Studies in History and Philosophy of Science 49 (2015) 58–68

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Model robustness as a confirmatory virtue: The case of climate science

Elisabeth A. Lloyd

History and Philosophy of Science Department, Goodbody Hall 130, Indiana University, Bloomington, IN, 47405, USA

ARTICLE INFO

Article history:
Received 23 April 2014
Received in revised form
30 September 2014
Available online

Keywords:
Robustness;
Confirmation;
Scientific inference;
Climate models;
Variety-of-evidence;
Causal attribution

ABSTRACT

I propose a distinct type of robustness, which I suggest can support a confirmatory role in scientific reasoning, contrary to the usual philosophical claims. In *model robustness*, repeated production of the empirically successful model prediction or retrodiction against a background of independently-supported and varying model constructions, within a group of models containing a shared causal factor, may suggest how confident we can be in the causal factor and predictions/retrodictions, especially once supported by a variety of evidence framework. I present climate models of greenhouse gas global warming of the 20th Century as an example, and emphasize climate scientists' discussions of robust models and causal aspects. The account is intended as applicable to a broad array of sciences that use complex modeling techniques.

1. Introduction

Philosophers Zach Pirtle et al. documented the fact that climate scientists tend to be attracted to robustness and to think it boosts confirmation of models.[2] In a recent qualitative survey of the contents of six leading climate journals since 1990, they found 118 articles in which the authors relied on a rough concept of agreement between climate model predictions/retrodictions to inspire confidence in their results (Pirtle, Meyer, & Hamilton, 2010, p. 3). I, too, defended robustness as an empirical strength of the huge general circulation models,[1] GCMs, in earlier discussions of the variety of evidence supporting those models (Lloyd, 2009, 2010, 2012).

But philosophers of science usually do *not* consider robustness to be an empirical or confirmatory virtue, that is, a virtue that indicates that a model or models are more likely to be used to represent accurate or true claims about the observable world (e.g., Calcott, 2011; Houkes & Vaesen, 2012; Orzack & Sober, 1993). In philosopher Jim Woodward's recent exploration of four different types of robustness, including what he calls 'inferential robustness,' it is confirmatory only in a very narrow (and admittedly scientifically extremely unrealistic) range of circumstances: inference to the robust claim involves the assumption that a "complete" set of models under consideration includes a "true" model, and the parallel in probabilistic terms (2006, pp. 219–224). In Woodward's lovely understatement, "its range of applications looks rather limited" (2006, p. 222).

Here, I pursue a view related to that of Richard Levins (1966), William Wimsatt (1981, 2007), Michael Weisberg (2006), Weisberg and Reisman (2008), and Jay Odenbaugh (2011, ms), (a group henceforth abbreviated as 'LWWO'), and I expand arguments first made in (Lloyd, 2009). I describe a distinct type of informal inference using robustness, which I call 'model robustness.' It is based not only on the agreement or convergence of the empirically correct *outcomes* or *predictions/retrodictions* of a group of models, but *also* on the independent *empirical* support for the

E-mail address: ealloyd@indiana.edu.

[1] A retrodiction is a model result that describes phenomena that have already occurred. The advantages of modeling past phenomena are many, especially in that such models can be compared to any empirical measurements, data, or observations of such phenomena, as well as to observable proxies for any processes or phenomena that are claimed to have occurred. I will refer henceforth to "predictions/retrodictions" to remind the reader that the models to which I refer in this paper all relate to phenomena that have already occurred.

[2] In the context of the Pirtle et al. study, they refer to both predictions and retrodictions.

http://dx.doi.org/10.1016/j.shpsa.2014.12.002
0039-3681/

variety of *assumptions and features* of a span of models that all share a common 'causal core.' The 'causal core' is a dependency among key variables or parameters of interest, common to all members of the model-type (a 'model-type' is where the models have in common a type of structure, sharing general characteristics, in which "certain parameters are left unspecified" (van Fraassen, 1980, 44)).

A model-type may be first equated, for our purposes, with the climate scientists' 'conceptual model,' in which key causal connections and processes are envisioned, but the details and/or parameters are not yet specified. Those causal ideas may be instantiated in an actual simulation model, (a GCM or simpler model), which specifies the previously-unspecified parameters, and which contains details we discuss in Section 3. The causal core of the model-type, i.e., the causal processes and explanations of interest, are endorsed directly and indirectly by both the successful predictions/retrodictions and the empirical support of assumptions of the models, and they are partly responsible for the predictions/retrodictions being correct. Thus, 'model robustness' involves all this direct and indirect empirical support for the *explanatory causal core* of the model-type, and by means of this causal core, the model prediction/retrodiction is also empirically supported. Note that this is very different from other philosophical meanings of 'robustness,' which are usually solely defined in terms of the convergent *predictions* of a group of possibly (or even preferably) unrelated models.

2. Introduction to robustness

The key insight comes from biology. Theoretical ecologist Richard Levins described robustness in a landmark book concerning theoretical biological methodology in 1968. There, he noted that when there are multiple, varying models of the same phenomenon in nature, the scientist often focuses on a common causal structure in the models, represented in Fig. 1 by the rough-edged bursting central node.

This causal core[3] reliably relates to a common outcome, **T**, regardless of the differing idealizations or assumptions, represented by the varying arrows, made in the various models. Ultimately, in the hands of philosophers Wimsatt, Weisberg, and Ken Reisman, Levins' insight is translated into a claim that a common structure in the models, the shared bursting node, represents a real world phenomenon or *cause* (Calcott, 2011, p. 284; Levins, 1966, p. 431; Weisberg, 2006, p. 737; Weisberg & Reisman, 2008, pp. 114–115; Wimsatt, 2007, p. 60).

Biologist Steve Orzack and philosopher Elliott Sober argue against Levins' view, saying that since his proposed robustness inference does not involve examining data, it is a distinct and non-empirical form of confirmation, one that they reject as ineffective for making predictive inferences (1993, pp. 541–544). Levins, on the other hand, argues that Orzack and Sober have mischaracterized robustness analysis, and insists that there are, contrary to their claims, central, *empirical* aspects of robustness. Specifically, Levins emphasizes the *empirical support present for the common core* in the models, as well as *for the various assumptions* appearing in the variety of models under investigation (Levins, 1993, p. 554; see Fig. 2).

In sum, Levins-style robustness analysis does indeed involve empirical evidence, but that observational and experimental evidence focuses on the model's *assumptions and core structure* (Fig. 2), not its prediction/retrodiction (see Fig. 3).

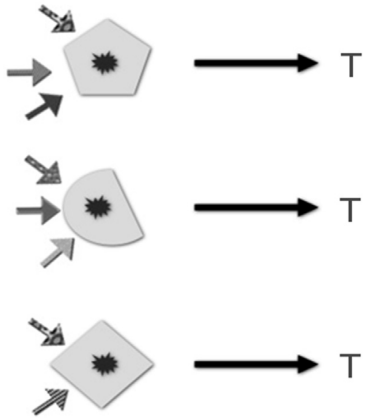

Fig. 1. Varying nodes with differing assumptions of various models, all predicting T.

Thus, we can see that Orzack and Sober had a different target, namely predictive inference to the model's outcome, as shown in Fig. 3, about which they were likely correct, although that is a topic for a different paper. Levins, in contrast, emphasized the key empirical evidence for the model structure under consideration from the other side of the arrow to the model outcome, as shown in Fig. 2. Thus, we can see that they were talking past one another on this point.

Still, that does little to establish a positive claim on Levins' behalf. Although the LWWO line of analysis has been extremely helpful by its frequent insistence on empirical support for assumptions of the model, (but see Odenbaugh & Alexandrova, 2011), they have not adequately described how or why the inference works to increase the confidence of the investigator in the causal core. Philosopher Brett Calcott, commenting on the Levins-Wimsatt approach, writes that although a series of models might be seen as robust, "by itself this is not enough to confirm anything. The models must be connected to the world, and this relies on making good on the resemblance they are meant to have with the phenomena in question" (Calcott, 2011, p. 287; Houkes & Vaesen, 2012).

