Assessing climate model projections: state of the art and philosophical reflections

2012, *Studies in History and Philosophy of Modern Physics*, 43(4), pp. 258-276.
https://www.sciencedirect.com/science/article/pii/S1355219812000536

Joel Katzav
The Department of Philosophy and Ethics, Eindhoven University of Technology, the Netherlands

Henk A. Dijkstra
Institute for Marine and Atmospheric Research, Utrecht University, the Netherlands

A. T. J. (Jos) de Laat
The Royal Netherlands Meteorological Institute, the Netherlands

Abstract

The present paper draws on climate science and the philosophy of science in order to evaluate climate-model-based approaches to assessing climate projections. We analyze the difficulties that arise in such assessment and outline criteria of adequacy for approaches to it. In addition, we offer a critical overview of the approaches used in the IPCC working group one fourth report, including the confidence building, Bayesian and likelihood approaches. Finally, we consider approaches that do not feature in the IPCC reports, including three approaches drawn from the philosophy of science. We find that all available approaches face substantial challenges, with IPCC approaches having as a primary source of difficulty their goal of providing probabilistic assessments.

1. Introduction

The climate system is the system of processes that underlie the behavior of atmospheric, oceanic and cryospheric phenomena such as atmospheric temperature, precipitation, sea-ice extent and ocean salinity. Climate models are designed to simulate the seasonal and longer term behavior of the climate system. They are mathematical, computer implemented representations that comprise two kinds of elements. They comprise basic physical theory – e.g., conservation principles such as conservation of momentum and heat – that is used explicitly to describe the evolution of some physical quantities – e.g., temperature, wind velocity and properties of water vapor. Climate models also comprise parameterizations. Parameterizations are substitutes for explicit representations of physical processes, substitutes that are used where lack of knowledge and/or limitations in computational resources make explicit representation impossible. Individual cloud formation, for example, typically occurs on a scale that is much smaller than global climate model (GCM) resolution and thus cannot be explicitly resolved. Instead, parameterizations capturing assumed relationships between model grid-average quantities and cloud properties are used.

The basic theory of a climate model can be formulated using equations for the time derivatives of the model's state vector variables, x_i, $i = 1, ..., n$, as is schematically represented by

$$\frac{\partial x_i}{\partial t} = F_i(x_1 ... x_n, y_1, ..., y_n, t) + G_i(t) \quad (1)$$

In Eqt. (1), t denotes time, the functions G_i represent external forcing factors and how these function together to change the state vector quantities, and the F_i represent the many physical, chemical and biological factors in the climate system and how these function together to change the state vector quantities. External forcing factors – e.g., greenhouse gas concentrations, solar irradiance strength, anthropogenic aerosol concentrations and volcanic aerosol optical depth – are factors that might affect the climate system but that are, or are treated as being, external to this system.

The x_i represent those quantities the evolution of which is explicitly described by basic theory, that is the evolution of which is captured by partial time derivatives. The y_i represent quantities that are not explicitly described by basic theory. So these variables must be treated as functions of the x_i, i.e., the y_i must be parameterized. In this case, the parameterizations are schematically represented in Eqt. (2).

$$y_i = H_i(x_1,...,x_n) \tag{2}$$

Given initial conditions $x_i(t_0)$ at time $t = t_0$ and boundary conditions, the climate model calculates values of the state vector at a later time $t = t_1$ in accordance with Eqt. (1).

Climate models play an essential role in identifying the causes of climate change and in generating projections. Projections are conditional predictions of climatic quantities. Each projection tells us how one or more such quantities would evolve were external forcing to be at certain levels in the future. Some approaches to assessing projections derive projections, and assess their quality, at least partly independently of climate models. They might, for example, use observations to decide how to extend simulations of present climate into the future (Stott et al., 2006) or derive projections from, and assess them on the basis of, observations (Bentley, 2010; Siddall et al., 2010). We focus on climate-model-based assessment. Such assessment is of the projections of one or more climate models and is assessment in which how good models are in some respect or another is used to determine projection quality. A climate model projection (CMP) quality is a qualitative or quantitative measure, such as a probability, that is indicative of what we should suppose about CMP accuracy.

It is well recognized within the climate science community that climate-model-based assessment of projection quality needs to take into account the effects of climate model limitations on projection accuracy (Randall et al., 2007; Smith, 2006; Stainforth et al., 2007a). Following Smith (2006) and Stainforth (2007a), we distinguish between the following main types of climate model limitations:

(a) External forcing inaccuracy – inaccuracy in a model's representation of external forcing, that is in the G_i in Eqt. (1).

(b) Initial condition inaccuracy – inaccuracy in the data used to initialize climate model simulations, that is in the $x_i(t_0)$.

(c) Model imperfection – limitations in a model's representation of the climate system or in our knowledge of how to construct this representation, including:

1. Model parameterization limitations – limitations in our knowledge of what the optimal or the appropriate parameter values and parameterization schemes for a model are. This amounts, in the special case where parameterizations are captured by Eqt. (2), to limitations in our knowledge of which functions H_i one should include from among available alternatives.

2. Structural inadequacy – inaccuracy in how a model represents the climate system which cannot be compensated for by resetting model parameters or replacing model parameterizations with other available parameterization schemes. Structural inaccuracy in Eqt. (1) is manifested in an insufficient number of variables x_i and y_i as well as in the need for new functions of these variables.

Parameterization limitations are illustrated by the enduring uncertainty about climate sensitivity and associated model parameters and parameterization schemes. A relatively recent review of climate sensitivity estimates underscores the limited ability to determine its upper bound as well as the persistent difficulty in narrowing its likely range beyond 2 to 4.5 °C (Knutti and Hegerl, 2008). The 21 GCMs used by Working Group One of the IPCC fourth report (WG1 AR4) illustrate structural inadequacy. These sophisticated models are the models of the World Climate Research Programme's Coupled Model Intercomparison Project phase 3 (CMIP3) (Meehl et al., 2007a). Some important sub-grid and larger than grid phenomena that are relevant to the evolution of the climate system are not accurately represented by these models, some are only represented by a few of the models and some are not represented at all. Parameterization of cloud formation, for example, is such that even the best available parameterizations suffer from substantial limitations (Randall et al., 2003). None of the models represent the carbon cycle, only some represent the indirect aerosol effect and only two represent stratospheric chemistry (CMIP3, 2007). The models also omit many of the important effects of land use change (Mahmood et al., 2010; Pielke, 2005). Many of their limitations, e.g., the limited ability to represent surface heat fluxes as well as sea ice distribution and seasonal changes, are the result of a combination of structural inadequacy and parameterization limitations (Randall et al., 2007, p. 616). CMIP3 simulations illustrate initial condition inaccuracy. Due to constraints of computational power and to limited observations, these simulations start from selected points of control integrations rather than from actual observations of historical climate (Hurrell et al., 2009).

The most ambitious assessments of projection quality, and these are primarily climate-model-based assessments, are those of WG1. The first three WG1 reports rely primarily on the climate-model-based approach that we will call the confidence building approach. This is an informal approach that aims to establish confidence in models, and thereby in their projections, by appealing to models' physical basis and success at representing observed and past climate. In the first two reports, however, no uniform view about what confidence in models teaches about CMP quality is adopted (IPCC 1990; IPCC 1996). The summary for policymakers in the WG1 contribution to the IPCC first assessment report, for example, qualifies projections using diverse phrases such as 'we predict that', 'confidence is low that' and 'it is likely that’ (IPCC 1990). A more systematic view is found in WG1's contribution to the third IPCC assessment report (WG1 TAR). It made use of a guidance note to authors which recommends that main results be qualified by degrees of confidence that are calibrated to probability ranges (Moss and Schneider, 2000). The summary for policymakers provided by WG1 TAR does assign projections such degrees of confidence. It expresses degrees of confidence as degrees of likelihood and takes, e.g., 'very likely' to mean having a chance between 90 and 99 %, and 'likely' to mean having a chance between 66 % and 90 %. The chapter on projections of future climate change, however, defines degrees of confidence in terms of agreement between models. A very likely projection, for example, is defined (roughly) as one that is physically plausible and is agreed upon by all models used (IPCC 2001).

WG1 AR4’s assessment of projection quality has two stages. First, confidence in models is established as in previous reports. This is mostly achieved in Chapter 8 – which describes, among other things, successful simulations of natural variability (Randall et al., 2007) – and in chapter 9 – which focuses on identifying the causes of climate change, but also characterizes model successes at simulating 20th century climate change (Hegerl et al., 2007). The second stage is carried out in Chapter 10 – which provides WG1 AR4’s global projections (Meehl et al., 2007b) – and Chapter 11 – which focuses on regional projections (Christensen et al., 2007). In these chapters, expert judgment is used to assign qualities to projections given established confidence in models and the results of formal, probabilistic projection assessment (Meehl et al., 2007b). WG1 AR4 is the first WG1 report that makes extensive use of formal assessment, though it recognizes that such approaches are in their infancy (Christensen et al., 2007; Randall et al., 2007). Both climate-model-based and partly climate-model-independent formal approaches are used.

Although WG1 AR4 assesses models using degrees of confidence, it does not assess projections in these terms. Nor does it equate projection likelihoods with degrees of agreement among models. It does, however, implement the advice to provide probabilistically calibrated likelihoods of projections (IPCC 2005). For example, unlike WG1 TAR, WG1 AR4 provides explicit likelihood estimates for projected ranges of global mean surface temperature (GMST) changes. It estimates that the increase in GMST by the end of the century is likely to fall within -40 to +60 % of the average GCM warming simulated for each emission scenario and provides broader uncertainty margins than the GCM ensemble in particular because GCMs do not capture uncertainty in the carbon cycle (Fig. 2).

The sophistication of WG1 AR4's assessments was enabled by the increasing ability to use multi-GCM and perturbed physics GCM ensembles. Thus, while WG1's first two reports relied on simple models to produce long term GMST projections, WG1 TAR and WG1 AR4 relied primarily on state-of-the-art GCM ensembles to assess these and other projections. WG1 AR4 nevertheless still relied on simpler models, including intermediate complexity and energy balance models (Randall et al., 2007).

In this review, we provide a critical discussion of the (climate-model-based) approaches to assessing projection quality relied on in WG1 AR4 and more recent work by climate scientists. In doing so, we build on the substantial climate science literature, including WG1 AR4 itself. We, however, extend this literature using the perspective of the philosophy of science. Our discussion does focus more than climate scientists themselves tend to on precisely why assessing projection quality is difficult, on what is required of an adequate approach to such assessment and on the limitations of existing approaches. We, nevertheless, also address some of the practical concerns of climate scientists. We outline three views of how to assess scientific claims that are drawn from the philosophy of science and consider how they might further assist in assessing projection quality. Important issues that space does not allow us to address are the special difficulties that assessment of regional projection quality raises. An issue that deserves more attention than we have given it is that of how uncertainty about data complicates assessing projection quality.

We begin (Section 2) by considering what kinds of qualities should be assigned to projections, especially whether probabilistic qualities should be assigned. We then (Section 3) discuss why assessing projection quality is difficult and outline criteria for adequate approaches to doing so. Using these criteria, we proceed to discuss (Sections 4–7) the approaches that were used in WG1 AR4, namely the confidence building, the subjective Bayesian and the likelihood approaches. Finally (Section 8), we discuss approaches that are not used, or are not prominent in, WG1 AR4, including the possibilist and three philosophy-of-science-based approaches.

2. Probabilistic and non-probabilistic assessment

Probabilistic assessment of projection quality will here be taken to include assigning probabilities or informative probability ranges to projections or projection ranges. Such assessment has been argued for on the ground that it is better suited to handling the inevitable uncertainty about projections than deterministic assessments are (Raisanen and Palmer, 2001). But philosophers of science, computer scientists and others point out that probabilities fail to represent uncertainty when ignorance is deep enough (Halpern, 2003; Norton, 2011). Assigning a probability to a prediction involves, given standard probability frameworks, specifying the space of possible outcomes as well as the chances that the predicted outcomes will obtain. These, however, are things we may well be uncertain about given sufficient ignorance. For example, we might be trying to assess the probability that a die will land on '6' when our information about the kind and bias of the die is limited. We might have the information that it can exhibit the numerals '1', '6' and '8' as well as the symbol '*', but not have any information about what other symbols might be exhibited or, beyond the information that '6' has a greater chance of occurring than the other known symbols, the chances of symbols being exhibited. The die need not be a six sided die. In such circumstances, it appears that assigning a probability to the outcome '6' will misrepresent our uncertainty.

