamarque, J. Bacmeister, D. A. Bailey, A. K. DuVivier, J. Edwards, L. K. Emmons, et al. 2020. “The Community Earth System Model Version 2 (CESM2).” *Journal of Advances in Modeling Earth Systems* 12 (2): e2019MS001916. https://doi.org/10.1029/2019MS001916.

Dethier, Corey. 2022. “When Is an Ensemble like a Sample? ‘Model-Based’ Inferences in Climate Modeling.” *Synthese* 200 (1): 52. https://doi.org/10.1007/s11229-022-03477-5.

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

Edwards, Paul N. 2011. “History of Climate Modeling.” *WIREs Climate Change* 2 (1): 128–39. https://doi.org/10.1002/wcc.95.

Eyring, Veronika, Mattia Righi, Axel Lauer, Martin Evaldsson, Sabrina Wenzel, Colin Jones, Alessandro Anav, et al. 2016. “ESMValTool (v1.0) – a Community Diagnostic and Performance Metrics Tool for Routine Evaluation of Earth System Models in CMIP.” *Geoscientific Model Development* 9 (5): 1747–1802. https://doi.org/10.5194/gmd-9-1747-2016.

Fraassen, B. C. van. 1980. *The Scientific Image.* Oxford: Clarendon.

Frigg, Roman, Erica Thompson, and Charlotte Werndl. 2015. “Philosophy of Climate Science Part II: Modelling Climate Change.” *Philosophy Compass* 10 (12): 965–77. https://doi.org/10.1111/phc3.12297.

Frisch, Mathias. 2015. “Predictivism and Old Evidence: A Critical Look at Climate Model Tuning.” *European Journal for Philosophy of Science* 5 (2): 171–90. https://doi.org/10.1007/s13194-015-0110-4.

Gates, W. Lawrence. 1992. “AN AMS CONTINUING SERIES: GLOBAL CHANGE--AMIP: The Atmospheric Model Intercomparison Project.” *Bulletin of the American Meteorological Society* 73 (12): 1962–70. https://doi.org/10.1175/1520-0477(1992)073<1962:ATAMIP>2.0.CO;2.

Gates, W. Lawrence, James S. Boyle, Curt Covey, Clyde G. Dease, Charles M. Doutriaux, Robert S. Drach, Michael Fiorino, et al. 1999. “An Overview of the Results of the Atmospheric Model Intercomparison Project (AMIP I).” *Bulletin of the American Meteorological Society* 80 (1): 29–56. https://doi.org/10.1175/1520-0477(1999)080<0029:AOOTRO>2.0.CO;2.

Gates, William Lawrence. 1979. *Report of the JOC Study Conference on Climate Models, Performance, Intercomparison, and Sensitivity Studies (Washington, DC, 3-7 April 1978)*. 22. International Council of Scientific Unions;[Geneva: obtained from the Word Meteorological Organization

Gettelman, A., C. Hannay, J. T. Bacmeister, R. B. Neale, A. G. Pendergrass, G. Danabasoglu, J.-F. Lamarque, et al. 2019. “High Climate Sensitivity in the Community Earth System Model Version 2 (CESM2).” *Geophysical Research Letters* 46 (14): 8329–37. https://doi.org/10.1029/2019GL083978.

Gettelman, A., and H. Morrison. 2015. “Advanced Two-Moment Bulk Microphysics for Global Models. Part I: Off-Line Tests and Comparison with Other Schemes.” *Journal of Climate* 28 (3): 1268–87. https://doi.org/10.1175/JCLI-D-14-00102.1.

Gettelman, A., H. Morrison, S. Santos, P. Bogenschutz, and P. M. Caldwell. 2015. “Advanced Two-Moment Bulk Microphysics for Global Models. Part II: Global Model Solutions and Aerosol–Cloud Interactions.” *Journal of Climate* 28 (3): 1288–1307. https://doi.org/10.1175/JCLI-D-14-00103.1.

