Confirmation and Climate Models
Elisabeth A. Lloyd and Wendy S. Parker

I — Elisabeth A. Lloyd
Varieties of Support and Confirmation of Climate Models

Today's climate models are supported in a couple of ways that receive little attention from philosophers or climate scientists. In addition to standard 'model fit', wherein a model's simulation is compared to observational data, there is an additional type of confirmation available through the variety of instances of model fit. When a model performs well at fitting first one variable and then another, the probability of the model under some standard confirmation function, say, likelihood, goes up more than under each individual case of fit alone. Thus, two instances of fit of distinct variables of a global climate model using distinct data sets considered collectively will provide stronger evidence for a model than either one of the instances considered individually. This has consequences for model robustness. Sets of models that produce robust results will, if their assumptions vary enough and they each are observationally sound, provide reasons to endorse common structures found in those models. Finally, independent empirical support for aspects and assumptions of the model provides an additional confirmational virtue for climate models, contrary to what is implied by some current philosophical writing on this topic.

Introduction. Building on the earlier work of Fritz Rohrlich (1990), Paul Humphreys (1995), and Ian Hacking, Eric Winsberg (1999*a*; 2003) has made a forceful case that we need an epistemology of simulation as well as an understanding of how simulation models are constructed and confirmed. We need such an epistemology, he argues, in order to develop standards for deciding when the results of the scientific technique of simulation have some degree of reliability. Winsberg details a number of the types of evidence the simulations can accumulate in relation to both theory and observational data, but more work remains to be done (see also Parker 2006; 2008). Meanwhile, completely independently, the need has evolved for looking at the inferences involved in developing, testing, and confirming climate models (simulations). The issue of the nature and degree of support for climate models is of timely concern because of the role of climate models in understanding present climate

Proceedings of the Aristotelian Society Supplementary Volume lxxxiii
doi: 10.1111/j.1467-8349.2009.00179.x

and projecting future climate. The models' reliability is frequently a focus of concern and question. In fulfilling part of Winsberg's mandate, I shall focus on the nature of the evidential support for the largest and most complex climate simulation models, and some reasons we may have for thinking them confirmed with relation to observational evidence.[1]

When discussing models and modelling, below, I assume that the simulations (models) are intended to represent particular aspects of the climate system, and that this representation will be judged as adequate or not relative to the purposes of the modellers or those who use the models (van Fraassen 2004, p. 794; Giere 2004; Parker 2006). One might require, for example, a certain level of confirmation or empirical support of the model in order for it to be used in projecting future climate, and particularly for its results to be of use in various climate policy contexts.[2] The forms of confirmation I'll consider will be fit, variety of evidence, robustness, and independent support for aspects of the model.[3]

I

Global Climate Models. Because the Earth rotates and the sun heats them, both the atmosphere and the ocean circulate, carrying heat around the globe and away from the equator and towards the poles on currents of air and water. Climate simulation modelling of this global atmospheric and ocean circulation involves the articulation of multiple physical theories simultaneously. The models represent mathematically the physical movements of gaseous and liquid masses, and the transfer, reflection, and absorption of energy. In the largest, most complex, Global Circulation Models (GCMs)—the variety of models we focus on here, of which there are about twenty to thirty 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

[1] This complements the work of others, who have analysed reasons for uncertainty regarding the models' results, e.g. Petersen (2000; 2006); Edwards (1999; 2001); Parker (2006; 2008).

[2] See Shackley (2001) for discussion of different goals and their association with different biases in climate modelling.

[3] There are many other issues relating to the confirmation of climate models; see, for example, Randall et al. (2007), Randall and Wielicki (1997), Parker (2006; 2008).

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

theories plus a 'moisture' equation represent the atmospheric system in terms of pressure, temperature, and density. Similar 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 and Parkinson 2005).

The state of the climate cannot, however, be fully derived from the physical theories, since we do not know the full set of physical laws guiding the system from one state to another at the scales governing all processes. Moreover, only the largest-scale processes are explicitly presented in the model, such as the movements of the air masses. The global atmosphere is divided into a grid, about 100–300 km on each side and 10–26 layers thick. The atmospheric processes are calculated within and across these volumes; whatever occurs within these large volumes is handled by parameterizations within the model, i.e. methods that attempt to take into account the important impacts of these 'subgrid processes' without simulating them explicitly. Important examples of parameterized subgrid processes include cloud formation and convection, each of which has significant effects on climate. Parameterizations are unavoidable in a model of global scale with phenomena of the multiple scales of climate, which ranges from the interactions going on at the molecular level all the way up to massive air motions.

Note that climate models are not themselves tests of the physical theories from which they are derived. Rather, they are composite instantiations or applications of them; physical theories converge in the single climate simulation model. Climate science requires such simulation models to properly represent how all these physical forces interact. No one expects to use climate models to test the validity of the principle of conservation of mass, nor would their success be used to confirm that principle. But theoretical knowledge is only one ingredient that modellers use to produce a model of the climate phenomena. The simulations also involve assumptions of a less theoretical nature, boundary values, numerical methods, and intuition, as well as the approximations already mentioned (Winsberg 2003).