Assigning probability ranges and probabilities to ranges can face the same difficulties as assigning probabilities to single predictions. In the above example, uncertainty about the space of possibilities is such that it would be inappropriate to assign the outcome '6' a range that is more informative than the unhelpful 'somewhere between 0 and 1'. The same is true about assigning the range of outcomes '1', '6' and '8' a probability.

One might suggest that, at least when the possible states of a system are known, we should apply the principle of indifference. According to this principle, where knowledge does not suffice to decide between possibilities in an outcome space, they should be assigned equal probabilities. Some work in climate science acknowledges that this principle is problematic, but suggests that it can be applied with suitable caution (Frame et al., 2005). Most philosophers argue that the principle should be rejected (Strevens, 2006a). We cannot know that the principle of indifference will yield reliable predictions when properly applied (North, 2010). If, for example, we aim to represent complete ignorance of what value climate sensitivity has within the range 2 to 4.5 °C, it is natural to assign equal probabilities to values in this range. Yet whether doing so is reliable across scenarios in which greenhouse gasses double depends on what climate sensitivity actually tends to be across such scenarios and it is knowledge of this tendency that is, given the assumed ignorance, lacking. Further, we can only define a probability distribution given a description of an outcome space and there is no non-arbitrary way of describing such a space under ignorance (Norton, 2008; Strevens, 2006a). What probability should we assign to climate sensitivity's being between 2 and 4 °C, given complete ignorance within the range 2 to 6 °C? 50 % is the answer, when the outcome space is taken to be the given climate sensitivity range and outcomes are treated as equiprobable. But other answers are correct if alternative outcome spaces are selected, say if the outcome space is taken to be a function not just of climate sensitivity but also of feedbacks upon which climate sensitivity depends. And in the supposed state of ignorance about climate sensitivity, we will not have a principled way of selecting a single outcome space.

Although the case of the die is artificial, our knowledge in it does share some features with our knowledge of the climate system. We are, for example, uncertain about what possible states the climate system might exhibit, as already stated in the case of climate sensitivity. A central question in what follows is to what extent our ignorance of the climate system is such that probabilistic assessment of projection quality is inappropriate.

Acknowledging that probabilistic assessment is inappropriate in some case is by no means then to give up on assessment. Assigning non-probabilistic qualities can commit us to less than assigning probabilities or probability ranges and thus can better represent uncertainty. Judging that it is a real possibility that climate sensitivity is 2 °C does not require taking a position on the full range of climate sensitivity. Nor need rankings of climate sensitivities according to plausibility do so. Other non-probabilistic qualities the assignment of which is less demanding than that of probabilities or probability ranges are sets of probability ranges and the degree to which claims have withstood severe tests (see Halpern (2003) for a discussion, and formal treatment, of a variety of non-probabilistic qualities. We discuss severe-test-based and real-possibility-based assessments in sections 8.4 and 8.1 respectively).

3. Why is assessing projection quality difficult?

Projections, recall, are predictions that are conditional on assumptions about external forcing. So errors in assumptions about external forcing are not relevant to assessing projection quality. Such assessment need only take into account the effects of initial condition inaccuracy and model imperfection. In the present section, we consider why these kinds of limitations make assessing projection quality difficult. This question is not answered just by noting that climate models have limitations. Scientific models are in general limited, but it is not generally true that assessing their predictions is a serious problem. Consider standard Newtonian models of the Earth-Sun system. Such models suffer from structural inadequacy. They represent the Earth and the Sun as point masses. Moreover, they tell us that the Earth and the Sun exert gravitational forces on each other, something that general relativity assures us is not strictly true. Still, assessing to what extent we can trust the predictions these models are used to generate is something we typically know how to do.

3.1 Initial condition inaccuracy and its impact on assessing projections

We begin by considering the difficulties associated with initial condition error. Work in climate science emphasizes the highly nonlinear nature of the climate system (Le Treut et al., 2007; Rial et al., 2004), a nature that is reflected in the typically nonlinear form of the F_i in Eqt. (1). Nonlinear systems are systems in which slight changes to initial conditions can give rise to non-proportional changes of quantities over time (Lorenz, 1963). This high sensitivity can make accurate prediction inherently difficult. Any errors in simulations of highly nonlinear systems, including even minor errors in initial condition settings, might be multiplied over time quickly. The high sensitivity to initial conditions also, as climate scientists note, threatens to make assessing prediction quality difficult. The way in which error grows over time in such systems cannot be assumed to be linear and might depend on how the system itself develops (Palmer, 2000; Palmer et al., 2005).

However, how serious a problem sensitivity to initial conditions is for assessing projection quality is not a straightforward matter. The known inaccuracy in model initial condition settings means that high sensitivity of the evolution of climatic quantities to initial conditions might be important. Yet, a climatic quantity the evolution of which is going to be highly nonlinear at one temporal scale may continue to exhibit approximately linear evolution on another such scale. Greenland ice volume may, for example, evolve linearly in time over the coming few decades but nonlinearly over more than three centuries (Lenton et al., 2008). If this is so, nonlinearity will only be a limited obstacle to assessing projections of Greenland ice volume. More generally, whether, and to what extent, a climatic process is nonlinear will depend on the desired projection accuracy, the quantity of interest, the actual period and region of interest and the temporal and spatial scale of interest (IPCC 2001). Thus, whether the highly nonlinear behavior of the climate system is a problem for assessing projection quality will have to be determined on a case by case basis.

3.2 Tuning and its impact on assessing projections

Further features of climate modeling complicate determining the impact of model imperfection on CMP quality. The first of these features is tuning. Tuning is the modification of parameterization scheme parameters so as to accommodate – create agreement with – old data. A prominent instance is the setting of parameters associated with the small-scale mixing processes in the ocean. Tuning to current day conditions is hard to avoid given the limited available data about the climate system. Moreover, climate scientists worry that when model success results from accommodation, it provides less confirmation of model abilities than success that results from out-of-sample prediction, that is from prediction that is made prior to the availability of the data but that nevertheless accurately captures the data (Knutti, 2008; Smith, 2006; Stainforth et al., 2007a). Prominently, there is the suspicion that accommodation threatens to guarantee success irrespective of whether models correctly capture those underlying processes within the climate system that are relevant to its long term evolution (Schwartz et al., 2007). This impacts assessing projection quality. Difficulty in assessing the extent to which a model's basic assumptions hold will give rise to difficulty in assessing its projections.

Work in the philosophy of science, however, shows that whether, and under what conditions, the accommodation of data provides reduced confirmation is an unresolved one (Barrett and Stanford, 2006). On the one hand, some philosophers do worry that accommodation raises the threat of generating empirical success irrespective of whether one’s theoretical assumptions are correct (Worrall, 2010). On the other hand, if we prioritize out-of-sample prediction over accommodation, evidence might be good evidence of the suitability of model *A* for generating a set of projections *R* for the late 21st century and not so good evidence for the suitability of model *B* for this purpose even though the models are intrinsically identical. This might occur because the developers of model *B* happen to learn, while those of *A* do not learn, of relevant evidence at the stage of model development. In such circumstances, the developers of *B* might end up accommodating the evidence while the developers of *A* successfully predict it. Resulting differing degrees of confidence in the models would, paradoxically, have to be maintained even if it were recognized that the models are intrinsically identical. If accommodated evidence as such is poor evidence, what determines whether evidence is good evidence for a model is the model's history and not just its intrinsic characteristics (see, e.g., Hudson (2007) for worries about the value of out-of-sample prediction).

Unfortunately, while the philosophy of science literature tells us that tuning might not be so bad, it still leaves open the possibility that it is problematic. So how tuning affects CMP accuracy still needs to be addressed.

Of course, different approaches to parameterization affect CMP quality differently. For example, stochastic parameterizations, i.e., parameterizations that introduce small but random variations in certain model parameters or variables, are arguably sometimes better than standard deterministic parameterizations (Palmer et al., 2005). The worries about tuning, however, arise for all available parameterization techniques.

3.3 The long term nature of projections and its impact on assessing projections

A second factor that, according to some climate scientists, complicates determining the impact of model imperfection is the fact that climate models cannot be tested repeatedly across relevant temporal domains (Frame et al., 2007; Knutti, 2008). We can repeatedly compare weather model forecasts with observations. Success frequencies can then be used to provide probabilistic estimates of model fitness for the purpose of generating accurate forecasts. Recently, some old CMPs have been directly assessed (Hargreaves, 2010). But many CMPs have fulfillment conditions that are never realized and, anyway, CMPs are generally too long term to allow repeated direct testing. Thus, it has been argued, it is hard to take the impact of many model implemented assumptions about long term climate into account in assessing model suitability for generating projections.

But the fact that we cannot test our models' predictions over the time scales of the predictions is not itself a difficulty. Consider predictions of Earth orbit variation induced changes in solar radiation at the top of atmosphere over the next million years. Here, predictions are generated using model implemented theory about orbital physics, including Newtonian mechanics and an understanding of its limitations (Laskar et al., 2004). This theory is what grounds confidence in the predictions, though the theory and the models based upon it are only tested against relatively short-term data. As the general views we will discuss about how scientific claims are assessed illustrate, there is no need to assume that estimates of a model's ability must be, or are, made on the basis of numerous observations of how well the model has done in the past.

3.4 Basic theory, recognized model imperfection and assessing projections

There are nevertheless two more factors other than tuning that complicate taking into account the effects of model imperfection in assessing projection quality. The first, which is not explicitly discussed in the climate science literature but which climate scientists no doubt recognize, is the combination of known model imperfection with the fact that the background knowledge used in constructing models provides a limited constraint on model construction.

Philosophers of science observe that theory provides essential information about model reliability (Humphreys, 2004). Newtonian physics, general relativity and other theories provide essential information about when, and to what extent, we can neglect aspects of the solar system in applying Newtonian theory to model the orbit of the Earth. The same, we have noted, is true of models of how changes in the Earth's orbit affect top of the atmosphere solar radiation. In the case of climate modeling, however, the extent to which theory can guide climate model construction and projection quality assessment is limited. After all, parameterization is introduced precisely because of a limited ability to apply explicit theory in model construction.

We do not, for example, have a quantitative theory of the main mechanisms of the stratospheric circulation. As a result, while our partial understanding of these mechanisms can be used in arguing that CMIP3 GCMs’ limited ability to represent the stratosphere adversely affects their simulations of tropospheric climate change, the way and extent to which it does so will remain a matter of ongoing investigation (as in, e.g., Dall' Amico (2010)).

A limited ability to apply theory in model construction will even make it difficult to decide what we can learn about CMP accuracy from whatever success models have. For easy, relatively theory neutral, ways of drawing conclusions from model successes are hard to come by given model imperfection.

Model imperfection implies that models will only have limited empirical success, as indeed is found in the case of climate models. The strongest claim reported by WG1 AR4 on behalf of simulated GCM multi-model annual mean surface temperatures is that, outside of data poor regions such as the polar regions, simulated temperatures were usually within 2 °C of observed temperatures. For most latitudes, the error in simulated zonally averaged outgoing shortwave radiation was about 6%.

Simulation of the strength of the Atlantic Meridional Overturning Circulation (MOC) suffers from substantial inaccuracies (Fig. 3). And the same is true of simulation of precipitation patterns, especially on regional scales (Randall et al., 2007). Such inaccuracies short-circuit a simple argument for assigning a high quality to CMPs, namely one that assigns them such a quality on the ground that they were generated by models which simulate data well across the board. Indeed, there is reason to think that increased ability to simulate the current mean climate state across large sets of climate variables is a limited constraint on CMP accuracy (Abe et al., 2009; Knutti et al., 2010). For example, it has been shown (Knutti et al., 2010) that the range of CMPs of precipitation trends is not substantially affected by whether it is produced by all the CMIP3 models or by a subset of high performing models. Assessment of a projection's quality requires correctly identifying which, if any, aspects of model performance are relevant to the projection's accuracy.

Further difficulty in figuring out what to infer from what model success there is arises from the well recognized interdependency of climatic processes. Changes in some climatic processes inevitably give rise to changes in others. Changes in cloud cover, land usage, soil hydrology, boundary layer structure and aerosols will, for example, affect surface temperature trends and vice versa. Thus, an accurate simulation of some quantity x will require an appropriate simulation of related quantities upon which x depends. And our assessment of the quality of a projection of x will have to take into account both the accuracy with which x has been simulated and the accuracy with which related quantities have been simulated. One cannot simply argue that since some models simulate a certain climatic quantity well, their projections of this quantity are good (Parker, 2009).