Gettelman, Andrew, and Richard B. Rood. 2016. *Demystifying Climate Models: A Users Guide to Earth System Models*. Edited by Andrew Gettelman and Richard B. Rood. Earth Systems Data and Models. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-48959-8_1.

Gleckler, P. J., D. A. Randall, G. Boer, R. Colman, M. Dix, V. Galin, M. Helfand, et al. 1995. “Cloud-Radiative Effects on Implied Oceanic Energy Transports as Simulated by Atmospheric General Circulation Models.” *Geophysical Research Letters* 22 (7): 791–94. https://doi.org/10.1029/95GL00113.

Hansen, J., G. Russell, D. Rind, P. Stone, A. Lacis, S. Lebedeff, R. Ruedy, and L. Travis. 1983. “Efficient Three-Dimensional Global Models for Climate Studies: Models I and II.” *Monthly Weather Review* 111 (4): 609–62. https://doi.org/10.1175/1520-0493(1983)111<0609:ETDGMF>2.0.CO;2.

Hartmann, Dennis L., and David A. Short. 1980. “On the Use of Earth Radiation Budget Statistics for Studies of Clouds and Climate.” *Journal of the Atmospheric Sciences* 37 (6): 1233–50. https://doi.org/10.1175/1520-0469(1980)037<1233:OTUOER>2.0.CO;2.

Hausfather, Zeke, Kate Marvel, Gavin A. Schmidt, John W. Nielsen-Gammon, and Mark Zelinka. 2022. “Climate Simulations: Recognize the ‘Hot Model’ Problem.” *Nature* 605 (7908): 26–29. https://doi.org/10.1038/d41586-022-01192-2.

Hegerl, Gabriele C, Francis W Zwiers, Pascale Braconnot, Nathan P Gillett, Yong Luo, Jose A Marengo Orsini, Neville Nicholls, et al. n.d. “Understanding and Attributing Climate Change.” In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Avery, M. Tignor, and H.L. Miller, 84. Cambridge University Press (U.K.; New York).

Held, Isaac M. 2005. “The Gap between Simulation and Understanding in Climate Modeling.” *Bulletin of the American Meteorological Society* 86 (11): 1609–14. https://doi.org/10.1175/BAMS-86-11-1609.

“Higher-Order Turbulence Closure and Its Impact on Climate Simulations in the Community Atmosphere Model in: Journal of Climate Volume 26 Issue 23 (2013).” n.d. Accessed August 19, 2022. https://journals-ametsoc-org.proxyiub.uits.iu.edu/view/journals/clim/26/23/jcli-d-13-00075.1.xml.

Hoesly, Rachel M., Steven J. Smith, Leyang Feng, Zbigniew Klimont, Greet Janssens-Maenhout, Tyler Pitkanen, Jonathan J. Seibert, et al. 2018. “Historical (1750–2014) Anthropogenic Emissions of Reactive Gases and Aerosols from the Community Emissions Data System (CEDS).” *Geoscientific Model Development* 11 (1): 369–408. https://doi.org/10.5194/gmd-11-369-2018.

Hotz, Robert. 2022. “Climate Scientists Encounter Limits of Computer Models, Bedeviling Policy.” *The Wall Street Journal*, February 6, 2022.

Hourdin, Frédéric, Jean-Yves Grandpeix, Catherine Rio, Sandrine Bony, Arnaud Jam, Frédérique Cheruy, Nicolas Rochetin, et al. 2013. “LMDZ5B: The Atmospheric Component of the IPSL Climate Model with Revisited Parameterizations for Clouds and Convection.” *Climate Dynamics* 40 (9): 2193–2222. https://doi.org/10.1007/s00382-012-1343-y.

Hourdin, Frédéric, Thorsten Mauritsen, Andrew Gettelman, Jean-Christophe Golaz, Venkatramani Balaji, Qingyun Duan, Doris Folini, et al. 2017. “The Art and Science of Climate Model Tuning.” *Bulletin of the American Meteorological Society* 98 (3): 589–602. https://doi.org/10.1175/BAMS-D-15-00135.1.