In this paper I describe how 'model robustness,' in the context of climate science, provides—against the usual philosophical claims, e.g. Woodward, 2006—a *confirmatory virtue*, through discussing the case of greenhouse gas models of Twentieth Century warming.[4] When we start with a LWWO-type approach, ensure the independent empirical support of model assumptions, in addition to the predictive success of the models, and add a bit of reasoning about variety of evidence, we can help supply a philosophical confirmatory framework for the reasoning about robustness being done by the climate scientists. Philosophers Wybo Houkes and Krist Vaesen agree with my previous set up and conclusion that, contrary to the traditional philosophical view (e.g. Woodward, 2006, etc.), confirmation through robustness *may* occur (2012; Lloyd, 2009, 2010), but offer no explanation regarding *how*, as I do here. The result is intended to apply to many scientific cases, where the structure of complex model types and causal foci appears.

An interpretive note regarding my treatment of models and confirmation: When discussing models and modeling, I assume that the models (and climate simulations, which I treat as large models, although they may, under different circumstances, be treated as distinct (Edwards, 2010)) are indicated as similar to, and intended to represent particular aspects of the real world climate

[3] More correctly, such a structure is a 'causal focus,' as it can represent parameterizations, parameter values, etc. But I will call it a 'causal core' here, as that is a common use.

[4] Weisberg and Reisman, in contrast, in their very useful discussion of the Lotka Volterra models, are not arguing for a confirmatory virtue (2008, p. 108).

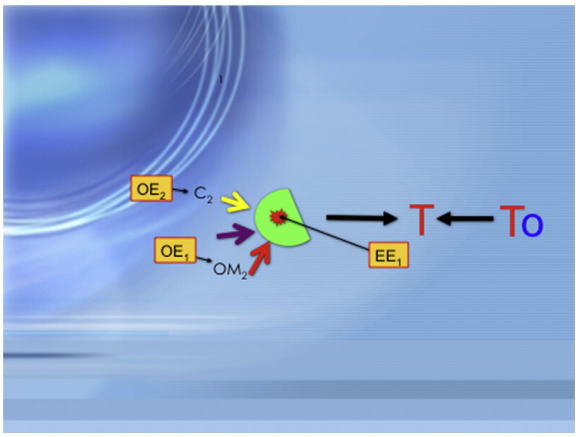

Fig. 2. Observational evidence (OEi) supporting assumptions, experimental evidence (EE) supporting core of model, temperature observations confirming predictions of T.

system, in specific respects and degrees, and also that this representation will be judged as adequate or not, relative to the purposes of the modelers or those who use the models, as specified by philosophers Wendy Parker (2006), Ronald Giere (2004) and Bas van Fraassen (2004, p. 794). I use shorthand such as "this climate model is supported or confirmed by such-and-so evidence," to represent these more complicated relationships.

2.1. 'Measurement robustness' and 'model robustness'

As Pirtle et al. (2010) documented, climate scientists may think, when confronted with results of a set of climate models that agree on a prediction/retrodiction (or range of such) of global mean surface temperature, **T**, that we should have more confidence in that prediction/retrodiction. So sometimes scientists do appear to think that robustness is a confirmatory virtue. In fact, they appeal to just such reasoning in the Nobel Prize-winning Report by the International Panel on Climate Change (IPCC)[5] in 2007: "models are unanimous in their prediction of substantial climate warming under greenhouse gas increases, and this warming is of a magnitude consistent with independent estimates derived from other sources, such as from observed climate changes and past climate reconstructions" (Randall et al. IPCC, 2007, p. 601). Note that the climate scientists include an appeal to the causal core of the climate models under consideration, i.e., the "greenhouse gas increases," in addition to the unanimity of the predicted causal result, this latter feature being the sole focus of other philosophers' analyses.

Climate scientist David Randall et al. are here offering independently-derived and independently-observed empirical evidence to reinforce the robust convergent predictions/retrodictions based on a particular family of causal (Greenhouse Gas, '**GHG**') models under conditions of measured increases in greenhouse gases. At the very least, this is similar to, but not identical with Woodward's 'measurement robustness' (see Section 7). 'Measurement robustness' refers to using multiple channels to infer and converge on the correct value (or range of values) of a variable, or the reduction of error by repetition in independent contexts.

Avogadro's number is the most familiar example of measurement robustness, sometimes also called 'heuristic robustness.' Jean

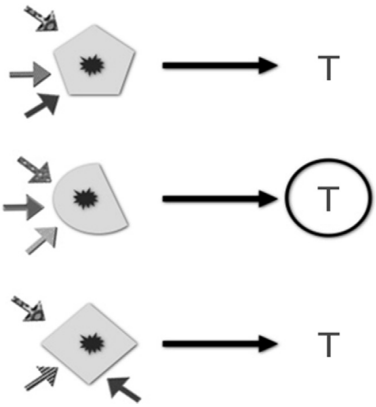

Fig. 3. A focus on model predictions/retrodictions.

Perrin used thirteen distinct and independent methods, measurements, reckonings and experiments to establish Avogadro's number in the early 20th Century. This was a robust convergence of outcomes that was taken to be too implausible a coincidence not to reflect something about the real world. The key to the persuasiveness of the case centered on the claim that the measurements were truly from distinct and independent methods and instruments (Cartwright, 1991; Salmon, 1984). Sylvia Culp offers some excellent examples from molecular biology (Culp, 1995).

In an extension of this type of reasoning, we can try to give a 'climate-model version' of measurement robustness, and to pose a set of various climate models as distinct 'experiments' or 'methods' for arriving at relevantly distinct outcomes, as exemplified by Hagedorn, Doblas-Reyes, and Palmer (2005, p. 227): "The basic idea of [collecting multiple climate models to compare outcomes] is to account for [an] inherent model error in using a number of independent and skilful models in the hope of a better coverage of the whole possible climate phase space." Here, the climate models under consideration need to be relevantly *independent*, as argued by a number of climate scientists, as well as by philosophers Parker (2011), Pirtle et al. (2010), Joel Katzav (2011), and Nancy Cartwright (1991). But it seems clear that many climate models are variants of others, and not usefully independent for these purposes, thus placing tight constraints on the climate-model-version of measurement robustness (Annan & Hargreaves, 2011; Edwards, 2010; Houkes & Vaesen, 2012; Knutti, 2008; Knutti, Abramowitz, et al., 2010; Knutti, Furrer, Tebaldi, Cermak, & Meehl, 2010; Muller & Von Storch, 2004, p. 33; Pirtle et al., 2010). Note that the sense in which different measurement methods for robustness in the case of Avogadro's number are *distinct* (e.g., they are causally independent; they have different possible sources of error, etc.), is quite different from the sense in which different climate models may be *independent*.[6]

More importantly, it is crucial to note that measurement robustness is usually used in aid of searching for a strong basis from which *to predict future temperatures* and other fu climate variables. I, on the other hand, am pursuing a distinct kind of robustness, 'model robustness,' used by climate scientists to *support causal explanations* of what has happened in the past history of the system. Our chief question is thus not about the robust effect or outcome *alone*, but *also* about *causes*: we want to know how reliable our understanding of the main causes of the robust effect is as it leads to the robust prediction. I emphasized the difference

[5] The IPCC is an international group of many hundreds of scientists, some hundreds of which are tasked with summarizing, by unanimous agreement among the authors of each chapter, the current state of climate science every five or six years.

[6] Thanks to an anonymous reviewer.

between using a model for explanatory causal purposes, versus merely predictive purposes, in (Lloyd, 2009) in an example using measles. Exposure to the measles virus *successfully explains the cause* in every case of a person's coming down with the disease, but it is a very poor *predictor* of coming down with disease; only some small percentage of those exposed to the virus, actually come down with measles. Hence model evaluation for different purposes can cast a very different light on the situation with any given group of models. And because of the terrible significance of climate predictions and projections,[7] from models or other calculations, so much emphasis has been put on projections for the future, that many have been distracted from evaluating the models for causal explanations used for predicting/retrodicting past and present climate change, the focus of our discussion here.