In one sense, a Global Circulation Model itself embodies a form of climate theory, presenting, as it does, how climate might work, and including our best understandings of some of the various proc-

Proceedings of the Aristotelian Society Supplementary Volume lxxxiii
doi: 10.1111/j.1467-8349.2009.00179.x

esses involved in climate. Each model presents a hypothesis about how climate works, and it is the simulation model itself that is confirmed or tested, and not usually the theory or equations from which it was created. As comprehensive models improve, they become the 'primary tools by which theory confronts observations', notes Isaac Held (2005, p. 1609). 'A model essentially embodies a theory', as David Randall and Bruce Wielicki said (1997, p. 400; see Murphy et al. 2004, p. 768; Lambert and Boer 2001, p. 83).

II

Fit. The most straightforward method of testing and confirming a model is to simulate (predict or retrodict) an outcome and compare the simulated state against observational data, which I call simply 'model fit'. As S. George Philander says, 'The best test for a model is its ability to simulate Earth's current and past climates' (1998, p. 199; see Rykiel 1996, p. 236). The simplest of these tests compare global mean temperature outcomes from a model against global mean temperatures estimated from observations (e.g. Meehl et al. 2004). Models are compared in their ability to simulate the observational records, and their ability to simulate the twentieth-century warming serves as a benchmark for model viability.

Other tests involve the deeper past. For instance, models are checked against the last millennium or two to test whether they can handle the time evolution over centuries of the climate system driven by the changes in forcings (causes or forces, such as solar warming) provided in the model (e.g. Mann et al. 2007). Further tests are done using the Last Glacial Maximum (*ca* 21,000 years before present), which was conceived as an experiment to examine climate response to the presence of large ice sheets, cold oceans, and lowered greenhouse gas concentrations. The mid-Holocene (*ca* 6,000 years before present) simulation tests the models' response to changes in solar radiation (Braconnot et al. 2007). The fact that the climate models are able to simulate the main climate variables in these past events with their very different forcings and feedbacks is taken as evidence that they demonstrate a good grasp of the fundamental physics of some of the forces that cause climate to change at global scales. This is also taken to show that the forces represented in models can handle values outside the ranges encountered recently. With

Proceedings of the Aristotelian Society Supplementary Volume lxxxiii
doi: 10.1111/j.1467-8349.2009.00179.x

these accomplishments in hand, the models should be better able to project future climate change.

There are a variety of ways to measure good fit in climate models, starting with simple statistical measures of the global mean temperature and moving on to measures of the monthly, seasonal, or regional temperatures. The WGNE (Working Group on Numerical Experimentation) of the World Climate Research Program has encouraged the development of standard diagnostics and established benchmark experiments through its model intercomparison projects (Randall et al. 2007; Gleckler et al. 2008).

III

Variety of Evidence. As just mentioned, there are many more ways of testing model fit beyond testing whether global climate models can simulate or predict global mean temperature accurately. For example, the climate scientists may want to know whether the variation either within a year or across years simulated or predicted by a model conforms to observations. The key for this section is to consider the importance of the fact that other variables are also accurately simulated by the model at the same time as, say, global mean temperature. For instance, models show significant skill in representing mean climate features such as large-scale distributions of the other variables of precipitation, radiation, wind, oceanic temperatures, and currents, in addition to temperature (Randall et al. 2007, p. 600). Additionally, models can simulate patterns of variability, in which the model is compared to changes in the climate variable over months or seasons. The global models can simulate patterns of variability such as the advance and retreat of major monsoon systems, seasonal shifts of temperatures, storm tracks, and rain belts (Randall et al. 2007). Gleckler et al. (2008, p. 12) concluded from their cross-model comparison that representation of the modes of variability was important to indicating whether a model had 'really captured the physics of the climate system'.

In an especially nice example of model fit, Santer et al. (2003) compared simulated and observed change in the height of the tropopause, the transition in the atmosphere that separates the lower atmosphere from the upper atmosphere. The model results were a good match to the observed changes in tropopause height. This is

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

not the sort of result that can be 'tuned', or adjusted in the model ahead of time before testing, since information about the change in the height of this boundary is not used in the development of models. Tuning involves the calibration or adjustment of the model parameters to produce a better fit with observations.

In another very important set of examples of good fit, the models are able to simulate or predict recent changes in the vertical temperature profile of the atmosphere, going from the surface warming up through the lower and upper tropospheric warming, and then cooling above. This pattern of surface and tropospheric warming combined with upper atmospheric cooling is a signature predicted from the greenhouse effect, and provides an especially rigorous spatial set of tests for the models (Karl et al. 2006; Allen and Sherwood 2008).