Jebeile, Julie, and Anouk Barberousse. 2021. “Model Spread and Progress in Climate Modelling.” *European Journal for Philosophy of Science* 11 (3): 66. https://doi.org/10.1007/s13194-021-00387-0.

Jebeile, Julie, and Michel Crucifix. 2020. “Multi-Model Ensembles in Climate Science: Mathematical Structures and Expert Judgements.” *Studies in History and Philosophy of Science Part A* 83 (October): 44–52. https://doi.org/10.1016/j.shpsa.2020.03.001.

Jeevanjee, Nadir, Pedram Hassanzadeh, Spencer Hill, and Aditi Sheshadri. 2017. “A Perspective on Climate Model Hierarchies.” *Journal of Advances in Modeling Earth Systems* 9 (4): 1760–71. https://doi.org/10.1002/2017MS001038.

Kawamleh, Suzanne. 2022. “Confirming (Climate) Change: A Dynamical Account of Model Evaluation.” *Synthese* 200 (2): 122. https://doi.org/10.1007/s11229-022-03659-1.

Kravitz, Ben, and Douglas G. MacMartin. 2020. “Uncertainty and the Basis for Confidence in Solar Geoengineering Research.” *Nature Reviews Earth & Environment* 1 (1): 64–75. https://doi.org/10.1038/s43017-019-0004-7.

Kuo, Yi-Hung, J. David Neelin, Chih-Chieh Chen, Wei-Ting Chen, Leo J. Donner, Andrew Gettelman, Xianan Jiang, et al. 2020. “Convective Transition Statistics over Tropical Oceans for Climate Model Diagnostics: GCM Evaluation.” *Journal of the Atmospheric Sciences* 77 (1): 379–403. https://doi.org/10.1175/JAS-D-19-0132.1.

Lenhard, Johannes, and Eric Winsberg. 2010. "Holism, Entrenchment, and the Future of Climate Model Pluralism." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, Special Issue: Modelling and Simulation in the Atmospheric and Climate Sciences, 41 (3): 253–62. https://doi.org/10.1016/j.shpsb.2010.07.001.

Li, Dan. 2022. "If a Tree Grows No Ring and No One Is around: How Scientists Deal with Missing Tree Rings." Climatic Change 174 (1): 6. https://doi.org/10.1007/s10584-022-03424-w.Lloyd, Elisabeth A. 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. https://doi.org/10.1016/j.shpsa.2012.02.001.

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

———. 2015b. "Adaptationism and the Logic of Research Questions: How to Think Clearly About Evolutionary Causes." *Biological Theory* 10 (4): 343–62. https://doi.org/10.1007/s13752-015-0214-2.

Lloyd, Elisabeth A., Melissa Bukovsky, and Linda O. Mearns. 2021. "An Analysis of the Disagreement about Added Value by Regional Climate Models." *Synthese* 198 (12): 11645–72. https://doi.org/10.1007/s11229-020-02821-x.

Lloyd, Elisabeth A., Richard C. Lewontin, and Marcus W. Feldman. 2008. "The Generational Cycle of State Spaces and Adequate Genetical Representation." *Philosophy of Science* 75 (2): 140–56. https://doi.org/10.1086/590196.

Maloney, Eric D., Andrew Gettelman, Yi Ming, J. David Neelin, Daniel Barrie, Annarita Mariotti, C.-C. Chen, et al. 2019. "Process-Oriented Evaluation of Climate and Weather Forecasting Models." *Bulletin of the American Meteorological Society* 100 (9): 1665–86. https://doi.org/10.1175/BAMS-D-18-0042.1.

Mann, Michael E. 2018. "Reconciling Climate Model/Data Discrepancies: The Case of the 'Trees That Didn't Bark.'" In *Climate Modelling: Philosophical and Conceptual Issues*, edited by Elisabeth A. Lloyd and Eric Winsberg, 175–97. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65058-6_7.