3. Global climate models

Global climate models represent mathematically the physical movements of gaseous and liquid masses, and the reflection, transfer and absorption of energy. In the largest, most complex GCMs—of which there are about 30–45 available today—the basic equations for the atmosphere at the heart of a climate model involve classical mechanics, thermodynamics, and fluid dynamics: a series of equations derived from these theories plus a "moisture" equation represent the atmospheric system in terms of pressure, temperature, and density. Similar sets of equations are used to represent the ocean in terms of pressure, temperature, density and salinity, and so on for the sea ice and land-surface system contributions to the climate system. A climate simulation model is constructed from physical and mathematical approximations of these "basic equations" that are solved by computer (Washington & Parkinson, 2005).

The greenhouse gas models we are discussing have central energy absorption and transfer or emission equations, which we can call "radiative causal cores," which represent these energetic processes (Pers. Comm., climate scientist Jeffrey Kiehl). The notion of a greenhouse gas 'causal core' of a model is a 'big tent' idea; there are many ways of modeling these basic energetic, radiative causal processes, with different modeling features and equations (see http://www-pcmdi.llnl.gov/ for schemes of updated models; Kiehl, 2007; Washington & Parkinson, 2005, pp. 79–91).

The state of the climate cannot, however, be fully derived from the physical theories, since we do not know the full set of physical processes guiding the system between states. Moreover, only the largest scale processes are explicitly represented in the model; smaller-scale processes cannot be modeled directly, and are handled by 'parameterizations.' These parameterizations are basically mini-models that coordinate small-scale parameters[8] with the large-scale variables of the climate model; they attempt to take into account the important impacts of 'subgrid processes' without simulating them explicitly.[9] Important examples of parameterized processes include cloud formation, convection, and ocean mixing, each of which has significant effects on climate (Edwards, 2010).[10] Thus, details and complexity about the ice, vegetation, soil and water vapor and the ornate interconnectedness of systems are represented in these GCMs. E.g., modelers can combine general pieces of theory from fluid dynamics, thermodynamics, and theories regarding radiation with, perhaps, precise ideas and measurements of how water vapor interacts with temperature in the context of a cloud from a parameterization. Ordinarily, the model instantiates at least several distinct generalized laws from different branches of physics, and the unique combination of these laws has usually never before been assessed. Thus, in climate science, "A model essentially embodies a theory," as David Randall and Bruce Wielicki say (1997, p. 400). Climate simulation modeling in this case can be an instance of theory articulation and application. We start with a model-type as the 'conceptual model,' and proceed to specify the parameterizations and details, in order to get a fully operational climate simulation model, from which we can determine whether the projected values conform to our expectations.

4. Robustness: an example from climate science

Let us examine multiple models of the same 20th Century climate system, keeping an eye on the types of evidence needed to help confirm a cause of the system's evolution and changes. A typical approach is to present six to a dozen or more of the huge climate models—including atmospheric, oceanic, and sometimes ice and/or land components—and compare the model results on specific experiments or parameter values; the models often produce the same range of outcomes for the specified values, and some are also supported empirically (Braconnet et al., 2007, p. 226; Gates et al., 1999; Gleckler, Taylor, & Doutriaux, 2008; Murphy et al., 2004). These collections of models are convenience-based, since the models are not generated in an orderly way or designed to explore specific parts of the possible model-space (Knutti, Abramowitz, et al., 2010). Such comparisons and compilations figure prominently in the 2007 IPCC Report (Meehl et al., 2007; Randall et al., 2007, pp. 594ff).

A clear example of convergence on a result is one in which all of the available climate models that incorporate greenhouse gases as a cause of climate change produce an increase in global mean surface temperature (GMST) in the late 20th Century. Fig. 4 shows 14 of them used in 58 simulations of 20th Century surface global temperature trends (Randall et al. IPCC, 2007, p. 600; see Knutti, Furrer et al., 2010, p. 7; cf. Parker, 2011).

The GMST change in these simulations is caused by both human causes like greenhouse gases and particulate pollution (aerosols),

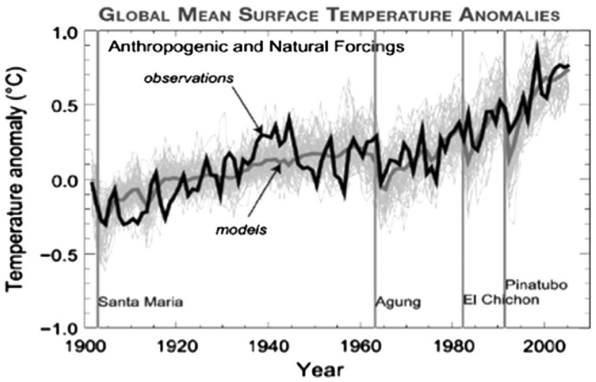

Fig. 4. 14 **GHG** GCMs, in 58 simulations of 20th Century GMST anomaly trends.

[7] A 'projection', in contrast to a 'prediction', incorporates or reflects information such as details of greenhouse gas emissions and other socio-economic variables in its construction, as represented, for instance, in the distinct "low-emissions" vs. "high emissions scenarios" in the 2007 IPCC report. This information is typically provided by building "integrated assessment models" or other socio-economic calculations, which are distinct from the climate models we are discussing here.

[8] Parameters are fixed values; variables define the system and can take different values over time.

[9] 'Subgrid processes' are those in which the physical processes involve variables and parameters on a scale smaller than the three-dimensional model grid cell, which typically may have a size of 100–300 km per side and be 1–10 km thick.

[10] See Lloyd (2009, 2010) for an elaboration of the evaluation and confirmation of complex models that focuses on independent support of all aspects of the model.

and natural causes such as solar events or volcano eruptions, which are marked in the graphs with their names. The volcanoes cool the atmosphere because of aerosol particles from the eruption shielding the earth from the sun. Each run of a model or simulation is represented by a thin pale line in Fig. 4, while the mean of the 58 runs is represented by the thick, smooth line. This mean model surface temperature anomaly is compared to the mean calculated from observational data, represented here by the jagged thick black line. The results of the modeling exercise include a good match between the model mean and the observational mean anomaly, representing a clear increase in temperature over the century of about 0.8 °C.

Let us consider, briefly, the variation in the model outcomes appearing in Fig. 4. The models vary in their assumptions and parameterizations—such as their representations of cloud cover or density—and also in some of their magnitudes and types of forcings (causes of change in the system), such as ozone changes or changes in land use.[11] Despite this model variation, all of the models in this model family share a core representation of greenhouse gases (**GHG**) as a radiative cause. We can consider this the common causal core shared by this entire **GHG** model-type under consideration, named **M**.[12] Our ultimate focus is on this shared causal core and the predictions produced from it.

Before proceeding, let us take a peek at what a model family would look like that does not include **GHG** radiative causation. Consider such a model family, which uses only natural causes, i.e., the volcano aerosols, called **N**.

We can see from Fig. 5, that the **N** model family results also converge, in that they produce temperature anomaly outcomes that are routinely lower than **M**-family results, and are also routinely empirically incorrect, since they do not track the measured mean temperature anomalies from the 20th Century.

In both of these model-types, **N** and **M**, let us look more closely at just what the "observed pattern of change," actually consists in (Hegerl et al., 2007, p. 702). First, I have oversimplified. The significant predictions/retrodictions of the models actually include complex spatio-temporal patterns of change in temperature, and vertical profiles (from Earth's surface to above the stratosphere), among other details, all of which I will simply call **T**. The spatio-temporal, structured temperature evidence will be **To** (for 'temperature observations').

These spatio-temporal patterns and vertical profiles of temperature are extremely important prediction/retrodictions from the greenhouse model type, **M**, since they differ sharply and significantly from predictions/retrodictions arising if there had been climate warming from solar or other natural forcings, internal variability, or stochasticity, and are so distinctive that they are sometimes referred to as a "fingerprint." Once these "fingerprints" in temperature are found across both a range of models and actual measurements and observations, this is considered confidence-inducing empirical evidence for increases in greenhouse gases having *caused* much of the late 20th Century warming (Hegerl et al., 2007; Hegerl & Zwiers, 2011, p. 13; Parker, 2010, p. 1092).