All these various cases of fit for the GCMs may seem to demonstrate the same thing, the fact that the models could predict the global mean temperature, but this is not so. Gleckler et al. took the twentieth-century simulations from a large model intercomparison study (CMIP3), compared the models to a group of data sets, and found that '[a]ccurate simulation of one variable does not in most cases imply equally accurate simulation of another' (Glecker et al. 2008, p. 8). Hence, if the model were in fact to simulate accurately another variable besides global mean temperature, this would be added information and additional credibility to the model. In other words, the fact that the model had simulated two variables, or been supported by a variety of instances of fit, would count in its favour.

This conclusion is supported by work by Branden Fitelson, who showed, using a Bayesian probabilistic framework based on Peircean notions, that two pieces of confirmatory evidence that are independent will provide stronger confirmation than either one of them provides individually (Fitelson 2001, p. S131). To say that the confirmatory evidence is independent means that the degree to which the first evidence, or instance of fit, supports the model doesn't depend on whether the second evidence or instance of fit has already occurred.[4] Thus, two instances of fit of distinct variables of a GCM using distinct data sets, for example, considered collectively will provide stronger evidence for a model than either one of the in-

[4] That is, given pieces of evidence E1 and E2, and hypothesis H, E1 and E2 are mutually confirmationally independent regarding H according to c iff both $c(\text{H}, \text{E1} \mid \text{E2}) = c(\text{H}, \text{E1})$ and $c(\text{H}, \text{E2} \mid \text{E1}) = c(\text{H}, \text{E2})$, where c is a confirmation function, such as a likelihood ratio measure (Fitelson 2001, p. S125).

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

stances considered individually. We have fulfilled these conditions when, for example, the ocean heat variable is tested against an ocean temperature observational data set and the pressure variable at given locations is tested against the observed pressures. Thus, a model with many instances of fit is much better supported and has a higher probability under a preferred confirmation function than a model with only one or two instances.[5]

I find this point underappreciated in the literature about climate modelling. In the Intergovernmental Panel on Climate Change (IPCC) report, for example, the various successes of climate models are listed, but the fact that this collectivity of successes provides stronger evidence than any of the instances considered one by one is not emphasized (e.g. Randall et al. 2007, pp. 591–3, 600–1). Rather, it looks as if 'the models are good at this, or good at that', but it signifies very little collectively. The summary documents of successes and challenges coming out of model intercomparison studies similarly make little note of this strength (e.g. Gates et al. 1999). Thus, these models appear to be better confirmed than they are sometimes given credit for, in the context of laundry-list characterizations of successes and weaknesses.

IV

Robustness. Consider the robust result that all global climate models simulate global warming of 0.5–0.7 degrees centigrade within 25% accuracy for the twentieth century (Houghton et al. 2001).[6] Jeffrey Kiehl (2007, p. 1) comments, 'This is viewed as a reassuring confirmation that models to first order capture the behavior of the physical climate system and lends credence to applying the models to projecting future climates.' A small variance among models, say Steve Lambert and George Boer, 'supports the assumption that they are capturing the processes that govern that variable and hence its climate' (Lambert and Boehr 2001, p. 88).

[5] John Earman, using a different approach, analyses how this type of variety of evidence increases support for a model or hypothesis (1992, pp. 77–9; see Lloyd 1988/1994). On Earman's approach, the variety of evidence is relativized to the background knowledge, including surrounding scientific theories. Earman describes how choosing a next experiment that is different can boost confirmation more than a next experiment that is similar to past ones.

[6] There are numerous other cases of robust findings, e.g. palaeoclimate models (Braconnot et al. 2007, p. 226) or perturbed physics models (Murphy et al. 2004).

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

Michael Weisberg has recently offered a philosophical account of robustness analysis, as a method for determining 'which models can reliably be used in explanations' (Weisberg 2006, p. 731). His approach is designed to distinguish whether a result depends on the 'essentials' of the model, or on its other assumptions. Weisberg's first step is (1) to study several models of the same phenomenon: if we find them leading to the same result, that result is a 'robust property'. We then (2) analyse the models, looking for a common structure that generates the robust property. We then link the two together into the conditional form of a robust theorem: '*Ceteris paribus*, if [common structure] obtains, then [robust property] will obtain.' In the third step, (3) we give an empirical interpretation of the mathematical structures combined in the conditional form. (Note that in the climate case, we've usually already got empirical assignments for our equations and variables that make up the common structure and robust property.) Finally, (4) we can conduct stability analyses of the robust theorem, in order to ascertain the conditions under which the connection between the common structure and the robust property wouldn't hold (Weisberg 2006, p. 738).

Suppose we were to apply Weisberg's steps to the collection of climate models:[7] we find that in all of them there is a significant role played by greenhouse gases in the late twentieth-century warming of the global climate, and that these are linked to the surface temperature rising in the equations, despite the fact that climate models vary in their assumptions about other aspects of climate. Thus, we would have an analysis isolating greenhouse gases linked to temperature rise (the common structure), and a robust theorem linking greenhouse gases to the robust property, the outcome of rising global mean temperature. But how can we be sure that greenhouse gases are the relevant cause? Weisberg then makes an implicit appeal to the variety of evidence. In order to infer to causes in the real world, he writes, 'The key comes in ensuring that a sufficiently heterogeneous set of situations is covered in the set of models subjected to robustness analysis' (Weisberg 2006, p. 739). If a sufficiently heterogeneous set of models for a phenomenon has the common structure, he continues, 'then it is very likely that the real-world phenomenon has a corresponding causal structure'. Moreover, this would allow

[7] Ryan Muldoon (2007, pp. 880–2) applies Weisberg's robustness analysis to climate modelling in a different way, showing how robustness strategies can be used to address the array of issues. He doesn't address the particular inference I consider here.