Easy, relatively theory neutral ways of assessing what to infer from limited model successes might also be hampered by structural instability, which is, like high sensitivity to changes in initial conditions, a feature of nonlinear systems. A system is structurally unstable when slight changes to its underlying dynamics would give rise to qualitatively different system evolutions. Components of the climate system do exhibit structural instability (Ghil et al., 2008; McWilliams, 2007). This means that minor observed errors in simulating current climate might, given model imperfection, lead to substantial errors in CMPs.

3.5 Unrecognized model imperfection and assessing projections

The final source of difficulty for assessing projection quality in light of model imperfection is the possibility, worried about by scientists from all fields, that our models are wrong in unrecognized ways. Empirically successful theories and models have often turned out to rest on mistaken assumptions about which theoretical – that is not directly observable – processes and entities explain observable phenomena (Laudan, 1981). This is true of theories and models of the climate system. Prior to the 1990s, for example, climate models that were used to provide spatial simulations of global surface temperatures did not include a representation of the role of aerosols in the climate system and this turned out to be a surprisingly substantial incompleteness in the simulations (Wigley, 1994). Moreover, current candidates for substantially underestimated forcing, feedbacks and internal variability exist (e.g., terrestrial biogeochemical feedbacks (Arneth et al., 2010) and feedbacks amplifying the effects of solar luminosity (Kirkby, 2007)).

Some philosophers have concluded, largely on the basis of the history of successful but superseded theories and models, that a theory or model's predictive success should not be used to justify belief in what the theory or model tells us about theoretical entities and processes (see, e.g., Stanford (2006)). On their view, theories and models should be taken to be no more than tools for predicting observable phenomena. The sad truth, however, is that it is currently unclear what we are entitled to assume about how complete empirically successful theories and models are (see Saatsi (2005) and Psillos (1999) for two of many further alternative perspectives on this unresolved issue). In particular, it is unclear what we are entitled to assume about how complete climate models and our knowledge of the climate system are, including about how complete our knowledge of climatic factors that are materially relevant to CMP accuracy is. This complicates assessment. For example, difficulty in estimating the completeness of GCMs' representations of the effects of solar luminosity fluctuations means difficulty in assessing projections of GMST trends.

3.6 Criteria of adequacy for approaches to assessing projections

Our discussion of why assessing projection quality is difficult helps to spell out criteria of adequacy for approaches to such assessment. Adequate approaches will, given initial condition inaccuracy, have to assess projection quality in light of the possible path dependent nature of error propagation. Given the inevitable use of parameterization, they will have to take the possible effects of tuning into account. They will also have to take the impact of model imperfection into account. Doing so involves paying attention to climate models’ limited ability to simulate climate, to the difficulty in determining which aspects of model empirical success are relevant to assessing which projections, to the interdependence of the evolution of climatic quantities along with the effect of this interdependence on error propagation and to possible structural instability. Doing so also requires attending to the history induced lack of clarity about unrecognized model imperfection. If the claim is that we are entitled to ignore the history of successful but superseded models and thus to cease worrying about unrecognized model imperfection, we need to be told why. Otherwise, the impact of unrecognized climate model limitations on the accuracy of their projections needs to be taken into account.

Since we know that only some of the projections of climate models will be accurate, an adequate approach to assessing projection quality will have to provide projection (or class of projections) specific assessments (Gleckler et al., 2008; Parker, 2009). It should judge the quality of a CMP on the basis of how fit the model or models which generated it are for the purpose of doing so, i.e., for the purpose of correctly answering the question the CMP answers.

4. The confidence building approach

We now discuss the confidence building approach to assessing projection quality. This approach, recall, focuses on what model agreement with physical theory as well as model simulation accuracy confirm. Better grounding in physical theory and increased accuracy in simulation of observed and past climate is used to increase confidence in models and hence in CMPs. Given the emphasis on grounding in physical theory, the reliance here is primarily on GCMs.

In the uncertainty assessment guidance note for WG1 AR4 lead authors, degrees of confidence in models are interpreted probabilistically. Specifically, they are calibrated to chance ranges, e.g., very high confidence in a model is interpreted as its having an at least 9 in 10 chance of being correct (IPCC 2005). The chance that a model is correct can be thought of as the model's propensity to yield correct results with a certain frequency, but neither the guidance note nor the report itself indicate how chances should be interpreted. Indeed, they do not indicate how the talk of chances of models' being correct relates to the talk of CMP likelihoods, and the report does not go beyond establishing increased confidence in models in order to assign them specific degrees of confidence. This last fact makes it unclear how the report's use of 'increased confidence' relates to the explication of degrees of confidence in terms of chances. Better grounding in physical theory is illustrated by the, at least partly theoretically motivated, inclusion in some GCMs of interactive aerosol modules (Randall et al., 2007)**.** Illustrations of improved simulation accuracy are given below.

4.1 Initial condition inaccuracy and the confidence building approach

WG1 AR4 states that many climatic quantities of interest, including those relating to anthropogenic climate change, are much less prone to nonlinear sensitivity to initial conditions than weather related quantities and are thus more amenable to prediction (Le Treut et al., 2007). This relative insensitivity to initial conditions is argued for primarily on the basis of GCM simulations in which initial conditions are varied. Notably, CMIP3 multi-model simulations of 20th century GMST, in which ranges reflect different initial condition runs of participating models, suggest little internal variability in GMST over periods of decades and almost none over the whole century (See Fig. 1 and (Hawkins and Sutton, 2009)).

WG1 AR4 acknowledges that confidence in simulations of response to changes in initial conditions depends on resolving worries about the effects of relevant model imperfection (Meehl et al., 2007b). But the claim is that these worries can be mitigated by examining how well GCMs simulate important sources of the climate system's nonlinear responses, e.g., the El Niño – Southern Oscillation (ENSO) and the MOC. Thus, the ability of GCMs to simulate observed nonlinear change in the

Atlantic MOC in response to fresh water influx has been used to argue that they can produce reliable projections of aspects of 21st century MOC behavior but that confidence in projections beyond the 21st century is very limited (Pitman and Stouffer, 2006).

Computational resources, however, only allowed a very limited range of initial conditions to be explored by CMIP3 GCMs (CMIP3, 2007)**.** As to the question of the extent to which GCM ability to simulate (in)sensitivity to initial conditions does help with assessment in light of model imperfection and tuning, it is addressed in the following sections. Here we only note that the need to address this question has been made pressing since WG1 AR4. Recent work suggests that GCMs do not adequately capture the structure of the climate system prior to abrupt changes in the past and are, in some circumstances, insufficiently sensitive to initial conditions. They can, for example, only simulate the cessation of the MOC under about 10 times of the best estimate of actual fresh water influx that has brought it about in the past (Valdes, 2011). There is, in addition, a spate of studies according to which CMIP3 GCMs substantially underestimate the extent to which 20th century GMST anomalies are due to internal variability, including initial condition variability, on multidecadal scales (Semenov et al., 2010; Swanson et al., 2009; Wu et al., 2011). Some work suggests that the underestimates extend to periods of 50 to 80 years in length (Wyatt et al., 2011).

Recognizing the potential significance of initial conditions to improving multidecadal CMPs, some recent work aims to take on the challenge of limited available data in order to initialize simulation runs to actual observed initial conditions (Hurrell et al., 2009). More extensive exploration of the impact of varying

GCM simulation initial condition settings is also being carried out (Branstator and Teng, 2010).

4.2 Parameterization, tuning and the confidence building approach

WG1 AR4 addresses the difficulty of assessing projection quality in light of tuning by taking increased simulation accuracy to increase confidence in models only when this accuracy is not a result of direct tuning, i.e., only when it is not the result of tuning a parameter for a certain quantity to observations of that quantity (Randall et al., 2007, p. 596). But tuning can be indirect. GCMs do not possess parameters for GMST trends, and thus cannot be directly tuned to observations of these trends. Nevertheless, there is (CCSP, 2009) substantial uncertainty about radiative forcings, and especially about aerosol forcing, allowing forcing parameters to be tuned to yield close agreement between simulated and observed 20th century mean GMST trends (Fig. 1). That this tuning occurs is, as is widely recognized within the climate science community, suggested by the observation that different models achieve such agreement by substantially different combinations of estimates of climate sensitivity and radiative forcing [CCSP, 2009; Knutti, 2008b].

The difficulty in assessing projection quality in light of parameterization limitations is partly, if implicitly, addressed by noting improvements in parameterization schemes since the publication of WG1 TAR. As schemes that incorporate a better understanding of the climate system and show better agreement with data become available, we acquire a better understanding of the limitations of older schemes and increase trust in model performance. Such improvement, however, leaves open the question of how to handle worries about tuning. Moreover, increased quality of parameterizations does not indicate how to assess the impact of the inevitable remaining underdetermination in parameterization choice on projection quality. Thus, it remains unclear how accurate CMPs actually are.

Another strategy that is not explicitly discussed in WG1 AR4, but which is consistent with the confidence building approach, is suggested by the idea that grounding in basic theory increases confidence in models. Perhaps, in some cases, the role of basic theory in generating CMPs is sufficient so as to eliminate, or substantially reduce, worries arising from the use of parameterizations. It has been argued that while simulating the feedback effect of increased water vapor inevitably makes use of parameterizations, this effect is dominated by processes that are represented by the equations of fluid dynamics and thus will continue to be accurately simulated by climate models (Dessler and Sherwood, 2009). It has also been suggested that, since GCMs use the equations of fluid dynamics, our ability to predict nonlinear MOC evolution that results from its fundamental properties is beginning to mature, unlike our ability to predict nonlinear evolution it might exhibit as a result of terrestrial ecosystems (Pitman and Stouffer, 2006).

One difficulty here is how to determine that properties represented by basic physical theory largely determine the evolution of projected quantities. Insofar as estimates that this is so rely on – as, e.g., Dessler and Sherwood (2009) rely on – climate model results, it is assumed that available parameterizations are adequate and the reliance on parameterization is not bypassed. Further, even if we have managed to isolate properties that are represented by basic theory and determine the evolution of a projected quantity, we cannot escape worries relating to the use of parameterization. Parameterization always plays an essential role even in descriptions of subsystems of the climate for which we possess basic equations. Basic equation discretization in GCMs brings with it grid-scale dependent parameterization, e.g., grid-scale dependent convection parameterization, of subgrid processes. How this discretization and associated parameterization affects CMP accuracy, especially in light of how it affects model ability to simulate highly nonlinear dynamics, needs adequate treatment.

4.3 Structural inadequacy and the confidence building approach

Increased model grounding in basic physical theory and increased accuracy in simulation results across a range of such results does indicate increased structural adequacy. Moreover, confidence building exercises do typically acknowledge a wide variety of model limitations. What we need, however, are arguments connecting increased success with the quality of specific classes of CMPs. This includes arguments addressing the issue of how total remaining inadequacy affects CMP quality.

Thus, for example, WG1 AR4 offers information such as that more state-of-the-art models no longer use flux adjustments, that resolution in the best models is improving, that more physical processes are now represented in models and that more such processes are explicitly represented (Randall et al., 2007). But we need arguments that connect these successes to an overall estimate of remaining structural inadequacy and tell us what this inadequacy means for the quality of specific classes of CMPs. It is one thing to be shown that simulated multi-model mean surface temperatures are, outside of data poor regions, usually within 2 °C of observed temperatures, another to be shown how this information bears on the quality of CMPs of mean surface temperature trends and yet another to be shown how it bears on the quality CMPs of mean precipitation trends.

While the needed arguments can be further developed, it remains to be seen how far they can be developed. Further, it is likely that these arguments will, to a substantial extent, be based on theory and expert judgment, thus limiting the extent to which the confidence building approach is model based.

4.4 The appeal to paleoclimate

An important distinction needs to be made between model ability to simulate 20th century climate and model ability to simulate paleoclimate. The latter provides opportunities for out-of-sample testing, as WG1 AR4 notes (Jansen et al., 2007, p. 440). Such testing is of particular significance as it has the potential to help in addressing the question of the extent to which tuning to current climate is a problem. Indeed, there is growing recognition of the importance of palaeodata, including of its importance for model assessment (Caseldine et al., 2010). In this context, there is an ongoing debate about whether to conclude that GCMs lack representations of crucial mechanisms/feedbacks because these models have difficulties in accurately simulating past warm, equable climates with a weak equator-to-pole temperature gradient (Huber and Caballero, 2011; Spicer et al., 2008).