Laboratory experiments done by John Tyndall in the mid-19th Century established the causal connection between CO2 (and other greenhouse gases) and an increase in atmospheric temperature (Hulme, 2009). But questions remained about whether these laboratory setups resembled the earth's real atmosphere enough to provide a *causal explanation* at the global scale. Today, support for

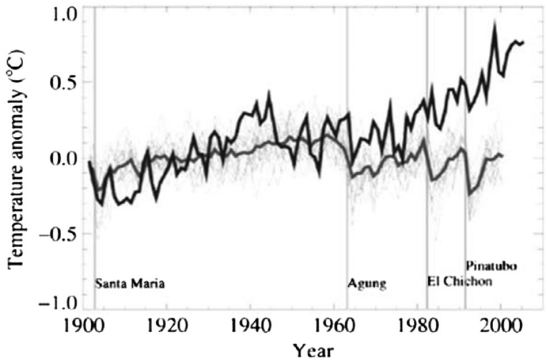

Fig. 5. 5 Non-**GHG** models, natural forcings only, in 19 simulations of 20th Century GMST anomaly trends.

the **GHG** explanation for recent raised temperatures comes from a wide variety of sources; note the repeated mentions of 'robustness' in climate scientists Gabriel Hegerl and Francis Zwiers' summary: The attribution results for large-scale temperature changes are supported by a large number of different lines of evidence and are robust to using very simple to fully complex models. They are also robust to using different methods and to using different assumptions and even approaches that avoid direct use of models altogether. They are also physically consistent with detection and attribution results from other climate variables, including, for example, tropopause height, vertical temperature of the atmosphere, atmospheric humidity, and to some extent, precipitation changes (Hegerl & Zwiers, 2011, p. 18).

Note that there is a broad sense of 'robust' being used here, and only part of this quote refers to what I call 'model robustness,' specifically, the phrase "attribution results" refers to the outcomes of analyses of the *causes* of recent (20th and 21st C.) *global warming phenomena*, including temperature changes. Hegerl and Zwiers claim that such causal explanations are robust, resistant, or impervious to all of these "different assumptions," factors, and contexts, etc., i.e., that the **GHG** model-type, configured in a wide variety of ways, succeeds in tests using a full variety of well-confirmed model assumptions, as desired for model robustness (Note that filling these conditions is far more complex and richly evidential than simple variety of evidence, for which it might perhaps be mistaken. We will address this difference in Section 6). In other words climate scientists believe that greenhouse gas causation works, and represents a real climate process, partly because using that causal core in their models produces simulations with convergent, successful predictions/retrodictions time and time again, supported by many distinct features of evidence, against a range of independently confirmable and distinct background causal assumptions. That is, this type of reasoning involves inferring causal processes through 'model robustness.' Quite simply, they believe that the physics are sound and well-confirmed.

5. Robustness as a confirmatory virtue

We can often see other climate scientists echoing the above picture of 'model robustness' reasoning as an indication that the models correctly represent a real or adequate causal account in the climate system. Climate scientist Jeffrey Kiehl, in discussing the fact that *all* **GHG** models simulate the global warming during the 20th Century with reasonable accuracy, notes, "[t]his is viewed as a reassuring confirmation that models to first order capture the

[11] Some models were run multiple times, initiated with different initial conditions.

[12] Model families are not necessarily hierarchical, they may be based also on sharing of simple parameters or parameterizations (Lloyd, 1994).

behavior of the physical climate system and lends credence to applying the models to projecting future climates" (Kiehl, 2007, p. 1).[13] Climate scientists Steve Lambert and George Boer also use the language of "capture" when discussing robust model results. A small amount of intermodel variance, they say, indicates agreement among models and "supports the assumption that they are *capturing the processes* that govern that variable and hence its climate" (2001, p. 88; emphasis added). And the reverse indicates the opposite about causal processes: "A large value of intermodal scatter, on the other hand, indicates disagreement and unreliability" (2001, p. 88). Hence, it is clear that the convergence of model projections is used as an indicator of the reliability of the causal processes in the models.

On sociologist of science Simon Shackley's analysis, there is a type of climate scientist who is pursuing a complete and realistic simulation model of climate, including the "capture" or representation in the models of basic predictive causal forces involved in changing climate, such as greenhouse gases, or volcanoes. I would add that they use, in addition to convergence of outcomes, the confirmation and variety of assumptions of various models, to support inferences involving such capture of causal forces (Hegerl et al., 2007; Lloyd, 2010; Shackley, 2001).

This type of inference was first characterized, to my knowledge, by ecologists. Richard Levins describes using assemblages of ecological models that share a common or "constant" core of assumptions, but also differ as widely as possible in other "variable" assumptions, and writes, "then the more the variable part spans the range of plausible assumptions, the more valid the claim that the conclusions shared by all of them depend on the constant part" (1993, pp. 554–555). He continues: "If we also have confidence that the constant part is true, then we have strong support for the claim that the conclusion is generally true" (1993, p. 555; see Weisberg & Reisman, 2008, pp. 119–120). This is an exact parallel description to model robustness in the greenhouse gas case.

In sum, the existence of convergence of key outcomes as well as empirical success of varied assumptions of a model is often seen as evidence in favor of a sort of rough or everyday realist interpretation of the model-type's causal core structures, or the reasonably accurate representation of specific causal processes described in the models (Calcott, 2011; Hegerl et al., 2007; Wimsatt, 2007, p. 60). As Weisberg writes, with a very realist bent, with robust outcomes and a common causal core among varied models, it is "very likely" that the real world has a "corresponding causal structure" (2006, p. 739). Thus, I urge extending the usual analytical philosophical focus on "robustness" from the product or *outcome* of the models to the *models themselves* (Lloyd, 2010). And although this line of emphasis has been alive ever since Levins and Wimsatt first wrote, and continues into recent discussions such as those by Weisberg and Odenbough, many are fuzzy about describing how or why such emphasis is desirable (e.g., Houkes & Vaesen, 2012). Explicating these lines of reasoning is useful not just for climate science, but for the many types of scientific inquiry that now involve complex and computational models, including aspects of ecology, evolutionary biology, engineering design, physics, economics and social sciences (Cartwright, 1991; Levins, 1993, 2006; Lloyd, 2009, 2010; Odenbaugh, 2011; Odenbaugh & Alexandrova, 2011; Rykiel, 1996; Winsberg & Lenhard, 2010;).

6. Model robustness

Thus, let us describe *model robustness* more specifically, using our example.

We start with our model type, **M**, which has greenhouse gases as a component in its causal core, **GHG**, robustly giving us increases in global mean temperature, **T** (as a spatio-temporally structured variable). Members of the family of model type, **M**, namely M1, M2, … Mn, are each different, but all are built with some **GHG** causal core. Each of these models also has an array of different parameterizations, parameter values, and assumptions, represented as, **Ai**, *indexed assumptions*, such as parameterizations of clouds or ocean mixing, and different parameter values or causes.

For simplicity, let us take a set of three distinct members of model type **M**: M1, M2, M3, all of which converge on robust prediction/retrodiction **T**, and say that M1 contains certain assumptions and aspects, **Ai**, which includes cloud parameterization C1 and ocean mixing parameterization OM1, among other aspects, and likewise M2 contains cloud parameterization C2 and ocean mixing parameterization OM2, among other aspects, etc. (The **C** and **OM** families of parameterizations are all types of minimodels, that also usually incorporate some empirical data into the minimodel, which are in turn embedded into the **M**s (Edwards, 2010).)