Proceedings of the Aristotelian Society Supplementary Volume lxxxiii
doi: 10.1111/j.1467-8349.2009.00179.x

us to infer that when we saw the robust property in a real system, 'it is likely that the core structure is present, and that it is giving rise to the property' (Weisberg 2006, p. 739).

In my view, Weisberg is appealing to a variety of evidence argument here, because he is appealing to a range of instances of fit of the model over different parameter values, parameter spaces or laws.[8] It is against this background of differing model constructions that the common structure occurs and causes the robust property to appear, and it is the degree of this variety of fit for which the model has been verified that determines how confident we should be in the causal connection. Weisberg deems this to be a part of robustness analysis, but it is a distinct inference from the usual robustness analysis, which involves inferences about the robust property, and not about the model(s) per se (e.g. Woodward 2006; Staley 2004). As such, it fits more naturally as a subtype of variety of evidence inferences, to which we can apply the probabilistic result demonstrating its confirmatory value.

Continuing our application of Weisberg's analysis to our case, we do find that the models covered a wide range of assumptions and conditions, and they all have this common structure of greenhouse gas causation. Hence, on his analysis, it is very likely that the real-world phenomenon has a corresponding causal structure (Weisberg 2006, p. 739; Muldoon 2007, p. 882).[9] Therefore, we could infer that greenhouse gas concentration increases cause global warming in the real world, as the attribution studies have also shown (Hegerl et al. 2007).[10]

But someone might respond: 'Of course the models will simulate twentieth-century warming; that's what they were designed and tuned to simulate or predict.' We may then want to look toward a more fine-grained type of inference the modellers are concerned

[8] In Weisberg's recent paper with Ken Reisman (forthcoming), they refer to a part of this range of variety as 'parameter robustness analysis', in which the parameter value is varied across a given range, also known as 'sensitivity analysis'. The variation of laws is called 'structural robustness analysis'.

[9] On Jim Woodward's categorization, inferring causal relationships requires manipulation and experimentation, which are generally impossible for climate, although possible for climate models (Woodward 2006, p. 235). Our inference is also different from Woodward's 'inferential robustness' (using the same data but different assumptions, one reaches the same conclusion, *S*), since our focus is on the model structure, not the robust property (Woodward 2006, p. 230).

[10] I would emphasize that this finding is parasitic on—but not reducible to—the empirical adequacy or fit of each individual model; their collection together instantiating a variety of evidence is doing the additional confirmatory work here.

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

with, namely, how much do greenhouse gases affect the temperature rise? Kiehl (2007) performed a study of 11 GCMs and examined the total anthropogenic climate forcing, which includes greenhouse gas components (which tend to warm the climate), and aerosol components (which both warm and cool the climate), as well as climate sensitivity, which is a reflection of how sensitive the climate is to the changes in atmospheric components.

He found that the models differ by as much as a factor of three in the aerosol forcing, a factor of two in climate sensitivity, and yet they still simulated the twentieth-century observations well. This raises a concern about whether models that differ so much can appropriately be used for making climate projections, given the uncertainties. Can we trust what they say about the causes and future of climate change, given that what one model says about the detailed forcings is rather different from what another model says?

On Kiehl's analysis, the fact that the models have been benchmarked against the present climate state means that they are better prepared to project the future climate state; but more importantly, the differences between the models will not make much of a difference when contrasted against the large contribution expected from greenhouse gas forcings in the future. Thus, he concludes, they do not invalidate the application of current climate models to projecting future climate (Kiehl 2007). We may also conclude further that the models are robust in their representations of greenhouse gases, and that given a robustness analysis and the inferences that follow from such an analysis, we would be safe in making inferences regarding greenhouse gas contributions to global climate change. The variation in the models provides the background against which the robust cause is evaluated; repeated instances of fit even against different assumptions, such as different aerosol forcings, count in favour of the greenhouse gas causation itself, as well as its representation in the models. Of additional philosophical significance is that the variety of evidence interpretation of robustness endows it with confirmational significance that it lacks under other interpretations (see esp. Woodward's survey of interpretations of robustness, Woodward 2006; Staley 2004).[11]

[11] Weisberg moves towards this confirmatory virtue, to the extent that he requires that the models possess 'low-level confirmation', as well as the array of background variability. that provides the conditions for variety of evidence. But on his view, robustness analysis 'does not confirm robust theorems; it identifies hypotheses whose confirmation derives from the low-level confirmation of the mathematical framework in which they are embedded' (Weisberg 2006, p. 741).