Although this may change in the future, the burden of assessing models in light of data nevertheless currently rests firmly on the ability of models to simulate recent climate. This is so for at least three reasons. First, simulation experiments with paleodata are still limited. WG1 AR4's appeal to such simulations is confined primarily to two instances. WG1 AR4 uses model ability to simulate aspects of the climate system during the Last Glacial Maximum (LGM) in order further to support the claim that models have captured the primary feedbacks operating in the climate system at the time (Jansen et al., 2007, p. 452). WG1 AR4 also uses model ability to simulate climate responses to orbital forcing during the mid-Holocene in order to improve confidence in model ability to simulate responses to such forcing (Jansen et al., 2007, p. 459). Second, most of the models WG1 AR4 relies on in generating projections are not among the models it relies on in discussing paleoclimate simulations (Schmidt, 2010). And when the same models are relied on in both contexts, model resolution usually varies across the contexts (Braconnot et al., 2007). Practical constraints mean lower resolution models have to be used to simulate paleoclimate. Thus it is unclear what the paleoclimate simulation successes allow us to conclude about model fitness for the purpose of generating projections. Third, there are substantial, unresolved issues about how uncertain paleoclimate reconstructions are, and thus about what we can learn from them (Snyder, 2010; Wunsch, 2010).

4.5 Inter-model results, robust projections and the confidence building approach

The confidence building approach is strengthened, both in WG1 AR4 and elsewhere, by noting that state-of-the-art GCMs provide a robust and unambiguous picture of the evolution of some large scale features of climate. Such multi-model results are supposed to increase confidence in projections. For example, state-of-the-art GCMs predict that GMST evolution will be roughly linear over much of this century, thus supposedly reducing worries about the sensitivity of such evolution to initial condition changes and to minor variations in model structure (Knutti, 2008).

How does the appeal to multi-model results help in assessing projection quality, as opposed to improving projection accuracy? We outline two views about how it does so and then critically discuss these views.

A common assumption in formal analyses of multi-model ensemble results, and to some extent in applications of the confidence building approach, is that model errors are independent of each other and thus tend to cancel out in calculations of multi-model means (Meehl et al., 2007b; Palmer et al., 2005; Tebaldi and Knutti, 2007). Indeed, there is empirical evidence that multi-model means are more accurate than are the results of individual models (see Gleckler et al. (2008) as well as, for further references, Knutti et al. (2010)). Given the assumptions of error independence and of error cancellation, one could argue that we can expect a reduction of error in ensemble means with increased model numbers and thus can take the number of models used in generating means to be an indicator of CMP quality (Tebaldi and Knutti, 2007).

In addition, or alternatively, one can assume that ensemble models are to some extent independent of each other in that they explore alternative model structures and parameterizations that are consistent with our knowledge of the climate system (Murphy et al., 2007). Ensemble projection ranges can then be viewed as at least partial explorations of our uncertainty about the climate system and can thus be used to tell us something about projection quality. One might suggest, in particular, that the greater the extent to which the range of uncertainty is explored by an ensemble, the greater the extent to which the projections/projection ranges it produces are robust or insensitive to uncertain assumptions and thus the more probable these results are (Weisberg (2006) describes the general logic behind appeals to robustness). Multi-model ensemble projection ranges are sometimes interpreted probabilistically, e.g., the range of generated projections is supposed to span the range of possibilities and each projection is assigned a probability equal to the fraction of models that generate it (as in Räisanen and Palmer (2001) and, to some extent, in WG1 TAR (IPCC 2001)).

The appeal to multi-model results does not, and is not intended to, address the issue of tuning or the difficulty of figuring out what to infer about the quality of specific CMPs from the partial empirical successes of models. Further, worries about the use of multi-model ensembles have been raised both within and without climate science.

Philosophers have pointed out that individual model error can only cancel out to a limited extent because limited knowledge and limited computational resources mean that where one model's error is not repeated by another model, the other model will probably have to introduce a different error (Odenbaugh and Alexandrova, 2011). Limited knowledge and limited computational resources also mean that substantial model imperfection will inevitably be shared across models in ensembles (Odenbaugh and Alexandrova, 2011). Multi-model ensembles in all fields of research accordingly inevitably leave us with substantial error the impact of which on results is not estimated. So, while coming to rely on multi-model ensembles might entitle us to be more confident in projections than we would have been otherwise, it does not appear to allow us to assign qualities that, like probabilities and informative probability ranges, involve specifying the full range of possible evolutions of projected quantities.

Climate scientists’ examination of GCM ensemble results confirms that such ensembles only provide limited improvement in agreement with empirical data and that much of the remaining disagreement arises from biases that are systematic across ensemble members (Knutti et al., 2010). For present day temperature, for example, half of the bias exhibited by the ensemble of models used by CMIP3 would remain even if the ensemble were enlarged to include an indefinite number of models of similar quality (Fig. 4). The observation that models share model imperfections is also acknowledged in climate science research, including in WG1 AR4. Climate modelers tend to aim at constructing the best models they can for their shared purposes and in doing so inevitably use shared knowledge and similar technology. As a result, climate models tend to be similar, sharing many of the same imperfections (Allen and Ingram, 2002; Knutti, 2010; Meehl et al., 2007b; Stainforth et al., 2007a; Tebaldi and Knutti, 2007).

A related problem is that, although model limitations are extensively examined in the literature, discussion of the extent to which models in specific multi-model ensembles differ in ways that are relevant to assessing projections is limited (Knutti et al., 2010).

Recognizing the limited extent to which model error cancels out, some climate scientists have suggested that we should not assume that the larger the ensemble the closer means are to representing reality. Instead, they suggest, one should assume that the correct climate and the climates simulated by models in an ensemble are drawn from the same distribution, e.g., from the standard normal (Gaussian) distribution. Under this new assumption, the failure of an increase in ensemble size to improve simulation results is no longer interpreted as indicating systematic bias. One can then, the suggestion is, assume that when a proportion *r* of an ensemble yield a given projection, *r* is the probability of that projection (Annan and Hargreaves, 2010). But the assumption that model probability distributions coincide with the real climate distribution cannot be made in general, as is illustrated in the case of the already mentioned GCM inability realistically to simulate historical Atlantic MOC collapse. Indeed, structural inadequacy that is known to be shared by ensemble models means that we know that the correct climate *cannot* be represented by current models.

Let us now look at the second argument for appealing to inter-model results in assessing projection quality, the one according to which multi-model ensembles allow us to explore our uncertainty. Since existing climate models share many uncertain assumptions, the projections/projection ranges multi-model ensembles produce do not reflect full explorations of our uncertainty (Parker, 2011; Pirtle et al., 2010).

Moreover, once again, such ensembles do not allow assigning projection qualities the assignment of which involves estimating the full range of possible evolutions of projected quantities.

The GCMs used by WG1 AR4 only sample some of the recognized range of uncertainty about aerosol forcing, perhaps because of the already mentioned tuning relating to this forcing. As a result, the spread of estimated temperature anomalies these models provide (Fig. 1) substantially underestimates the uncertainty about this anomaly and, accordingly, would be misleading as a guide to projection quality (Schwartz et al., 2007). So too, if we take the range of natural variability covered by the simulations represented in Fig. 1 to reflect our uncertainty about natural variability over the next three decades, we will assign a very low probability to the prediction that natural variability will substantially affect GMST trends over this period. Keeping in mind, however, that these models may well similarly and substantially underestimate internal variability over the next 30 years would lead us to reduce our confidence in this prediction. Worse, if we cannot estimate the probability that the ensemble is wrong (something the ensemble cannot help us with!) about internal variability here, we are not in a position to assign the prediction a probability.

A number of suggestions have been made within the climate science community about how partially to address the above worries about the use of multi-model ensembles. Assessments that are explicit about the extent to which climate models in any multi-model ensemble differ in ways that are relevant to assessing projection quality should be offered (IPCC 2010; Knutti et al., 2010). If, for example, internal variability in the MOC is an important source of uncertainty for projections of mean sea surface temperatures over the next 30 years and our ensemble is in the business of making such projections, it should be clear to what extent the simulations produced by the ensemble differ from each other in ways that explore how internal variability in the MOC might occur. Assessing projection quality relevant differences in models is a substantial task, one that goes well beyond the standard multi-model exercise.

In addition, while limited knowledge and resources, e.g., restrictions to certain grid resolutions, mean that there is no question of exploring all of existing uncertainty, provision of second and third best guess modeling attempts could provide a clearer picture of our uncertainty and its impact on CMP quality (Knutti et al., 2010; Smith, 2006).

A difficulty to keep in mind is that of determining how a model component that is shared by complex models that differ in complex ways affects CMP quality. Assessment of model components and their impact on model performance is a challenge that is – because of the need to evaluate models in light of background knowledge – part and parcel of assessing models fitness for purpose. This challenge is complicated when the projection is generated by complex models that implement common components but differ in other complex ways. For the same component may, as a result, function in different ways in different models (Lenhard and Winsberg, 2010). Examining how a parameterization of cloud microphysics affects CMPs may, for example, be hampered if the parameterization scheme is embedded in models that substantially differ in other parameterizations and/or basic theory.

The comparison of substantially differing models will also exacerbate existing challenges for synthesizing the results of multi-model ensembles. Climate scientists have noted that synthesizing the results of different models using a multi-model mean can be misleading even when, as in the case of the CMIP3 models, the models incorporate only, and only standard, representations of atmosphere, ocean, sea ice and land [Knutti et al., 2010]. For example, the CMIP3 multi-model mean of projected local precipitation changes over the next century is 50 % smaller than that which would be expected if we were to assume that at least one, we know not which, of the CMIP3 models is correct. So it seems that using a mean in this case is misleading about what the models describe (Knutti et al., 2010). Synthesizing the results of different models may be even more misleading where models differ substantially in how they represent processes or in which processes they represent, e.g., if some of the models do and some do not include representations of biogeochemical cycles (Tebaldi and Knutti, 2007). In such circumstances, for example, a mean produced by two models may well be a state that is impossible according to both models.

5. The subjective Bayesian approach

Perhaps the main approach to supplement the confidence building approach in WG1 AR4 is the subjective Bayesian approach. We first consider this formal, supplementary approach as it is used to assess projection quality in light of difficulties in parameter choice (Hegerl et al., 2006; Murphy et al., 2004). We then consider how it has been extended.

5.1 The subjective Bayesian approach to parameter estimation

A simple, but representative, application of the standard version of the Bayesian approach to parameter, including projection parameter, estimation involves calculating the posterior probability distribution function $P(F \mid \text{data}, M)$ using Bayes' theorem, as in Eqt. (3) (Frame et al., 2007). $P(F \mid \text{data}, M)$ specifies the probabilities of values of a parameter, F, given data and a model M. $P(\text{data} \mid F, M)$ is the likelihood of F and captures, as a function of values of F, the probability that the data would be simulated by *M*. In the Bayesian context, 'the likelihood of *F*' refers to a probability function for data rather than, as it would on the WG1 AR4 use of 'likelihood', to a probability range for *F*. The prior probability distribution function P(F | M) is the probability distribution function of *F* given only *M* and thus prior to consideration of the data. P(data) is a normalizing constant required to ensure that the probabilities sum up to 1.

$$P(F \mid \text{data}, M) = P(\text{data} \mid F, M)P(F \mid M)/P(\text{data}) \qquad (3)$$

The probabilities in Eqt. (3) are, on the subjective Bayesian approach, to be interpreted as precise, quantitative measures of strength of belief, so called 'degrees of belief'. What makes the subjective Bayesian approach subjective is that unconstrained expert opinion – the beliefs of certain subjects irrespective of whether they meet objective criteria of rationality such as being well grounded in empirical evidence – is used as a central source for selecting prior probability distributions. Still, the subjective Bayesian approach often uses uniform assignments of priors. In doing so, it borrows from what is usually called 'objective Bayesianism' (see Strevens (2006b) for a discussion of the different forms of Bayesian approaches to science).

Bayes' theorem allows us to take existing estimates of parameter uncertainty – here captured by P(F | M) – and to constrain these using information from perturbed physics experiments about how well a model simulates data as a function of parameter settings – information here captured by the likelihood function P(data | F , M). Assume experts provide prior probability distributions for parameters relating to total radiative and present-day indirect aerosol forcing and that we calculate the probability that a model gives, as a function of the parameters' values, to observed oceanic and atmospheric temperature change. Bayes' rule can then yield posterior probability distributions for the parameters (Fig. 5). Bayesian parameter estimation has tended to rely on models of intermediate complexity and on energy balance models.