Because of the variety of parameters, variables, and parameterizations used in the construction of M1…Mn, there is also a wide variety of empirical evidence that can be brought to bear on the assumptions, **Ai**s, of these individual models, in addition to scoring its empirical success in producing accurate global mean surface temperatures, **T**. For instance, one model may rely on empirical evidence supporting its parameter values in its modeling of the El Nino Southern Oscillation (ENSO), while another may rely heavily on the empirical support for a number of details, such as moisture content, drop size, etc., of its cloud parameterization. Yet another model's empirical success may rely firmly on its modeling of the thermodynamics in the causal core, along with ocean heating dynamics, and so on. Because the details of empirical support for these assumptions of the individual models—the values relating to parameterizations, variables, parameter values, and model structures—differ in the case of each individual model or model application, it is necessary to construct individualized sets of confirming empirical evidence for each model application in the set of robust models of the model family. Thus, the different **Ai**s are each supported by their own bodies of empirical evidence, even while they produce competing or conflicting detailed climate systems (Parker, 2006; Edwards, 2010; Rykiel, 1996; Odenbaugh, 2011). This collection of observational (OE) and experimental evidence (EE) for the model assumptions, parameterizations, etc., is called **OEi**, where i is the given model.

This last assumption about the availability of evidence for conflicting aspects of the different models may come as a surprise to many philosophers, but it is extremely plausible in the science (see Parker, 2006 for an excellent discussion of the fact that climate scientists pursue multiple, incompatible models simultaneously; compare Houkes & Vaesen, 2012, p. 351). The primary reason for this is that observations are incomplete, and there may not be enough observations to favor one of the versions of C1–C3 above the others, although there are generally data supporting each version. Thus, *different observations support the distinct, and often incompatible, parameterizations*. In addition, there are many distinct ways to measure and characterize most climate variables and parameters, leading to a multiplicity of different and conflicting descriptions of the same system. Finally, different theoreticians and modelers make different choices about the simplification and descretation (for computation) of the basic equations of the system they are modeling. There is no 'true' or 'right' way to simplify the system, but choices need to be made in order to build a model and represent the system (Giere, 2006; van Fraassen, 2008). This is simply how modeling must be done, given our human finitude.

[13] Kiehl goes on to critique this sort of inference to the models' goodness.

Now let us consider our predictions/retrodictions using models, M1-M3, which all include **GHG** forcings. We can run the simulations using identical inputs of greenhouse gas levels, and let us say that we obtain GMST, vertical temperature profiles, and spatio-temporal patterns, **T**, for the present day, or deep or recent past. And now we test **T** against the observations, **To**, and we discover that **T** is, to within our needs and purposes, accurate, as shown in Fig. 7. In such real-life inferences, we must be careful to propagate the uncertainties. For example: how much variation in the prediction/retrodiction is dependent on fitting the cloud model to the cloud data and its uncertainties? And similarly for any other parameterizations and their uncertainties. Still, we find we have a small set of models, M1-M3, that predict **T** in an empirically adequate way, and use the same **GHG** radiative core, but very different processes and **Ai**s to do so. That is, we have sampled a variety of processes in coordination with that common **GHG** radiative causal core, and no matter which ones we use with it, we still get **T** as an outcome.[14] Thus, **M** turns out to be a pretty strong model-type, and the **Ai**s and core are well-supported by other observational and experimental evidence, **OEi** (see Raisenen, 2007; Randall et al., 2007; Hegerl & Zwiers, 2011; Lloyd, 2009, 2010, 2012). Levins noted that the variety of assumptions may be considered as representing a "space of possible models" (1968, p. 7; thanks to an anonymous reviewer for the reference)." (It may appear, upon examining the 'model robustness' of the **GHG** models, that the various **Ai**s do not make any difference, because no matter which **Ai** is used, the result, **T**, is always coordinated with the causal core. But in actuality, if, say, the troublesome cloud parameterizations would be removed from the models, the models would 'blow up' (in modeling terms), producing no atmosphere or climate system at all. Thus, the variety of models in the model-type require the seemingly extra parameterizations and assumptions in order to run; they are neither optional nor functionless.)

In sum, we have so far (looking at Figs. 6 and 7):

(a) Given a model-type **M**, which is characterized by the inclusion of the **GHG** causal core, there is variety of different assumptions and parameterizations **Ai**s (including, e.g., **Ci** and **OMi**) composing the rest of the model, such that (**M** & **Ai**s) implies conclusion **T**;
(b) There is independent experimental and observational evidence for the **GHG** radiative causal core relationship with **T** embedded in **M**;
(c) There is independent empirical evidence, **To**, for **T**;
(d) There is some evidence for each **Ai**, but we do not know which **Ai** is the best;
(e) In sum, **T** is a robust result under the combination of the variety of assumptions and parameterizations, **Ai**s, which are themselves usually empirically supported, combined with any individual **Mi**, which includes the **GHG** causal radiative core, (**Mi** & **Ai**s).

With model robustness, we can thus identify the patterns of evidence that support a model-type and its causal core, while tracking the processes of reasoning used in climate modeling and the confirmation of climate models. Model robustness involves both the causal core of the individual models and the convergence of the model outcomes. The causal core is robust in that even if we change the parameterizations or details of the other aspects of the models, they are well supported empirically to specified degrees by

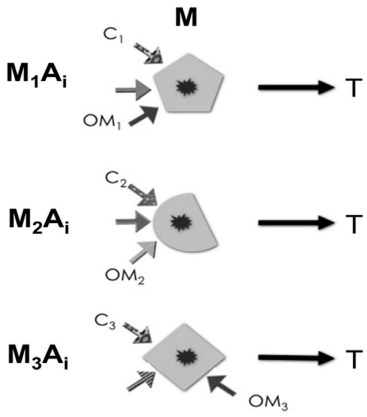

Fig. 6. Three models of model-type **M** with some **Ai**s specified: parameterizations of Clouds, **Ci**, and Ocean Mixing, **OMi**.

multiple aspects of the observational evidence. More generally, it is against the background of differing model constructions, M1, M2, … Mn, within the model family, **M**, that the radiative causal core, **GHG**, occurs and causes the robust property, **T**, to appear, and it is the degree of this variety-of-instances (the differing model constructions) for which the model has been verified, that affects how confident we should be in the causal connection, and its performance in accurate, and robust, prediction/retrodiction of **T**.

Ultimately, the notion of 'model robustness' is used in climate science in a nesting fashion. Robust model-types such as the **GHG** models discussed above utilize, in turn, robust mini-model-types as their internal parameterizations, e.g., cloud and convection parameterizations, and within each of these, there may be more robust model-types.[15] The fact that the **GHG** model-type utilizes, in its structure, such a set of robust assumptions (mini-model-types), makes it the strongest and most robust model-type currently available in climate science, and especially suitable to illustrate the concept of 'model robustness.'

Here we must pause to sort out a vexed issue, specifically, that the concepts of 'variety of evidence' and 'robustness' are very frequently confounded in the philosophical literature. A variety of evidence usually involves the support of multiple avenues of evidence for a single hypothesis or empirical claim. For example, the evolutionary hypothesis that birds descended from dinosaurs is supported not only by the fossil and embryological evidence, but also by repeated DNA data and analyses; thus, the bird descent hypothesis is supported by a variety of evidence, that is, multiple and distinct lines of evidence (Coyne, 1999). In our case, a single model is supported by a substantial variety of evidence. For example (see Fig. 2), the individual model, M2, is confirmed not just by its success in predicting **T**, by **To**, but also, crucially, by independent observational support for its parameterizations, among them, C2 and OM2, as well as experimental support for the **GHG** causal process itself, through Tyndall's and later laboratory experiments (Hulme, 2009). There is *a variety of observational and experimental evidence* supporting model M2, in Fig. 2.[16] Such an array of evidence is widely understood to offer increased support for such a model, all things being equal.

[14] See Weisberg and Reisman (2008), on varying the parameter values with "parameter robustness analysis," and varying the laws with "structural robustness analysis."

[15] Thanks due to climate scientist and statistician Doug Nychka.

[16] This is shorthand for the claim that this bundle of evidence supports: "Model M2 represents specific aspects of the real world, say, various structures contributing to and predicting/retrodicting global mean temperature, to specified degrees, for purposes x, y, or z."