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

V

Independent Support.

5.1. *Independent Empirical Support for Aspects of the Models.* Climate simulation models can also obtain a degree of confirmation through independent support of various aspects of the model, such as its laws, parameters, and parameterizations. Such confirmation is separate and additional to the confirmation or support that the model receives in virtue of its ability to simulate (predict or retrodict) successfully. This form of confirmation is extremely important in other modelling sciences such as evolutionary biology and ecology (Lloyd 1988/1994; Rykiel 1996; Winsberg 1999*b*), and I will argue in this section that it plays a significant (though contested) role in climate modelling as well (Randall et al. 2007, p. 594).

Simple examples of independent empirical support for a model or a set of models come in the form of a few empirical formulas that are believed, on theoretical grounds, to be universally applicable, and that have been determined or measured empirically, such as the Monin-Obukhov similarity functions, which represent the mixing behaviour of the atmosphere next to the land surface (Randall and Wielicki 1997, p. 399). Similarly, certain constants or parameters are measured once and inserted into many models, such as the physical properties of water and air, e.g. their optical properties. In this manner, values in the models are filled in, and the model is tied to the real world, and thus confirmed. Significantly, this confirmation from independent support of parameter values is distinct from the confirmation from successful prediction.

The most contested form of independent support for climate models pertains to the parameterizations. A parameterization is a mathematical model that calculates the net effect of unresolved (usually subgrid) processes on the processes that are directly calculated in the model. The parameterization can be developed from measurements of the unresolved processes, from which statistics are derived, which describe the net effect on the variables in the climate model. These statistical relationships can then be included in the climate model. Alternatively, deterministic models that simulate the statistics directly can be included in the climate model.

In other words, the parameterizations generally take account of details and interactions at the smaller scale, and translate them into consequences for the variables in the GCM. For example, a new and

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

simple model of aerosol evolution in the lower stratosphere for the volcanic aerosols was developed, based on a new volcanic forcing data set, which takes into account the monthly distribution and latitudinal differences. This newer parameterization led to improvement of the correlation of the GCM with the observations over the twentieth century (Ammann et al. 2003).

Or take the calculations of the radiation effects. In representing radiative transfer in the models, the black body radiation curve is used as well as two basic laws, Lambert's law and Kirchoff's law, which govern the behaviour of radiation in relation to gases. The application of these physical laws involves physical measurements of temperature, water vapour and other radiatively active gases such as carbon dioxide and ozone. These values are determined experimentally in the equations through the Atmospheric Radiation Measurement Program (ARM) and other field programmes, as well as being based on the latest lab studies (Held 2005).

But there are still many uncertainties about the parameterizations. For instance, relatively less is known about how aerosols affect clouds, and this leads to sizeable differences between the models (Kiehl 2007). Other uncertainties arise because the physics surrounding precipitation and clouds is not fully understood for these sub-gridscale processes. But there has been progress in cloud parameterization through specific programmes to develop parameterizations based on observations, such as the GEWEX (Global Energy and Water Cycle Experiment) cloud system study, and CERES (Clouds and the Earth's Radiant Energy Systems) observational programmes. Moreover, the available cloud parameterizations are based on physical theory, some of which has in turn been verified independently in the lab experiments (Washington and Parkinson 2005, p. 101).

According to Washington and Parkinson (2005, p. 98), the key to successful parameterization is the 'formulation of quantitative rules for expressing the location, frequency of occurrence, and intensity of the subgrid-scale processes in terms of the resolvable scale'. Note specifically that they do not require a tie-in with a physical law in order to have a successful parameterization, although they note that some methods are 'more physically sound' than others. Although modellers make parameterizations consistent with the basic laws of conservation of energy, mass, and momentum, which act as constraints, sometimes no further physical laws may be involved.

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

In my view, the fact that the cloud parameterizations are supported through the GEWEX and CERES as well as other observational programmes, and thus are based on specific cloud observational data, as well as being partially supported through lab experiments, means that there is reason to endorse them as independently verified aspects of the GCM that uses them, and therefore as confirmatory for that GCM. Simulation models that have more laws, parameters, and parameterizations (or other aspects of the model) that are independently supported empirically are thus better supported or confirmed than those that have fewer (Winsberg 1999*a*, p. 289; Randall and Wielicki 1997, p. 403; Randall et al. 2007, p. 594). This type of support is additional to the degree of confirmation attained by correct prediction or having the outcome(s) of the model approximately match the observational data. But there are those who appear to diminish the value or role of such empirical support for the models.