The Bayesian hope is that the constraints provided by simulation success on parameter estimates will increase the objectivity of such estimates. Moreover, Bayes' theorem provides, what the confidence building approach does not provide, a clear mechanism that relates simulation accuracy to conclusions about CMP quality, thus helping to address the problem of what to infer from available simulation accuracy given the existence of model imperfection.

Nevertheless, the standard version of the Bayesian approach to parameter estimation faces substantial problems. The standard interpretation of the probability distributions $P(F \mid M)$ and $P(F \mid \text{data}, M)$ is that they are probability distributions for F that are conditional on the correctness of a version of M. In the present context, what is being assumed to be correct is a model version in which one or more parameters are unspecified within a certain range. For the goal is to select parameter values from within a range of such values. Now, it is on the basis of the standard interpretation of $P(F \mid M)$ and $P(F \mid \text{data}, M)$ that standard justifications, using so-called Dutch Book arguments, for updating beliefs in accord with Bayes' theorem proceed. Dutch Book arguments generally assume that the, typically statistical, model versions upon which probabilities are conditional are correct. It is argued that, given this assumption, the believer would end up with beliefs that are not as true as they might have been, or would incur a financial loss, if his or her beliefs were not updated in accord with Bayes' theorem (see Jeffrey (1990) and Vineberg (2011) for examples). But if, as in the cases we are concerned with, the model version upon which distributions are conditional is not correct, applying Bayes' theorem may offer no advantage and may be a disadvantage.

Assume that our subject relies on a CMIP3 GCM to determine whether a specified fresh water influx will lead to a collapse in the MOC and that the specified influx is a tenth of that needed to get the model to simulate collapse. Assume also that some exploration of plausible parameter settings in the GCM does not alter results substantially. Applying Bayes's theorem on the assumption that the model is, up to plausible parameter modification, correct means that the probability we assign the outcome ‘collapse’ is 0. The modeler acquiesces to the theorem. Unfortunately, as we now know, the model's results are misleading here. In this case, not applying Bayes' theorem may lead to more realistic judgments.

Thus, the standard use of Bayes' theorem in parameter estimation requires an alternative to the standard interpretation of its conditional probabilities. We will also need an alternative to the standard justifications for applying Bayes' theorem.

Even if we have settled on some interpretation of the conditional posterior probabilities produced by Eqt. (3), there remains the question of what we can infer about reality from these probabilities. There remains, in other words, the question of what distribution of probabilities for F, $P(F)$, we should adopt given the conditional distribution $P(F \mid \text{data}, M)$. We might have a probability distribution for climate sensitivity that is conditional on the data and a model. But what should we infer from this about actual climate sensitivity? We cannot properly answer such questions until we have gone beyond assessing how parameter choice affects projection quality and have also assessed how structural inadequacy, parameterization scheme choice and initial condition inaccuracy do so (Rougier, 2007).

Rougier provides a non-standard version of the Bayesian approach to parameter estimation that has the substantial advantage of allowing us to factor in estimates of structural inadequacy into subjective Bayesian parameter estimates

(Rougier, 2007). Nevertheless, his work takes estimates of structural inadequacy as given and thus does not, by itself, tell us how more comprehensive assessments of projection quality are to be produced.

Additional difficulties for the Bayesian approach relate to the usage of prior probabilities. We rehearse two familiar worries about this usage. First, estimates of $P(F \mid M)$ are usually made after data that bears on the estimates is in hand and it is hard to estimate what probability distribution would be assigned to F independently of knowledge of this data. Failure properly to estimate $P(F \mid M)$ may lead to counting the same data twice, once in estimating priors and once in estimating likelihoods (Frame et al., 2007).

Second, while some climate scientists have argued that the explicit setting out of subjective priors by experts is desirable because it makes subjective judgments explicit (Hargreaves, 2010), philosophers of science have pointed out that it leaves open the question of the extent to which experts' views are evidence based and thus puts reliable and unreliable priors on a par (Sober, 2002). This issue becomes particularly worrying in the context of climate modeling. We know that prior selection may be based on results involving tuning and be required even when data underdetermines parameter value choice. So there is a risk that assigning a prior to a parameter value will beg the question against alternative choices and thus yield estimates of climatic variables we are by no means obliged to accept. The worry of question begging is exacerbated by arguments to the effect that the influence of likelihoods, and thus of data, on the shape and width of prior distributions is often minor (Frame et al., 2005).

A common way of trying to minimize the impact of the appeal to expert opinion is to represent the state of ignorance that existed prior to the consideration of likelihoods using uniform prior distributions within expert specified ranges. We have already seen that uniform distributions are not suitable for representing ignorance. Moreover, to assume a uniform prior distribution will often be to ignore knowledge we have of the relative plausibility of various prior assignments (Annan and Hargreaves, 2011; Rougier, 2007). So too, a uniform assignment of priors for one parameter will sometimes, because of the non-linear relationship between some model variables, provide a non-uniform prior assignment to another (Frame et al., 2005). It has been suggested that best practice given the worries about prior selection is to provide readers with posteriors as well as likelihoods. This would somewhat clarify the role data actually have had in determining posteriors (Frame et al., 2007).

Another way in which the influence of priors might be minimized is by repeated updating of posteriors in response to new evidence over time. As already noted, however, evidence with which to test models is mostly limited to familiar 20th century datasets. There is thus currently limited scope for successive updating of priors.

As to the idea that the appeal to likelihoods in deriving posterior probabilities will provide an objective constraint on parameter selection, it also has problems. Likelihoods measure agreement with data, irrespective of whether such agreement results from tuning (Katzav, 2011). In addition, we have seen that an adequate assessment of projection quality needs to take into account not only agreement with data, but also how error for each simulated quantity develops over projection scenarios as a function of error associated with other such quantities. Finally, there are various likelihood metrics or ways of measuring agreement with data. Choice between these and how such choice affects posteriors is only beginning to be explored (see, e.g., Ishizaki et al. (2010)).

5.2 The subjective Bayesian approach and multi-model ensembles

The subjective Bayesian approach has been extended to assessing multi-GCM ensemble output. This extension, which will be called the subjective Bayesian MM approach, involves taking an ensemble and producing a statistical model of its simulation results. Comparing the statistical model and available data yields a likelihood function that captures the probability the ensemble gives to the data. Bayes' theorem can then be used, in conjunction with the likelihood function and estimates of prior probability distributions for the statistical model's parameters, in order to produce a posterior probability distribution for these parameters (Furrer et al., 2007a; Furrer et al., 2007b; Leith and Chandler, 2010; Tebaldi et al., 2005; Tebaldi and Knutti, 2007).

Some variants of the subjective Bayesian MM approach give each ensemble model equal weight in calculating ensemble posterior probability distributions (Leith and Chandler, 2010). Other variants weight the contribution of each ensemble model to posteriors as a function of how well the model simulates aspects of the climate system (Tebaldi et al., 2005).

Many analyses, e.g., those in WG1 TAR and some of those in WG1 AR4, of multi-model ensemble results produce projections that are just averages of individual model results and that have uncertainty ranges which reflect inter-model variability. This does not yield probabilistic estimates of multi-model ensemble results. The subjective Bayesian MM approach does yield such estimates. The hope is that doing so helps to take into account structural inadequacy and limited knowledge of how to select parameterization schemes. The subjective Bayesian MM approach does not explicitly tackle the issue of how initial condition inaccuracy affects CMP quality.

The subjective Bayesian MM approach suffers from many of the problems of the subjective Bayesian approach to parameter estimation. The subjective Bayesian MM approach faces the problems that arise from the use of prior probabilities. It also suffers from the problems relating to the choice of likelihood metrics and the failure to take into account how error for each simulated quantity develops as a function of error associated with other such quantities. Even weighting models in assessing projection quality is not a clear advantage given that the data used to do so may well have already been used in model construction.

Finally, there remain the issues of how to interpret the conditional probabilities used in Bayes' theorem given model imperfection and of how the conditional probabilities produced by Bayes' theorem relate to unconditional probabilities. On the subjective Bayesian MM approach, one updates priors on the assumption that the statistical model of multi-model ensemble results is correct. However, given that we know that multi-model ensemble results are biased, this assumption is false. And any inference from probabilities that are conditional upon data and an ensemble to unconditional probabilities can only be made given a full assessment of the effects of initial condition error and model imperfection on CMP accuracy. We have seen, however, that multi-model ensembles do not provide such an assessment.

6. The likelihood approach

One response to the subjective Bayesian approach’s difficulties with subjective prior probabilities is to try to avoid the use of priors all together. This is what the likelihood approach does using GCMs. It aims to produce probability distributions for parameters solely in light of how well models simulate data as a function of parameter settings, that is solely in light of likelihood functions such as P(data | *F*, *M*) (Allen et al., 2006). Doing so requires not discounting any parameter settings prior to simulation and thus providing likelihood functions that span a much broader range of parameter values than is usual. This has become possible, though usually only in experiments that perturb the parameters of a single model structure, with the distributed computing techniques used by climateprediction.net (Frame et al., 2007). The results of such attempts are distributions that are less biased due to those parameters that are perturbed, but that are far broader than those otherwise produced.

An application of the likelihood approach is as follows: we take the climate sensitivities of each of a multi-thousand climateprediction.net ensemble of GCM variants and estimate the true climate sensitivity to be a weighted sum of these sensitivities. The weight of each sensitivity is determined by the probability the variant it belongs to gives to observations of a number of climatic quantities, including mean sea level temperature, precipitation and surface heat fluxes (Piani et al., 2005).

The likelihood approach can also be used to minimize the impact of structural inadequacy and uncertainty about choice of parameterization scheme on CMP accuracy. It can do so by producing assessments that are only based on the best simulations available for specific parameter settings (Sanderson et al., 2008). But focusing on best results does not take into account how they are affected by initial condition inaccuracy, tuning or aspects of model imperfection other than parameter choice uncertainty. The same is true of what might be called the multi-model likelihood approach. This approach uses correlations between GCMs' predictions of trends for a quantity and related observations formally to select the best predictions (Boe et al., 2009; Shukla et al., 2006).

7. Putting it all together

As we have noted, WG1 AR4 often uses expert judgment that takes the results of the approaches we have been discussing, as well as partly model-independent approaches, into consideration in assigning final projection qualities. Insofar as final assignments are model based, however, the shared limitations of the approaches we have been discussing remain untouched. In particular, insofar as final assessments are model based, they face serious challenges when it comes to assessing projection quality in light of structural inadequacy, tuning and initial condition inaccuracy. Moreover, they continue to be challenged by the task of assigning probabilities and informative probability ranges to projections.

8. Assessing projections: what else can be done?

We now examine approaches that differ from those that play center stage in WG1 AR4. The first approach, the possibilist approach, is described in the climate science literature but is primarily non-probabilistic. The remaining approaches are philosophy-of-science-based approaches. There are currently four main, but not necessarily mutually exclusive, philosophical approaches to assessing scientific claims. One of these is the already discussed subjective Bayesian approach. The other three are those that are discussed below.

8.1 The possibilist approach

On the possibilist approach, we should present the range of alternative projections provided by models as is, insisting that they are no more than possibilities to be taken into account by researchers and decision makers and that they provide only a lower bound to the maximal range of uncertainty (Stainforth et al., 2007a; Stainforth et al., 2007b). Climate model results should, accordingly, be presented using plots of the actual frequencies with which models have produced specific projections (as in Fig. 6). At the same time, one can supplement projected ranges with informal, though sometimes probabilistic, assessments of confidence in projections that appeal, as the confidence building approach appeals, to inter-model agreement and agreement with physical theory (Stainforth et al., 2007a).

Informal approaches to assessing projection quality must address the same central challenges that quantitative approaches must address. So, insofar as the possibilist position allows informal probabilistic assessments of projection quality, it must address the difficulties that all probabilistic approaches face. However, one could easily purge the possibilist approach of all probabilistic elements and assess projections solely in terms of their being possibilities. Moreover, there are obvious ways to develop purely possibilistic assessment further. Purely possibilistic assessment can, in particular, be used to rank projections. Possibilities can, for example, be ranked in terms of how remote they are.