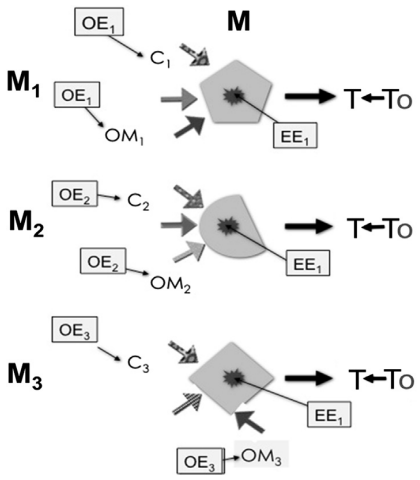

Fig. 7. **GHG** models, **M**, with a variety of evidence.

So, what exactly do we add to confirmation by including *more* models? In other words, if we already have a variety of evidence for a particular model, what good does *robustness* do?

Suppose we really want to know about the reliability of the model-type itself, especially the value of the causal core of a model-type of which the individual model is a member? Having a variety of evidence for a good model is a strong start. One issue, though, is that we have a whole collection of such good models. But in addition, one standard answer to the question about the value of a model-type's causal core itself, is that we can test the causal core in a variety of different contexts, with different parameterizations and causal assumptions, **Ai**s, which builds confidence in the core's efficacy, accuracy and reality, as noted by Hegerl and Zwiers (2011; Weisberg, 2006; Wimsatt, 2007; Hegerl et al., 2007; Knutti, Furrer et al., 2010). This treatment simultaneously gives us reasons to think that when models predict/retrodict and converge on **T**, the causal core of the model is also confirmed indirectly. It is tricky, because part of this sort of 'model robustness' inference may, in specific cases, fit more naturally as a subtype of variety-of-evidence inferences, taking each Mi as a *randomly chosen member* of the specific model-type under consideration, **M**. Note that we are frequently not considering an entire model-type here, but only a well-defined subset of it.

We can imagine that each model is an "experiment" for purposes of a variety of evidence argument. These "experiments" are in the form of random, distinct, independently confirmed, models, M1, M2, … Mn, and their supporting observational and experimental evidence, in which **GHG** is part of the radiative causal core, and *other assumptions*, **Ai**s, such as formulations of equations, values of forcings, or parameterizations, of the individual models *vary*. Significantly, each random model is well-supported by a variety of empirical and experimental evidence, making it a satisfactory candidate to serve as evidence or an "experiment." This situation includes, as we have discussed, that many of the **Ai**s are often independently empirically supported, as well as the causal core itself having independent experimental and/or observational evidence of its own.

It is also important that these models have to be distinct, or "heterogeneous enough," as Weisberg and Levins emphasized. But they usually are, as each **Mi** uses both distinct causes in conjunction with **GHG**, as well as distinct parameterizations. Even small differences in parameterizations are enough to cause relatively large differences in model outputs, due to feedbacks and nonlinear effects (Houghton et al., 2001). For example, as the climate warms, the (parameters representing the) amount of snow and ice cover decrease, which leads to more sunlight being absorbed because of the loss of reflectivity of the surface, thus establishing a positive feedback of warming in the climate system; a small variable change leads to a small parameter change which has a relatively large effect. Characteristically, as well, parameterizations of clouds make a variety of very different assumptions about the distribution of different types of clouds in a grid-box, thus representing cloud phenomena very differently in the models, for a variety of purposes, all of which can have radically different effects on the outcomes of the GCM as a whole (IPCC 2007; http://www-pcmdi.llnl.gov/projects/cloud_feedbacks/index.php).

There may also be a concern that the variety of models in the robustness analysis above is constrained to a particular model-type, i.e., with a **GHG** causal core, and then constrained further by requiring empirical support for values of assumptions in the models. But the climate scientists do not need to supply an especially rich variety to the causal core hypothesis. If they supply any variety *at all*, it is *confirmatory*. This grants more confirmatory power to robustness than non-LWWO philosophers of science have granted in detailed discussions before, e.g., with their scientifically unrealistic requirement of the guarantee of truth and completeness among the models (Orzack & Sober, 1993; Woodward, 2006).[17]

In sum, this is a way in which the **GHG** causal core itself can have its confidence and reliability raised through its repeated successes in producing accurate predictions/retrodictions of late 20th and early 21st C. global mean temperature, **T**, in conjunction with a variety of independently empirically supported model assumptions. Model robustness describes a pattern of models and evidence, which is described within a variety-of-evidence inference, as telling us *more* than any given piece or subset of pieces of evidence as used in these inferences, and as giving us increased confidence first in the causal core, and ultimately in the model outcomes. This increased confidence resulting from a variety-of-evidence inference can be interpreted according to the aims of the scientific endeavor: the empirical virtues of the independently well-confirmed models that vary (in our case, they are sketched by the **Mi**s, supported by the **OEi**s, see Fig. 7), where the models can play the role of evidential "experiments," raise the confidence connecting the causal core, **GHG**, of the model-type, **M**, to the 20th and early 21st Century warming outcomes, to specified degrees and respects, and assuming a particular purpose (see Section 1).[18] This contradicts Houkes and Vaesen's conclusion that robustness cannot add "credibility for the model family and, therefore, warrant for the instantiation of the causal structure" (2012, p. 351; see pp. 351–354).

Thus, we can add a further point to our previous summary list. From points (a) through (e), we had concluded that generally, it is against the background of differing model constructions, M1, M2, … Mn, within the model family, **M**, that the radiative causal core, **GHG**, occurs and causes the robust property, **T**, to appear, and it is the degree of this variety-of-instances (the differing model constructions) for which the model has been verified, that affects how confident we should be in the causal connection, and its performance in accurate, and robust, prediction/retrodiction of **T**. Now,

[17] In Weisberg' view, robustness analysis "does not confirm robust theorems; it identifies hypotheses whose confirmation derives from the low-level confirmation of the mathematical framework" (2006, 741). Still, I see my proposal as a friendly extrapolation of the Levins–Wimsatt–Weisberg and Odenbaugh approaches cited above, especially when we focus on the causal core of the model-type.

[18] Note that we still need to keep variety of evidence separate from model robustness. Model robustness is not equivalent to variety of evidence, rather it is supported by variety of evidence reasoning in part of its inferences.

we add: (f) Model robustness involves inferences from patterns of models and evidence: we infer from the facts that a number of randomly chosen, distinct, *independently well-confirmed* models share a common causal factor or core, and all have some common consequence **T**, where **T** is found to be empirically supported by observations, to the increased belief or confidence that **T** is an empirically adequate description for given purposes in a given context, and that **T** is (partially) caused by or attributable to the common causal factor or core, in combination with *any* of the various independently-supported assumptions, **Ai**s, of a given model.

7. Alternative types of philosophical robustness

Let us finally compare the proposed 'model robustness' with other ideas. Woodward, in his very useful (2006) review of varieties of robustness, writes of 'inferential robustness,' which may seem to closely resemble part of what I describe here. The issue concerning 'inferential robustness' starts with a set of data, D, which we would like to use to reach some conclusion, S, "about the truth of some hypothesis or the value of some parameter of interest" (Woodward, 2006, p. 219). Woodward writes, "[s]uppose that doing so requires use of additional assumptions, that there are a number of different, competing possibilities Ai regarding these assumptions, and that available background knowledge provides no strong reason to prefer one of the Ai over the others." Woodward notes that many authors have argued that if for a range of Ais, the dataset D reliably produces S, then this is a strong reason to believe S, which he takes as a case of his 'inferential robustness.' Earlier, I called cases like this, 'climate-model-versions' of 'measurement robustness,' since a variety of models were used to infer a result without reference to a causal core. In either case, it is clear that the central difference with 'model robustness' is that 'inferential robustness' centers around convergent model *predictions* or *parameter values*, without also focusing on the *causal cores* of models and especially the *observational evidence* supporting model assumptions or the causal cores In model robustness, *all of* the causal cores, empirical support, and predictions are considered.