5.2. *Philosophical Treatments*. The biggest assumption in philosophical discussion of the issue of independent support of a climate model is that derivation from theory is better than empirical support. One approach to clarifying this issue is taken by philosopher and climate scientist Arthur Petersen (2006), who, following Paul Humphreys (1995), divides models into the 'model structure'—the mathematical equations—and the 'model parameters'—the constants in the mathematical equations—which need to be determined from empirical data (Petersen 2006, p. 22). The 'model structure', or mathematical core of the model, on this view, ought not to be determined or affected by empirical data. This is a particularly problematic position when it comes to parameterizations, because '[parameterization] can be described as the integration of observationally-derived approximation or heuristics into the model core', according to philosopher Paul Edwards (1999, p. 449). Edwards objects to the inclusion of data-laced parameterizations into the models, on the basis that it violates the 'reductionist imperative of the physical sciences' (Edwards 2001, p. 59). In other words, such 'physical science practice normally attempts to explain large-scale phenomena as an outcome of smaller-scale processes' (Edwards 2001, p. 59). But the incorporation of parameterizations that involve large-scale empirical statistical data in the simulations violates this reductionist mandate. It seems that Edwards thinks that climate

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

science must be reductionist in order to meet the standards of the physical sciences, although he never defends this supposition, nor explains how the earth sciences fit in his vision.

Petersen sometimes follows Edwards in showing disdain for independent empirical support for aspects of the model. He complains that most simulation models contain a number of parameterizations 'not fully based on general theory' (Petersen 2000, p. 168). This theoretical basis is one of four key features of 'methodological rigour' listed by Petersen to evaluate climate models. The other features include the assessment of empirical basis and testing, comparison to other simulations, and peer review of simulations. (The IPCC, on the other hand, evaluates models according to a variety of comparisons with observational data and reconstructions of past and present climate, as well as model intercomparisons; theoretical basis or derivation is not emphasized as a criterion for evaluation—Randall et al. 2007, pp. 594–6; see Washington and Parkinson 2005, p. 98, quoted above.)

Petersen says that many simulationists assume that 'the more ad hoc corrections a model contains the worse it is'. (Petersen's 'ad hoc' corrections include parameterizations.) Modellers 'should strive to provide an independent justification for these corrections (preferably through deriving them from theory through approximation). This would ensure that the model is based as much as possible on theory instead of letting the model become nagged by auxiliary hypotheses that are not independently justifiable' (Petersen 2006, p. 38). Note that this assumes that independent justification can only come from theory, and not from empirical support.

But Petersen also wants to bring in the methodology proposed by Randall and Wielicki to counteract the arbitrariness and lack of theoretical basis he sees in some parameterizations. They suggest deriving parameterizations by means of observations and testing the parameterizations for many different conditions. The parameterizations should then be left unchanged when the model is tested as a whole, rather than tuning the parameters interactively, in order to have the outcomes of the model match the observational data (Petersen 2006; Randall and Wielicki 1997, p. 404). But this basically amounts to independent testing and confirmation of aspects of the model, as I argued for in the previous section. So Petersen, as compared to Edwards, seems to hold a gradated view about this issue: he wants the parameterization to be derived by theory; but if that's

Proceedings of the Aristotelian Society Supplementary Volume lxxxiii
doi: 10.1111/j.1467-8349.2009.00179.x

not available, he wants the kind of empirical support that Randall and Wielicki (and I) advocate.

Edwards has further objections to the inclusion of observation-laden parameterizations in the model (in addition to his reductionist one), and therefore to any attempt to see their ties to the world as a form of confirmation of the model. For instance, he thinks that the model needs to be kept completely segregated from the data that are used to test the model. Prima facie, this may seem a fair requirement, but his strict application of the rule is not shared, for example, by those analysing how model confirmation works in other fields involving complex models. Take, for example, Edward Rykiel's analysis of model validation in ecology. Rykiel sets out three basic steps in the model testing and confirmation process, starting with the 'Design' step, in which the scientists develop the model's assumptions based on theory, observations, general knowledge, and intuition. 'Implementation' is the second step: 'Empirically test the model's assumptions where possible.' And 'Operation' is the last, in which the input–output relationships of the model and real system are compared (Rykiel 1996, p. 236). Note that in the first step, observational data are used in the design of the model, and step two amounts to gaining independent empirical support for aspects of the model (see also Lloyd 1988/1994, ch. 8). Significantly, the practice of 'data-splitting'—the procedure where part of a data set is used to build or calibrate a model and the other part used to test it—is designed to deal with just the problem that Edwards is worried about, and is widely used in ecological and climate modelling. The fact that Rykiel's modellers in ecology are also tackling the modelling of extremely complex systems with no single overarching theory may be relevant to this testing and confirmation strategy.

Edwards's third concern about the so-called 'data-laden' parameterizations has to do with their non-theoretical nature, a concern that he shares with Petersen. We might better understand the source of the philosophers' anxiety about the physical derivation of aspects of the model by consulting the paper by Ernan McMullin on idealization cited by Petersen (2006). According to McMullin (1985), it's important to the realism of a claim that, when applying a theory, corrections made to a theory not be ad hoc, and must be well-motivated from the point of view of theory, or suggested by the theory, in order for the theory to be confirmed or validated. If, on the other hand, the corrections were to be ad hoc, the resulting description of

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

reality would be unable to provide any evidence for the truth of the theory.