Marsh, Patrick T., Harold E. Brooks, and David J. Karoly. 2007. "Assessment of the Severe Weather Environment in North America Simulated by a Global Climate Model." *Atmospheric Science Letters* 8 (4): 100–106. https://doi.org/10.1002/asl.159.

Mauritsen, Thorsten, Bjorn Stevens, Erich Roeckner, Traute Crueger, Monika Esch, Marco Giorgetta, Helmuth Haak, et al. 2012. "Tuning the Climate of a Global Model." *Journal of Advances in Modeling Earth Systems* 4 (3). https://doi.org/10.1029/2012MS000154.

Mayernik, Matthew S. 2021. "Credibility via Coupling: Institutions and Infrastructures in Climate Model Intercomparisons:" *Engaging Science, Technology, and Society* 7 (2): 10–32. https://doi.org/10.17351/ests2021.769.

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

Meehl, Gerald A., George J. Boer, Curt Covey, Mojib Latif, and Ronald J. Stouffer. 2000. "The Coupled Model Intercomparison Project (CMIP)." *Bulletin of the American Meteorological Society* 81 (2): 313–18.

Morrison, Monica Ainhorn. 2021. "The Models Are Alright: A Socio-Epistemic Theory of the Landscape of Climate Model Development." Ph.D., United States -- Indiana: Indiana University.

Accessed August 30, 2021. https://www.proquest.com/docview/2489342331/abstract/CDF0E73D2F944EEPQ/1.

National Academy of Sciences, Climate Research Board. 1979. *Carbon Dioxide and Climate: A Scientific Assessment (Jule Charney, Chair).* Washington, DC: National Academy of Sciences.

Neale, R. B., and B. J. Hoskins. 2000. "A Standard Test for AGCMs Including Their Physical Parametrizations: I: The Proposal." *Atmospheric Science Letters* 1 (2): 101–7. https://doi.org/10.1006/asle.2000.0019.

Neale, Richard B, Andrew Gettelman, Sungsu Park, Chih-Chieh Chen, Peter H Lauritzen, David L Williamson, Andrew J Conley, et al. n.d. "Description of the NCAR Community Atmosphere Model (CAM 5.0)," 289.

Notz, Dirk, F. Alexander Haumann, Helmuth Haak, Johann H. Jungclaus, and Jochem Marotzke. 2013. "Arctic Sea-Ice Evolution as Modeled by Max Planck Institute for Meteorology's Earth System Model." *Journal of Advances in Modeling Earth Systems* 5 (2): 173–94. https://doi.org/10.1002/jame.20016.

Odenbaugh, Jay. 2018. "Building Trust, Removing Doubt? Robustness Analysis and Climate Modeling." In *Climate Modelling: Philosophical and Conceptual Issues*, edited by Elisabeth A. Lloyd and Eric Winsberg, 297–321. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65058-6_11.

O'Loughlin, Ryan. 2021. "Robustness Reasoning in Climate Model Comparisons." *Studies in History and Philosophy of Science Part A* 85 (February): 34–43. https://doi.org/10.1016/j.shpsa.2020.12.005.

Oreopoulos, Lazaros, Eli Mlawer, Jennifer Delamere, Timothy Shippert, Jason Cole, Boris Fomin, Michael Iacono, et al. 2012. "The Continual Intercomparison of Radiation Codes: Results from Phase I." *Journal of Geophysical Research: Atmospheres* 117 (D6). https://doi.org/10.1029/2011JD016821.

Parker, Wendy S. 2011. "When Climate Models Agree: The Significance of Robust Model Predictions." *Philosophy of Science* 78 (4): 579–600. https://doi.org/10.1086/661566.

———. 2018a. "The Significance of Robust Climate Projections." In *Climate Modelling: Philosophical and Conceptual Issues*, edited by Elisabeth A. Lloyd and Eric Winsberg, 273–96. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65058-6_9.

———.2018b. "Climate Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/climate-science/.