The purged possibilist approach would still face challenges. Presenting CMPs as possibilities worthy of consideration involves taking a stance on how CMPs relate to reality. For example, if we are presented with an extreme climate sensitivity range of 2 to 11 K (Fig. 6) and are told that these are possibilities that should not have been neglected by AR3 WG1's headline uncertainty ranges (Stainforth et al., 2005), a claim is implicitly being made about which climate behavior is a real possibility. It is implied that these possibilities are unlike, for example, the possibility that the United States will more than halve its budget deficit by 2015. Thus a possibilist assessment of projection quality needs to be accompanied by an examination of whether the projections are indeed real possibilities. The same considerations apply to 'worst case scenarios' when these are put forward as worthy of discussion in policy settings or research. The threat that arises when we do not make sure that possibilities being considered are real possibilities is that, just as we sometimes underestimate our certainty, we will sometimes exaggerate our uncertainty.

Nevertheless, the challenges facing purely possibilistic assessment are substantially more manageable than those facing probabilistic assessment. To say that something is a real possibility at some time *t* is, roughly, to say that it is consistent with the overall way things have been up until *t* and that nothing known excludes it (see Deutsch (1990) for a similar definition). A case for a projection's being a real possibility can, accordingly, be made just by arguing that we have an understanding of the overall way relevant aspects of the climate system are, showing that the projection's correctness is consistent with this understanding and showing that we do not know that there is something that ensures that the projection is wrong. There is, as observed in discussing probabilistic representations of ignorance, no need to specify a full range of alternatives to the projection here. Further, state-of-the-art GCMs can sometimes play an important role in establishing that their projections are real possibilites. State-of-the-art GCMs' projections of GMST are, for example and given the extent to which GCMs capture our knowledge of the climate system, real possibilities.

8.2 The critical approach

The first philosophy-of-science-based approach that is not discussed in the IPCC reports and that will be discussed here is the critical approach (Freedman, 2009; Longino, 1990). According to this approach, scientific claims are rational to the extent that they result from open, critical discussion. Longino offers a prominent view of what such discussion involves. She holds that open critical discussion occurs within a community to the degree that the community has recognized avenues for criticism of evidence, methods, assumptions and reasoning; the community's members share standards of criticism; the community is responsive to criticism and intellectual authority is shared equally among qualified members (Longino, 1990). Petersen offers what can be thought of as a version of the critical approach, one that is designed to assist in, among other things, assessing CMP quality. He provides procedures, and a classification of types of uncertainty, that are supposed to help systematizing qualitative assessments of model assumptions and thus to facilitate open, critical discussion of the quality of model-based-claims (Petersen, 2012).

The motivation for the critical approach is twofold. On the one hand, according to its proponents, critical discussion allows overcoming individual subjective bias. On the other hand, there are no available standards beyond our current standards by which scientific claims can be judged. So, it is argued, rationality cannot amount to more than the application of available standards of critical discussion and the acceptance of the deliverances of these standards.

The critical approach is not really an alternative to the approaches used in WG1 AR4. Rather it is a framework that tells us in what conditions the deliverances of these approaches are acceptable. Petersen’s framework could, for example, be used to guide applying the confidence building approach.

Further, according to the critical approach, we can recognize that an assessment of the quality of a projection is limited while nevertheless accepting the projection. For, on this approach, where acceptance of models' fitness for the purpose generating projections is a result of open, critical discussion, accepting the models' projections is reasonable even if the discussion in question has substantial limitations, e.g., if the impact of unknown structural inadequacy on the projections has not been taken into account. The critical approach would thus, for example, warrant trust in state-of-the-art GCMs for the purpose of generating the GMST projections presented in Fig. 2, subject to expert correction in light of known GCM limitations and provided that the trust results from open, critical discussion.

Acceptance of models' fitness for purpose can, however and as Longino's criteria for such criticism state, only be the result of open, critical discussion if there are shared standards for assessing fitness for purpose. In the absence of shared standards, agreement will be the result of the arbitrary preference of some standards over others rather than the uptake and assessment of relevant alternatives. In the case of CMP assessment, what we need for acceptance of model fitness for purpose to be the result of open, critical discussion is agreement about issues such as whether assessment should be probabilistic, whether it should be formal and so on. The present paper makes it clear, however, that it would be premature to agree on these issues and, indeed, that there is no such agreement.

A more general worry about the critical approach is that, by itself, it leaves unaddressed the question of when the results of open, critical discussion are reliable (Crasnow, 1993). Unless we have an assessment of how reliable current critical discussion of model fitness for purpose is, it is unclear why we should accept the results of such discussion.

8.3 Inference to the best explanation and climate model evaluation

The next philosophy based approach to assessing projection quality is the inference to the best explanation (IBE) approach (Lipton, 2004). In discussing the confidence building approach we saw model confidence being increased on the basis of improvement in model virtues such as agreement with background knowledge (including grounding in basic theory), increased realism, agreement with observations and model scope – that is, roughly, the number of distinct classes of facts a model simulates. An additional model virtue that is appealed to in climate modeling (Shackley, 1997) but is not explicitly discussed in WG1 AR4 is simplicity – which is directly tied to the number and complexity of model assumptions. Yet WG1 AR4 does not, recall, tell us how to map combinations of model virtues onto non-comparative assessments of model confidence. It tells us when confidence should be increased on the basis of model virtues but not when confidence should be high. The IBE approach does and does so in a way that aims to take structural inadequacy into account.

Theories and models explain phenomena in the sense that they provide derivations or simulations that show how phenomena are caused or fit into broader patterns of phenomena (Bokulich, 2011). Thus, GCMs can be said to explain GMST trends and rising sea levels because the simulations they provide show how these phenomena causally depend on anthropogenic greenhouse gas trends. How good the explanations of a model or theory are depends on what combination of virtues the model or theory has. How good a climate model's explanations are, for example, depends on how accurate its simulations are, how detailed its descriptions of climatic mechanisms are, the extent to which it can simulate climate in different periods and so on. This allows proponents of the IBE approach to propose that how confident we should be in a theory or model depends on how good the explanations it provides are, and thus on how good its virtues make its explanations (Lipton, 2004; Thagard, 1978).

That is, it allows the proposal that IBE determines how confident we should be in our explanations. IBE, as applied to models, is just that form of inference which involves:

(i) the possession of alternative explanations of a body of data, where each alternative explanation rests on a model that explains the data;

(ii) a determination of which of the available alternative models that explain the data provides the best available explanation of the data, i.e., of which of these models has the best combination of explanatory virtues;

(iii) an inference to the approximate truth of that model which provides the best available explanation, provided that the model explains the data well enough (this standard presentation of IBE has been adapted from Katzav (2012)).

Since very successful theories do turn out to suffer from unexpected imperfections, even the most optimistic proponents of the IBE approach only allow that the very best explanations are good enough. Explanations that are good enough are usually identified with explanations that are not only empirically successful, simple, of wide scope and well grounded in background knowledge but that also provide confirmed novel predictions, that is confirmed predictions of phenomena that were out-of-sample when they were made *and* unexpected at the time. The idea behind this stringent definition is that, while it is true that the history of science provides examples of successful theories and models that have turned out to be fundamentally wrong, those theories or models which generate confirmed novel predictions arguably tend to survive, at least as approximations, in later theories (see Psillos (1999, pp. 101-111) for a standard discussion). Newtonian mechanics is one of the most successful theories ever, and it lead to its share of novel and confirmed predictions. Of course, like the already mentioned Newtonian Earth-Sun models,, Newtonian mechanics appears to be fundamentally wrong in many ways. But Newtonian mechanics can still be argued to be approximately true. After all, general relativity does show that we can recover the central equations of Newtonian mechanics given the right approximations.

Unfortunately, IBE does not provide a way of assessing the quality of specific classes of CMPs from climate model successes. The IBE approach, like the confidence building approach in WG1 AR4, provides a way of establishing confidence in models as wholes (Katzav, 2012).

Further, how accurate a climate model is depends not only on how good its explanations are but also on how well its parameterization schemes have been engineered to compensate for our limited ability to model climate. So confidence in a climate model, or in its fitness for some purpose, should not depend solely on the quality of its explanations (Katzav, 2012). As to the question whether, in any case, climate models' explanations are good enough to warrant inferring their approximate correctness, it is too complex to be addressed here.

We also need to note the dispute about whether IBE should be relied on. When asked why we should think that IBE allows us to infer the approximate correctness of models when the future might provide us with surprises about model imperfection, proponents of IBE answer that we can only explain the success of our models by supposing that they are approximately true. The success of models would, otherwise, be a miracle (see, e.g., Musgrave (1988) and Worrall (2010)). Winsberg, however, provides examples of highly successful principles that do not appear to be approximately true (Winsberg, 2006). Opponents of IBE point out, further, that the justification of IBE is itself a kind of IBE and thus begs the question of whether IBE is acceptable (Laudan, 1981). The justification aims to get us to trust IBE on the grounds that the best explanation for the successes of a model is its approximate truth. Some, partly in light of the circular justification of IBE, aim to eschew IBE all together. Others, accepting that IBE cannot future proof our estimates of how good our models are, weaken IBE so that it is a form of inference that allows us to rank models according to explanatory capacity but that leaves open the question of how our best models relate to the truth. Yet others insist that IBE is fine roughly as it is, arguing that it is impossible, on pain of an infinite regress, to provide non-circular justification of all basic inferential principles and that IBE is a good candidate fundamental principle for justifying models and theories (see Psillos (1999) for a discussion of some of these views).

8.4 Severe testing, climate models and climate model projections

The remaining approach to assessing scientific claims that we will discuss is the severe testing approach. The idea behind the severe testing approach is that the deliberate search for error is the way to get to the truth. Thus, on this approach, we should assess scientific claims on the basis of how well they have withstood severe testing or probing of their weaknesses (Mayo, 1996; Popper, 2005; Rowbottom, 2011). There are a variety of definitions of ‘severe test’. One prominent definition is Mayo's (Mayo, 1996; Parker, 2008). It, however, requires that for a test of a claim to be severe it must be very unlikely that the claim would pass the test if the claim were false, a requirement that very few tests of climate model fitness for purpose fulfill and thus which would render the severe testing approach largely unhelpful here. We, accordingly, explore the usefulness of the main alternative definition, which is Popper's.

According to Popper, an empirical test of a theory or model is severe to the extent that background knowledge tells us that it is improbable that the theory or model will pass the test. Background knowledge consists in established theories or models other than those being tested (Popper, 2002, p. 150). Popper offers the 1919 test of general relativity's prediction of the precise bending of light in the Sun's gravitational field as an example of a severe test. The observed bending was improbable and indeed inexplicable in light of background knowledge at the time, which basically consisted in Newtonian mechanics. For similar reasons, the precise precession of Mercury also provided a severe test of general relativity.

A crucial difference between the severe testing approach and the approaches pursued by WG1 AR4 is that the severe testing approach never allows mere agreement, or increased agreement, with observations to count in favor of a claim. That simulation of observed phenomena has been successful does not tell us how unexpected the data are and thus how severely the data have tested our claims. If, for example, the successful simulation is the result of tuning, then the success is not improbable, no severe test has been carried out and no increased confidence in model fitness for purpose is warranted. Notice, however, that the fact that claims are tested against in-sample data is not itself supposed to be problematic as long as the data does severely test the claims [Mayo, 1996]. This is illustrated by the prediction of the precession of Mercury. The prediction was not novel or even out-of-sample. It was well measured by Le Verrier in 1859 and was known by Einstein when he constructed his theory (Earman and Glymour, 1978)**.** Another crucial difference between the severe testing approach and those pursued by WG1 AR4 is that the severe testing approach is not probabilistic. The degree to which a set of claims have withstood severe tests, what Popper calls their degree of corroboration, is not a probability.

How might one apply a (Popperian) severe testing approach to assessing projection quality? What we need, from a severe testing perspective, is a framework that assigns a degree of corroboration to a CMP, *p*, as a function of how well the model (or ensemble of models), *m*, which generated *p* has withstood severe tests of its fitness for the purpose of doing so. Such severe tests would consist in examining the performance of some of those of *m*'s predictions the successes of which would be both relevant to assessing *m*'s fitness for the purpose of generating *p* and improbable in light of background knowledge. Assessing, for example, a GCM's projection of 21^{st} century GMST would involve assessing how well the GCM performs at severe tests of relevant predictions of 20^{th} century climate and/or paleoclimate. That is it would involve assessing how well the GCM performs at simulating relevant features of the climate system that we expect will seriously challenge its abilities. A relevant prediction will be one the accuracy of which is indicative of the accuracy of the projection of 21^{st} century GMST. Examples of relevant features of the climate the accurate simulation of which will be a challenge to IPCC-AR5 models are the effects of strong ENSO events on the GMST, effects of Atlantic sea surface temperature variations (associated with the MOC) on the GMST and special aspects of the GMST such as its late 30s and early 40s positive trends. That these data will challenge IPCC-AR5 models is suggested by the difficulty CMIP3 models have in adequately simulating them (Katzav, 2011).