But applications of the physical theories involved in climate models (such as mechanics or thermodynamics) are not intended to be tests or confirmations of those theories, thus, it seems that the McMullin-type worry about ad-hoc-ness does not apply here. The climate models are, simply, applications of those physical theories, and what is being tested instead is a description of the climate system.[12] I submit that much philosophical concern, in particular that voiced by Peterson and Edwards, is guided by a mistaken emphasis on attempting to ensure a correct lineage for testing physical theories, rather than keeping an eye on the ball, which is in testing climate science descriptions of the climate system, where empirically supported parameterizations actually strengthen climate models. Independent evidence for assumptions and aspects of a model in climate science add reasons to believe that the model is empirically adequate or realistic, which goes beyond the predictive success of the model's outcome.

VI

Conclusion. In sum, today's climate models are supported empirically in several ways that receive little explicit attention. Understanding these ways aids in the philosophical project of unpacking the knowledge claims made on behalf of simulation models and the evidence brought in their favour. Model fit provides a foundation for evaluation of climate models, and a wide variety of metrics are under discussion and applied to evaluate this measure. Variety of evidence provides an important summary that amplifies the importance of collective instances of fit, and acts as a supplement to the robustness of model results that can turn it into a real confirmatory virtue, rather than simply an analysis of model relations. Finally, independent empirical evidence can provide significant support for aspects of the model that are not otherwise defended, and goes

[12] Giving an accurate mathematical characterization of a phenomenon is a significant accomplishment, even if the underlying physics are not yet understood (cf. Edwards 2001). There might be some tension within the community of climate scientists about the importance of giving the physical basis of processes versus giving an empirically adequate accounting of it, which I cannot address here (see Shackley 2001).

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

beyond any support offered by predictive success or good fit. Previous philosophical and scientific examinations of some of these evidential relations, especially in the context of climate models, may have tended to underestimate their importance or force.[13]

Department of History and Philosophy of Science
130 *Goodbody Hall*
Indiana University
Bloomington, IN 47405
USA
ealloyd@indiana.edu

References

Allen, Robert J., and Steven C. Sherwood 2008: 'Warming maximum in the tropical upper troposphere deduced from thermal winds'. *Nature Geoscience*, 1(6), pp. 399–403.

Ammann, Caspar M., Gerald A. Meehl, Warren M. Washington, and Charles S. Zender 2003: 'A monthly and latitudinally varying volcanic forcing dataset in simulations of 20th-century climate'. *Geophysical Research Letters*, 30(12), 1657, pp. .

Braconnot, P., B. Otto-Bliesner et al. 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, pp. 261–77.

Earman, John 1992: *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory.* Cambridge, MA: MIT Press.

Edwards, Paul N. 1999: 'Global Climate Science, Uncertainty and Politics: Data-laden Models, Model-filtered Data'. *Science as Culture*, 8(4), pp. 437–72.

——2001: 'Representing the Global Atmosphere: Computer Models, Data and Knowledge about Climate Change'. In Clark A. Miller and Paul N. Edwards (eds.), *Changing the Atmosphere: Expert Knowledge and Environmental Governance*, pp. 31–65. Cambridge, MA, MIT Press.

Fitelson, Branden 2001: 'A Bayesian Account of Independent Evidence with Applications'. *Philosophy of Science*, 68, pp. S123–S140.

[13] I would like to thank climate researchers Caspar Ammann, William Collins, Jeffrey Kiehl, Doug Nychka, Kevin Trenberth, Tom Wigley, and especially Linda Mearns of the National Center for Atmospheric Research, as well as Richard Somerville of the Scripps Institute, for their assistance regarding climate models; all mistakes are, of course, my own. I also thank Kathryn Carter, Branden Fitelson, Michael Weisberg and Eric Winsberg for their helpful comments.

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

Gates, W. Lawrence, James S. Boyle et al. 1999: 'An Overview of the Results of the Atmospheric Model Intercomparison Project (AMIP I)'. *Bulletin of the American Meteorological Society*, 80(1), pp. 29–55.

Giere, Ronald N. 2004: 'How Models Are Used to Represent Reality'. *Philosophy of Science*, 71, pp. 742–52.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux 2008: 'Performance metrics for climate models'. *Journal of Geophysical Research*, 113, D06104.

Hegerl, Gabriele C., Francis W. Zwiers et al. 2007: 'Understanding and Attributing Climate Change'. In Susan Solomon, Dahe Qin, Martin Manning et al. (eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 663–745. New York: Cambridge University Press.

Held, Isaac M. 2005: 'The Gap Between Simulation and Understanding in Climate Modeling'. *Bulletin of the American Meteorological Society*, 86, pp. 1609–14.

Houghton, John et al. 2001: *Climate Change 2001: The Scientific Basis*. Cambridge: Cambridge University Press.

Humphreys, Paul 1995: 'Computational science and scientific method'. *Minds and Machines*, 5, pp. 499–512.

Karl, Thomas R., Susan J. Hassol, et al. (eds.) 2006: 'Temperature Trends in the Lower Atmosphere. Synthesis and Assessment Product 1.1: US Climate Change Science Program, Washington, DC.