Pitari, Giovanni, Valentina Aquila, Ben Kravitz, Alan Robock, Shingo Watanabe, Irene Cionni, Natalia De Luca, Glauco Di Genova, Eva Mancini, and Simone Tilmes. 2014. "Stratospheric Ozone Response to Sulfate Geoengineering: Results from the Geoengineering Model Intercomparison Project (GeoMIP)." *Journal of Geophysical Research: Atmospheres* 119 (5): 2629–53. https://doi.org/10.1002/2013JD020566.

Ramanathan, V., R. D. Cess, E. F. Harrison, P. Minnis, B. R. Barkstrom, E. Ahmad, and D. Hartmann. 1989. "Cloud-Radiative Forcing and Climate: Results from the Earth Radiation Budget Experiment." *Science* 243 (4887): 57–63. https://doi.org/10.1126/science.243.4887.57.

Randall, David A., Cecilia M. Bitz, Gokhan Danabasoglu, A. Scott Denning, Peter R. Gent, Andrew Gettelman, Stephen M. Griffies, et al. 2018. "100 Years of Earth System Model Development." *Meteorological Monographs* 59 (1): 12.1-12.66. https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0018.1.

Schmidt, Gavin A., and Steven Sherwood. 2015. “A Practical Philosophy of Complex Climate Modelling.” *European Journal for Philosophy of Science* 5 (2): 149–69. https://doi.org/10.1007/s13194-014-0102-9.

Schmidt, Gavin A., David Bader, Leo J. Donner, Gregory S. Elsaesser, Jean-Christophe Golaz, Cecile Hannay, Andrea Molod, Richard B. Neale, and Suranjana Saha. 2017. “Practice and Philosophy of Climate Model Tuning across Six US Modeling Centers.” *Geoscientific Model Development* 10 (9): 3207–23. https://doi.org/10.5194/gmd-10-3207-2017.

Schmidt, Kjeld. 2012. “The Trouble with ‘Tacit Knowledge.’” *Computer Supported Cooperative Work (CSCW)* 21 (2): 163–225. https://doi.org/10.1007/s10606-012-9160-8.

Schneider, S. H. 1979. “Verification of Parameterizations in Climate Modeling.” In *Report of the Study Conference on Climate Models: Performance, Intercomparison and Sensitivity Studies*, edited by W. Lawrence Gates, 728–51. World Meteorological Organization, Global Atmospheric Research Program, GARP Publications Series no. 22, 2 vols.

Schneider, Stephen H. 1972. “Cloudiness as a Global Climatic Feedback Mechanism: The Effects on the Radiation Balance and Surface Temperature of Variations in Cloudiness.” *Journal of the Atmospheric Sciences* 29 (8): 1413–22. https://doi.org/10.1175/1520-0469(1972)029<1413:CAAGCF>2.0.CO;2.

———. 1975. “On the Carbon Dioxide–Climate Confusion.” *Journal of Atmospheric Sciences* 32 (11): 2060–66. https://doi.org/10.1175/1520-0469(1975)032<2060:OTCDC>2.0.CO;2.

Schneider, Stephen H., and Robert E. Dickinson. 1974. “Climate Modeling.” *Reviews of Geophysics* 12 (3): 447–93. https://doi.org/10.1029/RG012i003p00447.

Sengupta, Sailes, and James S. Boyle. 1998. “Using Common Principal Components for Comparing GCM Simulations.” *Journal of Climate* 11 (5): 816–30. https://doi.org/10.1175/1520-0442(1998)011<0816:UCPCFC>2.0.CO;2.

Steele, Katie, and Charlotte Werndl. 2013. “Climate Models, Calibration, and Confirmation.” *The British Journal for the Philosophy of Science* 64 (3): 609–35. https://doi.org/10.1093/bjps/axs036.

Sun, Ying, Susan Solomon, Aiguo Dai, and Robert W. Portmann. 2006. “How Often Does It Rain?” *Journal of Climate* 19 (6): 916–34. https://doi.org/10.1175/JCLI3672.1.