The above ideas about applying the severe testing approach will, as a step towards their operationalization, be elaborated on somewhat and put more formally. *p* is corroborated by data just in case the data are probable in light of *p* but improbable in light of background knowledge, *B*. Symbolically, *p* is corroborated by data just in case $P(data \mid B) < 0.5$ and $C(p \mid \text{data}, B)$ satisfies

$$C(p \mid \text{data}, B) \propto P(data \mid p, B) - P(data \mid B) > 0 \qquad (4)$$

Here $P(data \mid p, B)$ is the probability of the data in light of p and B, and $P(data \mid B)$ is the probability of the data in light of B alone. $C(p \mid \text{data}, B)$ itself results when the right hand side of (1) is normalized so as to yield a result that is between 1 and -1, where 1 signifies the highest degree of corroboration and -1 signifies the highest degree of falsification (Popper, 1983).

Now, we want to assign a degree of corroboration to p as a function of the fitness of m for the purpose of generating p. So one could identify $\mathrm{P}(data \mid p, B)$ with the probability that m gives to data which are relevant to testing m's fitness for the purpose of generating p, that is with $\mathrm{P}(data \mid q, m)$, where q is m's prediction about the relevant data. One could also identify $\mathrm{P}(data \mid B)$ with the probability given to the relevant data by an established rival, $m1$, to m, that is with $\mathrm{P}(data \mid q1, m1)$, where $q1$ is $m1$'s prediction for the data. Thus, in the context of assessing m's suitability for generating p, (4) could be interpreted as:

$$C(p \mid \text{data}, m, m1) \propto P(data \mid q, m) - P(data \mid q1, m1) > 0 \qquad (5)$$

If one's focus is on assessing IPCC-AR5 projections of 21st century GMST, it is natural to identify the probability background knowledge gives to data with the probability the CMIP3 ensemble gives to them. Accordingly, one could, for example, calculate the degree of corroboration of the projection of GMST of a particular AR5 GCM for the 21st century in light of the model's simulation of data relating to ENSO strength by calculating the difference between the probability the model gives to these data – $P(data \mid q, m)$ in (5) – and the probability the CMIP3 ensemble gives to them – $P(data \mid q1, m1)$ in (5).

How might the severe testing approach help us with the difficulties involved in assessing projection quality? The severe testing approach allows us to bypass any worries we might have about tuning since it only counts success that does not result from tuning, success that surely does exist, in favor of CMPs (Katzav, 2011). The severe testing approach can thus, at least, be used as a check on the results of approaches that do not take tuning into account. If, for example, the subjective Bayesian approach assigns a high probability to a projection and the severe testing approach gives the projection a high degree of corroboration, we can at least have some assurance that the probabilistic result is not undermined by tuning.

Underdetermination in choice between parameters/available parameterization schemes might also be addressed by the severe testing approach. Substituting different parameterization schemes into a model might result in varying degrees of corroboration, as might perturbing the model’s parameter settings. Where such variations exist, they allow ranking model fitness for purpose as a function of parameter settings/parameterization schemes. Similarly, degrees of corroboration can be used to rank fitness for purpose of models with different structures. The resulting assessment has, like assessment in terms of real possibilities, the advantage that it is less demanding than probabilistic assessment or assessment that is in terms of truth or approximate truth. Ranking two CMPs as to their degrees of corroboration, for example, only requires comparing the two CMPs. It does not require specifying the full range of alternatives to the CMPs. Nor does it require that we take some stand on how close the CMPs are to the truth, and thus that we take a stand on the effects of unknown structural inadequacy on CMP accuracy. Popper's view is that a ranking in terms of degrees of corroboration only provides us with a ranking of our conjectures about the truth. The most highly corroborated claim would thus, on this suggestion, be our best conjecture about the truth. Being our best conjecture about the truth is, in principle, compatible with being far from the truth.

Consider now some of the limitations of the severe testing approach. To begin with, while the fact that the severe testing approach is, in some respects, less demanding than other approaches has its advantages, it also have its disadvantages. Suppose we rank a claim according to degree of corroboration. What does this imply for the usability of the claim in research and in decision making? Popper’s suggestion that the most highly corroborated claim is our best conjecture about the truth suggests a role for corroboration in the context of research. But when is our best conjecture close enough to the truth to be relevant to practice, e.g., to decision making (Salmon, 1981)? Popper’s response is not straightforward (Miller, 2005). However, one can make use of Popper’s idea that claims should be assessed by severe tests without buying into the rest of his views about science. The beginnings of an alternative response is as follows: the overall degree of corroboration of a claim depends not just on how the claim has done at this or that single test, but also on how broadly it has been tested. A claim's degree of corroboration is thus correlated with the extent to which the claim is consistent with the overall way things are and, therefore, with the extent to which the claim is a real possibility. A high enough degree of corroboration will, accordingly, allow us to conclude that a claim is a real possibility and that it should be used in decision making.

Another basic worry is that our description of the severe testing approach presupposes that we are able to determine, prior to using the severe testing approach, whether data are relevant to assessing fitness for purpose. This includes sometimes being able to determine, independently of the severe testing approach, that inaccuracy in simulating a quantity is not substantially relevant to the accuracy of projections of other quantities. But being able to provide such determinations is something we required of adequate approaches to assessing projection quality.

9. Conclusion

There remain substantial difficulties for WG1 AR4's (climate-model-based) approaches to assessing projection quality, particularly because they aim at probabilistic assessment. Indeed, worries about probabilistic assessment of projection quality are increasingly being raised by those working on projection quality assessment (Parker, 2010; Smith, 2006; Stainforth et al., 2007a).

The commonly used versions of the subjective Bayesian approach leave us, because of their limited ability to represent known climate model imperfection, with a puzzle about why Bayesian updating should be used. Rougier's version does allow a more complete representation of model imperfection, though it does not actually provide us with a way of assessing such imperfection. The likelihood approach was only briefly discussed. It is limited to assessment that takes uncertainty about parameter choice into account. The confidence building approach has the advantage of flexibility. It can, since confidence need not be expressed probabilistically, provide non-probabilistic assessments. So too, the argumentation it uses can in principle be extended to providing assessments of fitness for purpose, though it currently tends to stop at assessing models as such.

In examining approaches not used in WG1 AR4, we saw that the similarity between the confidence building and IBE approaches suggests that IBE might be used to extend the confidence building approach. The many who do not share in the skepticism about IBE will be tempted to use the criterion of explanatory goodness in order to establish the approximate correctness of climate models. At the same time, we saw that the IBE approach does not help us to select which CMPs we are entitled to be confident in. We also saw that considering explanatory quality alone is not the appropriate way of assessing climate model performance. The critical approach provides not so much a way of assessing projection quality as one of systematizing such assessments and legitimizing its results. The legitimization it would provide, however, is problematic because of the lack of agreement about how to assess projection quality and because of the need to address the question of when consensus is a guide to truth.

The possibilist and severe testing approaches are promising in that they propose specific non-probabilistic measures of CMP quality. The severe testing approach has the additional advantage that it provides a way of trying to get a handle on the effects of tuning on CMP accuracy. As we have noted, however, both possibilist and severe testing approaches face problems.

Some of the difficulties that arise in assessing projection quality are difficulties that would arise irrespective of actual projection accuracy. Tuning may well not affect the ability of models reliably to generate some important class of projections. But our uncertainty about the very practice of tuning means that, even if the projections in question are accurate and reliably generated, we will find it difficult to decide whether they are accurate. Similarly, the non-linear nature of the climate system may well not adversely affect the accuracy of some class of projections. But our uncertainty about whether non-linearity is pertinent to the projections will mean that we will find it difficult to decide whether they are accurate. This is frustrating, but does not alter the predicament we find ourselves in with respect to developing adequate approaches to assessing projection quality.

References

Abe, M., H. Shiogama, J. C. Hargreaves, J. D. Annan, T. Nozawa and S. Emori, Correlation between Inter-Model Similarities in Spatial Pattern for Present and Projected Future Mean Climate, *Sola*, *5*, 133-136, 2009.

Allen, M. R., J. Kettleborough and D. A. Stainforth, Model error in weather and climate forecasting, in Predictability of weather and climate, edited by T. Palmer and R. Hagedorn, pp. 391-427, Cambridge University Press, Cambridge, 2006.

Allen, M. R. and W. J. Ingram, Constraints on future changes in climate and the hydrologic cycle, *Nature*, *419*(6903), 224-232, 2002.

Annan, J. D. and J. C. Hargreaves, Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, *37*(2), L02703, 2010.

Annan, J. D. and J. C. Hargreaves, On the generation and interpretation of probabilistic estimates of climate sensitivity, *Climatic Change*, *104*(3-4), 423-436, 2011.

Arneth, A., S. P. Harrison, S. Zaehle, K. Tsigaridis, S. Menon, P. J. Bartlein, J. Feichter, A. Korhola, M. Kulmala, D. O'Donnell, G. Schurgers, S. Sorvari and T. Vesala, Terrestrial biogeochemical feedbacks in the climate system, *Nature Geosci*, *3*(8), 525-532, 2010.

Barrett, J. and P. K. Stanford, Prediction, in The Philosophy of Science: An Encyclopedia, edited by J. Pfeifer and S. Sarkar, pp. 589-599, Routledge, New York, 2006.

Bentley, M. J., The Antarctic palaeo record and its role in improving predictions of future Antarctic Ice Sheet change, *Journal of Quaternary Science*, *25*(1), 5-18, 2010.

Boe, J., A. Hall and X. Qu, September sea-ice cover in the Arctic Ocean projected to vanish by 2100, *Nature Geosci*, *2*(5), 341-343, 2009.

Bokulich, A., How scientific models can explain, *Synthese*, *180*(1), 33-45, 2011.

Braconnot, P., B. Otto-Bliesner, S. Harrison, S. Joussaume, J. Y. Peterchmitt, A. be-Ouchi, M. Crucifix, E. Driesschaert, T. Fichefet, C. D. Hewitt, M. Kageyama, A. Kitoh, A. Laine, M. F. Loutre, O. Marti, U. Merkel, G. Ramstein, P. Valdes, S. L. Weber, Y. Yu and Y. Zhao, Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum - Part 1: experiments and large-scale features, *Climate of the Past*, *3*(2), 261-277, 2007.

Branstator, G. and H. Teng, Two Limits of Initial-Value Decadal Predictability in a CGCM, *J. Climate*, *23*(23), 6292-6311, 2010.

Caseldine, C. J., C. Turney and A. J. Long, IPCC and palaeoclimate - an evolving story?, *Journal of Quaternary Science*, *25*(1), 1-4, 2010.

CCSP, Atmospheric Aerosol Properties and Climate Impacts. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research. [Mian Chin, Ralph A. Kahn, and Stephen E. Schwartz (eds.)]. National Aeronautics and Space Administration, Washington, D.C., USA., 2009.

Christensen, J. H., B. Hewitson, A. Busuioc, A. Chen, X. Gao, I. Held, R. Jones, R. K. Kolli, W. T. Kwon, R. Laprise, V. Magaña Rueda, L. Mearns, C. G. Menéndez, J. Räisanen, A. Rinke, A. Sarr and P. Whetton, Regional Climate Projections, in Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. M. Averyt, M. Tignor and H. L. Miller, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

CMIP3, World Climate Research Programme's Coupled Model Intercomparison Project phase 3 multi-model dataset (on line), *http://www-pcmdi. llnl. gov/ipcc/about_ipcc. php*, 2007.

Crasnow, S. L., Can Science be Objective? Longino's Science as Social Knowledge, *Hypatia*, *8*, 194-201, 1993.

Dall' Amico, M., P. Stott, A. Scaife, L. Gray, K. Rosenlof and A. Karpechko, Impact of stratospheric variability on tropospheric climate change, *Climate Dynamics*, *34*(2), 399-417, 2010.

Dessler, A. E. and S. C. Sherwood, Atmospheric Science: A Matter of Humidity, *Science*, *323*(5917), 1020-1021, 2009.