Let us now consider Woodward's 'derivational robustness,' in which we have a model that predicts observed facts, T. The model contains some assumption, A, which might concern the value of a parameter, x, or the relation between two parameters, x and y, and A is involved in the derivation of T.

Suppose that A were replaced by a different assumption, A′, as shown above in Fig. 8, such as a different value for x, or (x and y). Is it still possible to derive the same result, T (or approximately T)? Or would the model, with A′, derive a different outcome, T'? If the former, then the model might be thought of as "providing a 'robust' derivation" of T. If the latter, then the "model might be thought of as sensitive or non-robust with respect to the derivation of [T] from A″ (Woodward, 2006, p. 232). This approach comes closer to my notion of 'model robustness' and LWWO's notions, since it involves varying parameters of the model; nevertheless, because it does not emphasize or discuss independent *empirical evidence* for those varied parameter values, it differs fundamentally from the proposed 'model robustness.'[19]

Moreover, we can see, from Fig. 8, that Woodward is using the same model for the comparisons of the assumptions, not different models, as I do in 'model robustness.' (Compare with Fig. 6.) Also, Woodward does not discuss a core of the model, because he is not discussing a model-type. Finally, the focus is on comparing the *predictions* or *retrodictions* of the two models; this is indeed one aspect of model robustness, but it also focuses on the causal focus or core of the model-type, all absent in Woodward's derivational robustness.

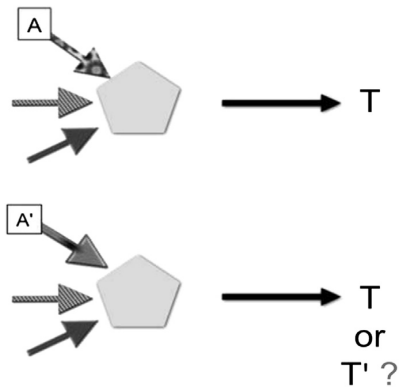

Fig. 8. Derivational robustness (Woodward, 2006, pp. 231–233).

These individual points may be best appreciated with further reflection on Fig. 8. With derivational robustness, one model has assumption A, the other has assumption A′, and still both models may produce T as their outcomes. So far, the two types of robustness, derivational and model robustness seem identical. But here the analysis for derivational robustness officially stops: the model is called 'derivationally robust.' But the model robustness type of reasoning and analysis goes several steps further: first, the empirical, observational, or experimental evidence for assumptions A and A′ are collected, if available; the prediction/retrodiction of the model, T, is tested for empirical validity; both models are analyzed, to determine if there is any shared causal factor (or 'core') which is essential to model robustness; and any observational and experimental evidence for this model 'core' is collected, if there is one. This evidence for the two models and their common core, if there is one, is then evaluated using any and all available tools, including a variety of evidence approach. Inferences may then be made regarding the common core of the two models, if there is one, based on the available evidence from *all of* successful predictions/retrodictions *and* independent evidence for the model assumptions, A and A′, *and* a core. We can see, then, that model robustness involves a great deal more evidential and structural detail about the models than does derivational robustness.

In concluding this section, let us return to measurement robustness, i.e., the reduction of error by repetition in independent contexts, where the exemplar case is Avogadro's number. As we discussed, in order for reasoning about measurement robustness to go through, in the case of climate modeling, the models under consideration need to be relevantly distinct or independent (see Section 2). Model robustness, in contrast, does *not* require an equivalent level of independence as the much-discussed measurement robustness, and in fact, expects a certain level of conformity, in that, e.g., all models in the model family must share the **GHG** causal core. In fact, we might anticipate increased unity and accuracy in their portrayal of this causal core, as models develop over time. Still, differences between the various models from different modeling groups are sometimes inevitable, given the entrenched nature within groups of some of the aspects—such as certain parameterizations—of the models. In other words, certain levels of *both* model independence, as well as model conformity

[19] Kuorkikoski and Lehtinen (2009) explore the confirmatory properties of derivational robustness, which under their analysis shares its inferential virtues with measurement robustness (pp. 559ff).

and dependence, are assured due to the social and epistemic conditions of model construction (Gates et al., 1999; Gleckler et al., 2008; Hegerl et al., 2007; Knutti, 2008; Muller & Von Storch, 2004).

8. Conclusion

Let us begin our conclusion by looking further at what the climate scientists themselves say. In a paper on best modeling practices as guidelines for the IPCC, a panel of top climate scientists, led by Knutti, Abramowitz, et al. (2010),[20] advised that a robust or convergent outcome from a collection of models may not necessarily indicate truth or empirical strength, echoing the philosophers. Nevertheless, they stated that if a group of models represents relevant *processes* in truly *different* ways, and these multiple models *still produce the same outcome*, e.g., **T**, then "*confidence in a result [i.e., model outcome] may increase*"[21] (Knutti, Abramowitz, et al., 2010, p. 10, emph. added). Confidence will also increase in cases when the processes "that determine the result are well understood," where this "understanding" includes empirical support for this robust modeling of the causal processes, such as in the case of greenhouse gases (Knutti, Abramowitz, et al., 2010, p. 10). These are precisely the sorts of informal inferences I have characterized as 'model robustness.'

In essence, 'model robustness' involves the standard convergence of predictions/retrodictions of multiple instantiations of variants of the model-type, as well as exploration and empirical confirmation of an array of empirical model assumptions, which can be seen as aspects of random, well-supported experiments, when we use variety of evidence inferences to support the core structure. Thus, a ***confirmational aspect of robustness***, tying model robustness to the reliability of causal aspects of the model, as well as to successful predictions or retrodictions, may be in place. This *model robustness* I have outlined is exemplified in distinctive, common, and significant types of inference in climate science. It is also appropriate for evolutionary and ecological models (Levins, 1966, 2006; Odenbaugh, 2011), and for other sciences such as physics and economics, especially when their models appeal to empirical evidence for the model assumptions (Cartwright, 1991; Odenbaugh & Alexandrova, 2011).

Within philosophy of science, robustness has most frequently been considered a heuristic but non-confirmatory virtue. Yet the practice of contemporary climate science has increasingly relied on robustness inferences. This practice presents a worthwhile challenge for philosophers of science, to analyze whether robustness in some forms and contexts can be confirmatory. Here, I have sketched how the reasoning from climate scientists about both convergent results concerning past and present climate, and the causal cores of the models producing those convergent results, can be understood as a case of variety-of-evidence inference using the independently well-supported, varied models as random, experiments, in order to affirm its confirmatory power. I focus on one set of examples of such reasoning, in which the shared and empirically sound predictions/retrodictions of a family of diverse and independently-supported greenhouse gas models based on an independently-and experimentally-supported common causal core are taken to increase confidence in that causal core as a good explanation of the robust and verified model predictions/retrodictions (Knutti, Abramowitz, et al., 2010). I conclude that philosophers should break the habit of denying that robustness is confirmatory, and understand and analyze this case as confirmatory, just as the climate scientists do.

[20] Thanks to Reto Knutti and Gabi Hegerl for discussion on this matter.

[21] It is understood that these parameterizations are empirically supported in the models, to some extent or another (Lloyd, 2009, 2010; Pers. Comm. climate scientists Reto Knutti, Jeffrey Kiehl).

Acknowledgments

I would like to thank climate researchers Caspar Ammann, Jeffrey Kiehl, Doug Nychka, Claudia Tebaldi, Warren Washington, Tom Wigley, and especially Linda Mearns, of the National Center for Atmospheric Research, as well as Richard Somerville, of the Scripps Institute, James Annan of the Research Institute for Global Change, Japan Agency for Marine-Earth S&T, Gabriele Hegerl of the University of Edinborough, Reto Knutti, of the ETH Zurich, Myles Allen of Oxford University and UK's Met Office, Peter Stott, of the UK's Met Office, and Francis Zwiers, of the Pacific Climate Impacts Consortium, for their assistance regarding climate models; all mistakes are of course my own. I also thank Brenden Fitelson, Ron Giere, Andrew Hamilton, Ryan Ketcham, Noretta Koertge, Gordon McOuat, Jay Odenbaugh, Wendy Parker, Zachary Pirtle, Trevor Pearce and his colleagues at the University of Chicago, Kari Theurer, Trin Turner, Sean Valles, Bas van Fraassen, Michael Weisberg, Eric Winsberg, and especially Steve Lawrie and Jim Woodward for their helpful comments.