Tebaldi, Claudia, and Reto Knutti. 2007. “The Use of the Multi-Model Ensemble in Probabilistic Climate Projections.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365 (1857): 2053–75. https://doi.org/10.1098/rsta.2007.2076.

Touzé-Peiffer, Ludovic, Anouk Barberousse, and Hervé Le Treut. 2020. “The Coupled Model Intercomparison Project: History, Uses, and Structural Effects on Climate Research.” *WIREs Climate Change* 11 (4): e648. https://doi.org/10.1002/wcc.648.

Voosen, Paul. 2021. “U.N. Climate Panel Confronts Implausibly Hot Forecasts of Future Warming.” 2021. http://www.science.org/content/article/un-climate-panel-confronts-implausibly-hot-forecasts-future-warming.

Washington, Warren. 2006. *Odyssey in Climate Modeling, Global Warming, and Advising Five Presidents*. Edited by Mary C. Washington. lulu.com.

Weart, Spencer. 2020. “The Discovery of Global Warming - A History.” The Discovery of Global Warming. 2020. https://history.aip.org/climate/pdf/Gcm.pdf.

Wilson, Joseph. 2021. “Two Exploratory Uses for General Circulation Models in Climate Science.” *Perspectives on Science* 29 (4): 493–509. https://doi.org/10.1162/posc_a_00380.

Wimsatt, William C. 2007. *Re-Engineering Philosophy for Limited Beings. Piecewise Approximations to Reality.* Cambridge, MA and London, England: Harvard University Press.

Winsberg, Eric. 2018a. “What Does Robustness Teach Us in Climate Science: A Re-Appraisal.” *Synthese*, November. https://doi.org/10.1007/s11229-018-01997-7.

———. 2018b. *Philosophy and Climate Science*. Cambridge University Press.

Zhang, G. J., and Norman A. McFarlane. 1995. “Sensitivity of Climate Simulations to the Parameterization of Cumulus Convection in the Canadian Climate Centre General Circulation Model.” *Atmosphere-Ocean* 33 (3): 407–46. https://doi.org/10.1080/07055900.1995.9649539.

Zhang, Guang J. 2002. “Convective Quasi-Equilibrium in Midlatitude Continental Environment and Its Effect on Convective Parameterization.” *Journal of Geophysical Research: Atmospheres* 107 (D14): ACL 12-1-ACL 12-16. https://doi.org/10.1029/2001JD001005.

TABLE 8. Changes of model physics from Model I to Model II.

Test run	Physics change	Major effect
I–6	Coriolis/metric terms at pole	Strengthened polar cell
I–10	Drag in top model layer	Reduced stratospheric winds; realistic tropopause at high latitudes
I–13, 14	9 layers in vertical	Improved definition of jet stream and tropopause; more longwave generation
I–24	1 k-distribution for each gas	Faster computation; higher accuracy
I–25	Realistic surface emissivities	No large effect
I–29	No subgrid-scale temperature variation for moist convection	Increased EKE; reduced upper level humidity and temperature; narrowed Hadley cell
I–34	Moist convection can start below condensation level	Stronger high-latitude winter temperature inversion at low levels
I–36	Large-scale rain every 5 h	Increased large-scale cloud cover
I–40	Local $T = -40°C$ for saturation over ice	Less cirrus clouds at low latitudes
I–42, 43	Cloud optical thickness modified	Reduced net heat into ground
I–44	Snow density decreased	Warmer ground in winter
I–45	Ground thermal conductivity changed	Reduced vertical temperature gradient in ground
I–46, 47, 49	Altered hydrology based on vegetation; intermediate run-off formulation	Early summer moisture increased and temperature decreased
I–50	Realistic vegetation masking depths	Reduced albedo in snow-covered areas
I–51	Ground albedo based on vegetation	Small albedo increase in subtropics
I–52	Modified ocean ice coverage	Local effects on T and evaporation
I–54	Modified ocean temperatures	No large effect

Figure 1. Changes of model physics from Model I to Model II (excerpted from Hansen (1983)).