Kiehl, Jeffrey T. 2007: 'Twentieth-century climate model response and climate sensitivity'. *Geophysical Research Letters*, 34, L22710.

Lambert, S. J., and G. J. Boer 2001: 'CMIP1 evaluation and intercomparison of coupled climate models'. *Climate Dynamics*, 17, pp. 83–106.

Lloyd, Elisabeth A. 1988/1994: *The Structure and Confirmation of Evolutionary Theory*. Westport, CT: Greenwood Press. Reprinted Princeton, NJ: Princeton University Press.

Mann, Michael, Scott Rutherford, Eugene Wahl, and Caspar Ammann. 2007: 'Robustness of proxy-based climate field reconstruction methods'. *Journal of Geophysical Research*, D12109.

McMullin, Ernan 1985: 'Galilean Idealization'. *Studies in History and Philosophy of Science*, 16, pp. 247–73.

Meehl, Gerald A., Warren M. Washington, Caspar M. Ammann et al. 2004: 'Combinations of Natural and Anthropogenic Forcings in Twentieth-Century Climate'. *Journal of Climate*, 17, pp. 3721–7.

Muldoon, Ryan 2007: 'Robust Simulations'. *Philosophy of Science*, 74, pp. 873–83.

Murphy, James M., David M. H. Sexton, David N. Barnett et al. 2004: 'Quantification of modelling uncertainties in a large ensemble of climate change simulations'. *Nature*, 430, pp. 768–72.

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

Parker, Wendy S. 2006: 'Understanding Pluralism in Climate Modeling'. *Foundations of Science*, 11, pp. 349–68.

——2008: 'Computer simulation through an error-statistical lens'. *Synthese*, 163, pp. 371–84.

Petersen, Arthur C. 2000: 'Philosophy of Climate Science'. *Bulletin of the American Meteorological Society*, 81(2), pp. 265–71.

——2006: *Simulating Nature: A Philosophical Study of Computer-simulation Uncertainties and their Role in Climate Science and Policy Advice*. Amsterdam: Het Spinhuis.

Philander, S. George 1998: *Is the Temperature Rising? The Uncertain Science of Global Warming*. Princeton, NJ: Princeton University Press.

Randall, David A., and Bruce A. Wielicki 1997: 'Measurements, Models, and Hypotheses in the Atmospheric Sciences'. *Bulletin of the American Meteorological Society*, 78(3), pp. 399–406.

——Richard A. Wood et al. 2007: 'Climate Models and Their Evaluation'. In Susan Solomon, Dahe Qin, Martin Manning et al. (eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 589–662. New York: Cambridge University Press.

Rohrlich, Fritz 1990: 'Computer Simulation in the Physical Sciences'. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 2, pp. 507–18.

Rykiel, Eric J., Jr. 1996: 'Testing ecological models: the meaning of validation'. *Ecological Modelling*, 90, pp. 229–44.

Santer, Ben D., Robert Sausen et al. 2003: 'Behavior of tropopause height and atmospheric temperature in models, reanalyses, and observations: Decadal changes'. *Journal of Geophysical Research*, 108(D1), p. 4002.

Shackley, Simon 2001: 'Epistemic Lifestyles in Climate Change Modeling'. In Clark A. Miller and Paul N. Edwards (eds.), *Changing the Atmosphere: Expert Knowledge and Environmental Governance*, pp. 107–33. Cambridge, MA: MIT Press.

Staley, Kent 2004: 'Robust Evidence and Secure Evidence Claims'. *Philosophy of Science*, 71, pp. 467–88.

van Fraassen, Bas C. 2004: 'Science as Representation: Flouting the Criteria'. *Philosophy of Science*, 71, pp. 794–804.

Washington, Warren M., and Claire L. Parkinson 2005: *Introduction to Three-dimensional Climate Modeling*. New York: University Science Books.

Weisberg, Michael 2006: 'Robustness Analysis'. *Philosophy of Science*, 73, pp. 730–42.

——and Ken Reisman forthcoming: 'The Robust Volterra Principle'. *Philosophy of Science*.

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x

Winsberg, Eric 1999*a*: 'Sanctioning Models: The Epistemology of Simulation'. *Science in Context*, 12(2), pp. 275–92.
——1999*b*: 'The Hierarchy of Models in Simulation'. In Lorenzo Magnani, Nancy J. Nersessian and Paul Thagard (eds.), *Model-Based Reasoning in Scientific Discovery*, pp. 245–69. New York: Kluwer Academic/Plenum Publishers.
——2003: 'Simulated Experiments: Methodology for a Virtual World'. *Philosophy of Science*, 70, pp. 105–25.
Woodward, Jim 2006: 'Some varieties of robustness'. *Journal of Economic Methodology*, 13(2), pp. 219–40.

Proceedings of the Aristotelian Society Supplementary Volume LXXXIII
doi: 10.1111/j.1467-8349.2009.00179.x