References

Annan, J., & Hargreaves, J. C. (2011). Understanding the CMIP3 multi-model ensemble. *Journal of Climate, 24*, 4529-4538.
Braconnet, P., Otto-Bliesner, B., Harrison, S., Joussaume, S., Peterchmitt, Y., Abe-Ouchi, A., … Zhao, Y. (2007). Results of PMIP2 coupled simulations of mid-Holocene and last glacial maximum – part 1: Experiments and large-scale features. *Climate of the Past, 3*, 261-277.
Calcott, B. (2011). Wimsatt and the robustness family: Review of Wimsatt's re-engineering philosophy for limited beings. *Biology and Philosophy, 26*, 281-293.
Cartwright, N. (1991). Replicability, reproducibility and robustness: Comments on Harry Collins. *History of Political Economy, 23*, 143-155.
Coyne, J. (1999). *Why evolution is true*. New York, NY: Viking.
Culp, S. (1995 Sep). Objectivity in experimental inquiry: Breaking data-technique circles. *Philosophy of Science, 62*(3), 438-458.
Edwards, P. (2010). *A vast machine*. Cambridge, MA: MIT Press.
van Fraassen, B. C. (1980). *The Scientific Image*. Oxford, UK: The Clarendon Press.
van Fraassen, B. C. (2004). Science as representation: Flouting the criteria. *Philosophy of Science, 71*, 794-804.
van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford, UK: Oxford University Press.
Gates, W. L., Boyle, J. S., Covey, C., Dease, C. G., Doutriaux, C. M., Drach, R. S., … Williams, D. (1999). An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I). *Bulletin of the American Meteorological Society, 80*, 29-55.
Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science, 71*, 742-752.
Giere, R. N. (2006). *Scientific perspectivism*. Chicago, IL: University of Chicago Press.
Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research, 113*, D06104.
Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus, 57A*, 219-233.
Hegerl, G. C., & Zwiers, F. W. (26 May 2011). Use of models in detection and attribution of climate change. *WIREs Climate Change*. http://dx.doi.org/10.1002/wcc.121.
Hegerl, G. C., Zwiers, F. W., Braconnet, P., Gillett, N. P., Luo, Y., Orsini, J. A. M., … Stott, P. A. (2007). Understanding and attributing climate change. In S. Solomon, D. Qin, M. Manning, … (Eds.), *Climate change 2007: Report of the Intergovernmental Panel on Climate Change* (pp. 663-745). New York, NY: Cambridge University Press.
Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., Van der Linden, P. J., Dai, X., … Johnson, C. A. (2001). *Climate change 2001: The scientific basis*. Cambridge, UK: Cambridge University Press.
Houkes, W., & Vaesen, K. (2012). Robust! Handle with care. *Philosophy of Science, 79*, 345-364.
Hulme, M. (2009 May). On the origin of 'the greenhouse effect': John Tyndall's 1859 interrogation of nature. *Weather, 64*, 121-123.
Katzav, J. (2011). Should we assess climate model predictions in light of severe tests? *EOS, 92*, 2-3.
Kiehl, J. T. (2007). Twentieth century climate model response and climate sensitivity. *Geophysical Research Letters, 34*, L22710.
Knutti, R. (2008). Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society A, 366*, 4647-4664.

Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P. J., Hewitson, B., & Mearns, L. (2010). Good practice guidance paper on assessing and combining multi model climate projections. In T. F. Stocker, D. Qin, G.-K. Plattner, … (Eds.), *Meeting report of the Intergovernmental Panel on Climate Change Expert meeting on assessing and combining multi model climate projections*. IPCC Working Group I Technical Support Unit, Bern, Switzerland: University of Bern.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate, 23*, 2739-2758. http://dx.doi.org/10.1175/2009JCLI3361.1.

Kuorkikoski, J., & Lehtinen, A. (2009). Incredible Worlds, Credible Results. *Erkenntiniss, 70*, 119-131.

Lambert, S. J., & Boer, G. J. (2001). CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dynamics, 17*, 83-106.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist, 54*(4), 421-431.

Levins, R. (1968). *Evolution in changing environments*. Princeton, NJ: Princeton University Press.

Levins, R. (1993). A response to Orzack and Sober: Formal analysis and the fluidity of science. *The Quarterly Review of Biology, 68*, 547-555.

Levins, R. (2006). Strategies of abstraction. *Biology and Philosophy, 21*, 741-755.

Lloyd, E. A. (1994). *The Structure and Confirmation of Evolutionary Theory*. Princeton, NJ: Princeton University Press.

Lloyd, E. A. (2009). Varieties of Support and Confirmation of Climate Models. *Proceedings of the Aristotelian Society Supplementary Volume LXXXIII*, 217-236.

Lloyd, E. A. (2010). Confirmation and Robustness of Climate Models. *Philosophy of Science, 77*, 971-984.

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

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., & Taylor, K. E. (2007). The WCRP CMIP3 multimodel dataset – A new era in climate change research. *Bulletin of the American Meteorological Society, 88*, 1383-1394.

Muller, P., & Von Storch, H. (2004). *Computer modelling in atmospheric and oceanic sciences*. New York, NY: Springer.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature, 430*, 768-772.

Odenbaugh, J. (2011). True lies: Realism, robustness, and models. *Philosophy of Science, 78*, 1177-1188.

Odenbaugh, J. Ms. Building trust, removing doubt? Robustness analysis and climate modeling (under review).

Odenbaugh, J., & Alexandrova, A. (2011). Buyer beware: Robustness analysis in economics and biology. *Biology and Philosophy, 26*, 757-771.

Orzack, S. H., & Sober, E. (1993). A critical assessment of Levins's 'The strategy of model-building in population biology' (1966). *Quarterly Review of Biology, 68*(4), 533-546.

Parker, W. S. (2006). Understanding pluralism in climate modeling. *Foundations of Science, 11*, 349-368.

Parker, W. S. (2010). Comparative process tracing and climate change fingerprints. *Philosophy of Science, 77*, 1083-1095.

Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science, 78*, 579-600.

Pirtle, Z., Meyer, R., & Hamilton, A. (2010). What does it mean when climate models agree? A case for assessing independence among general circulation models. *Environmental Science and Policy, 13*(5), 351-361.

Raisenen, J. (2007). How reliable are climate models? *Tellus, 59A*, 2-29.

Randall, D. A., & Wielicki, B. A. (1997). Measurements, models and hypotheses in the atmospheric sciences. *Bulletin of the American Meteorological Society, 78*(3), 399-406.

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., … Taylor, K. E. (2007). Climate models and their evaluation. In S. Solomon, D. Qin, M. Manning, … (Eds.), *Climate change 2007: Report of the Intergovernmental Panel on Climate Change* (pp. 589-662). New York, NY: Cambridge University Press.

Rykiel, E. J., Jr. (1996). Testing ecological models: The meaning of validation. *Ecological Modelling, 90*, 229-244.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Shackley, S. (2001). Epistemic lifestyles in climate change modeling. In C. A. Miller, & P. N. Edwards (Eds.), *Changing the atmosphere: Expert knowledge and environmental governance* (pp. 107-133). Cambridge, MA: MIT Press.

Washington, W. M., & Parkinson, C. L. (2005). *Introduction to three-dimensional climate modeling*. New York, NY: University Science Books.

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science, 73*, 730-742.

Weisberg, M., & Reisman, K. (2008). The Robust Volterra principle. *Philosophy of Science, 75*(1), 106-131.

Wimsatt, W. (1981). Robustness, reliability and overdetermination. In M. Brewer, & B. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 124-163). San Francisco, CA: Jossey-Bass.

Wimsatt, W. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.

Winsberg, E., & Lenhard, J. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Modern Physics, 41*(3), 253-262.

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology, 13*, 219-240.