Deutsch, H., Real Possibility, *Nous*, *24*(5), 751-755, 1990.

Earman, J. and C. Glymour, Einstein and Hilbert - 2 Months in History of General Relativity, *Archive for History of Exact Sciences*, *19*(3), 291-308, 1978.

Frame, D. J., B. B. B. Booth, J. A. Kettleborough, D. A. Stainforth, J. M. Gregory, M. Collins and M. R. Allen, Constraining climate forecasts: The role of prior assumptions, *Geophysical Research Letters*, *32*(9), L09702, 2005.

Frame, D. J., N. E. Faull, M. M. Joshi and M. R. Allen, Probabilistic climate forecasts and inductive problems, *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences*, *365*(1857), 1971-1992, 2007.

Freedman, K., Diversity and the Fate of Objectivity, *Social Epistemology*, *23*(1), 45-56, 2009.

Furrer, R., R. Knutti, S. R. Sain, D. W. Nychka and G. A. Meehl, Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis, *Geophysical Research Letters*, *34*(6), L06711, 2007a.

Furrer, R., S. R. Sain, D. Nychka and G. A. Meehl, Multivariate Bayesian analysis of atmosphere - Ocean general circulation models, *Environmental and Ecological Statistics*, *14*(3), 249-266, 2007b.

Ghil, M., M. D. Chekroun and E. Simonnet, Climate dynamics and fluid mechanics: Natural variability and related uncertainties, *Physica D: Nonlinear Phenomena*, *237*(14-17), 2111-2126, 2008.

Gleckler, P. J., K. E. Taylor and C. Doutriaux, Performance metrics for climate models, *Journal of Geophysical Research-Atmospheres, 113*(6), D06104, 2008.

Halpern, J. Y., Reasoning about uncertainty, MIT Press, London and Cambridge Massachusetts, 2003.

Hargreaves, J., Skill and uncertainty in climate models, *Wiley Interdisciplinary Reviews: Climate Change*, 556-564, 2010.

Hawkins, E. and R. Sutton, The Potential to Narrow Uncertainty in Regional Climate Predictions, *Bull. Amer. Meteor. Soc., 90*(8), 1095-1107, 2009.

Hegerl, G. C., T. J. Crowley, W. T. Hyde and D. J. Frame, Climate sensitivity constrained by temperature reconstructions over the past seven centuries, *Nature, 440*(7087), 1029-1032, 2006.

Hegerl, G. C., F. W. Zwiers, P. Braconnot, N. P. Gillett, Y. Yuo, J. A. Marengo Orsini, N. Nicholls, J. E. Penner and P. A. Stott, Understanding and Attributing Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Huber, M. and R. Caballero, The early Eocene equable climate problem revisited, *Clim. Past, 7*(2), 603-633, 2011.

Hudson, R., What is Really at Issue with Novel Predictions?, *Synthese, 155*(1), 1-20, 2007.

Humphreys, P., Extending ourselves: computational science, empiricism and scientific method, Oxford University Press, Oxford, 2004.

Hurrell, J., G. Meehl, D. Bader, T. Delworth, B. Kirtman and B. Wielicki, A Unified Modeling Approach to Climate System Prediction, *Bull. Amer. Meteor. Soc., 90*(12), 1819-1832, 2009.

IPCC 1990, Climate Change: The IPCC Scientific Assessment, Cambridge University Press, Cambridge, UK and New York, 1990.

IPCC 1996, Climate Change 1995: The Science of Climate Change. Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK and New York, 1996.

IPCC 2001, Climate change 2001: The scientific basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, 2001.

IPCC 2005, Guidance Notes for Lead Authors of the IPCC Fourth Assessment Report on Addressing Uncertainties (on line), *http://www. ipcc. ch/meetings/ar4-workshops-express-meetings/uncertainty-guidance-note. pdf*, 2005.

IPCC 2007, Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, Pachauri, R.K and Reisinger, A. (eds.)]. IPCC, Geneva, Switzerland, 104 pp., 2007.

IPCC 2010, 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, University of Bern, Bern, 2010.

Ishizaki, Y., T. Nakaegawa and I. Takayabu, Comparison of Three Bayesian Approaches to Project Surface Air Temperature Changes over Japan Due to Global Warming, *Sola, 6*, 21-24, 2010.

Jansen, E., J.Overpeck, K.R.Briffa, J.-C.Duplessy, F.Joos, V.Masson-Delmotte, D.Olago, B.Otto-Bliesner, W.R.Peltier, S.Rahmstorf, R.Ramesh, D.Raynaud, D.Rind, O.Solomina, R.Villalba and D.Zhang, Palaeoclimate, in *Climate Change* 2007: *The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. M. Averyt, M. Tignor and H. L. Miller, Cambridge University Press, Cambridge, UK and New York, 2007.

Jeffrey, R. C., The logic of decision, The University of Chicago Press, Chicago and London, 1990.

Katzav, J., Should we assess climate model predictions in light of severe tests?, *Eos, Transactions American Geophysical Union, 92(23)*, 195, 2011.

Katzav, J., Hybrid models, climate models and inference to the best explanation, *British Journal for the Philosophy of Science, doi: 10.1093/bjps/axs002*, 2012.

Kirkby, J., Cosmic Rays and Climate, *Surveys in Geophysics, 28*(5), 333-375, 2007.

Knutti, R., Should we believe model predictions of future climate change?, *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences, 366*(1885), 4647-4664, 2008.

Knutti, R., The end of model democracy?, *Climatic Change, 102*, 395-404, 2010.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak and G. A. Meehl, Challenges in Combining Projections from Multiple Climate Models, *J. Climate, 23*(10), 2739-2758, 2010.

Knutti, R., T. F. Stocker, F. Joos and G. K. Plattner, Constraints on radiative forcing and future climate change from observations and climate model ensembles, *Nature, 416*(6882), 719-723, 2002.

Knutti, R. and G. C. Hegerl, The equilibrium sensitivity of the Earth's temperature to radiation changes, *Nature Geosci*, *1*(11), 735-743, 2008.

Laskar, J., P. Robutel, F. Joutel, M. Gastineau, A. C. M. Correia and B. Levrard, A long-term numerical solution for the insolation quantities of the Earth, *Astronomy & Astrophysics, 428*(1), 261-285, 2004.

Laudan, L., A Confutation of Convergent Realism, *Philosophy of Science, 48*(1), 19-49, 1981.

Le Treut, H., R. Somerville, U. Cubasch, Y. Ding, C. Mauritzen, A. Mokssit, T. Peterson and M. Prather, Historical overview of climate change, in *Climate Change* 2007: *The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. M. Averyt, M. Tignor and H. L. Miller, Cambridge University Press, New York, 2007.

Leith, N. A. and R. E. Chandler, A framework for interpreting climate model outputs, *Journal of the Royal Statistical Society Series C-Applied Statistics, 59*, 279-296, 2010.

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

Lenton, T. M., H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf and H. J. Schellnhuber, Tipping elements in the Earth's climate system, *Proceedings of the National Academy of Sciences of the United States of America, 105*(6), 1786-1793, 2008.

Lipton, P., *Inference to the best explanation*, Routledge, London and New York, 2004.

Longino, H. E., Science as Social Knowledge: Values and Objectivity in Scientific Inquiry, Princeton University Press, Princeton, 1990.

Lorenz, E., Deterministic Nonperiodic Flow, *Journal of Atmospheric Science, 20*, 103-141, 1963.

Mahmood, R., R. A. Pielke, K. G. Hubbard, D. Niyogi, G. Bonan, P. Lawrence, R. McNider, C. McAlpine, A. Etter, S. Gameda, B. D. Qian, A. Carleton, A. Beltran-Przekurat, T. Chase, A. I. Quintanar, J. O. Adegoke, S. Vezhapparambu, G. Conner, S. Asefi, E. Sertel, D. R. Legates, Y. L. Wu, R. Hale, O. W. Frauenfeld, A. Watts, M. Shepherd, C. Mitra, V. G. Anantharaj, S. Fall, R. Lund, A. Trevino, P. Blanken, J. Y. Du, H. I. Chang, R. E. Leeper, U. S. Nair, S. Dobler, R. Deo and J. Syktus, Impacts of Land Use/Land Cover Change on Climate and Future Research Priorities, *Bull. Amer. Meteor. Soc., 91*(1), 37-46, 2010.

Mayo, D. G., Error and the Growth of Experimental Knowledge, The University of Chicago Press, Chicago and London, 1996.

McWilliams, J. C., Irreducible imprecision in atmospheric and oceanic simulations, *Proceedings of the National Academy of Sciences, 104*(21), 8709-8713, 2007.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer and K. E. Taylor, The WCRP CMIP3 multimodel dataset - A new era in climate change research, *Bull. Amer. Meteor. Soc., 88*(9), 1383-1394, 2007a.

Meehl, G. A., T. F. Stocker, W. D. Collins, P. Friedlingstein, A. P. Gaye, J. M. Gregory, A. Kitoh, R. Knutti, J. M. Murphy, A. Noda, S. C. B. Raper, I. G. Watterson, A. J. Weaver and Z.-C. Zhao, Global Climate Projections, in *Climate Change* 2007: *The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. M. Averyt, M. Tignor and H. L. Miller, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007b.

Miller, D., Out of error: further essays on critical rationalism, Ashgate Publishing Limited, Aldershot, 2005.

Moss, R. and S. H. Schneider, Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment and reporting, in Guidance papers on the cross cutting issues of the third assessment report of the IPCC. Technical report., pp. 33-51, World Meteorological Organization, Geneva, 2000.

Murphy, J. M., B. B. B. Booth, M. Collins, G. R. Harris, D. M. H. Sexton and M. J. Webb, A Methodology for Probabilistic Predictions of Regional Climate Change from Perturbed Physics Ensembles, *Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 365*(1857), 1993-2028, 2007.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb and M. Collins, Quantification of modelling uncertainties in a large ensemble of climate change simulations, *Nature, 430*(7001), 768-772, 2004.

Musgrave, A., The ultimate argument for scientific realism, in Relativism and realism in science, edited by R. Nola, pp. 229-252, Kluwer, Dordrecht, 1988.

North, J., An empirical approach to symmetry and probability, *Studies in History and Philosophy of Modern Physics, 41*(1), 27-40, 2010.

Norton, J. D., Ignorance and indifference, *Philosophy of Science, 75*(1), 45-68, 2008.

Norton, J. D., Challenges to Bayesian Confirmation Theory, in Philosophy of Statistics, vol. 7, edited by P. S. Bandyopadhyay and M. Forster, pp. 391-440, Elsevier, New York, 2011.

Odenbaugh, J. and A. Alexandrova, Buyer beware: robustness analyses in economics and biology, *Biology and Philosophy, 26*(5), 757-771, 2011.

Palmer, T. N., Predicting uncertainty in forecasts of weather and climate, *Reports on Progress in Physics, 63*(2), 71-116, 2000.

Palmer, T. N., G. J. Shutts, R. Hagedorn, E. Doblas-Reyes, T. Jung and M. Leutbecher, Representing model uncertainty in weather and climate prediction, *Annual Review of Earth and Planetary Sciences*, *33*, 163-193, 2005.

Parker, W. S., Computer simulation through an error-statistical lens, *Synthese*, *163*(3), 371-384, 2008.

Parker, W. S., Confirmation and Adequacy-for-Purpose in Climate Modelling, *Aristotelian Society Supplementary Volume*, *83*(1), 233-249, 2009.

Parker, W. S., When Climate Models Agree: The Significance of Robust Model Predictions, *Philosophy of Science*, *78*(4), 579-600, 2011.

Parker, W. S., Predicting weather and climate: Uncertainty, ensembles and probability, *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, *41(3)*, 263-272, 2010.

Petersen, A. C., Simulating Nature: A Philosophical Study of Computer Simulation Uncertainties and their Role in Climate Science and Policy Advice, CRC Press, Boca Raton, FL, 2012.

Piani, C., D. J. Frame, D. A. Stainforth and M. R. Allen, Constraints on climate change from a multi-thousand member ensemble of simulations, *Geophys. Res. Lett.*, *32*(23), L23825, 2005.

Pielke, R. A., Land use and climate change, *Science*, *310*(5754), 1625-1626, 2005.

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

Pitman, A. J. and R. J. Stouffer, Abrupt change in climate and climate models, *Hydrology and Earth System Sciences*, *10*(6), 903-912, 2006.

Popper, K. R., Realism and the aim of science, Routledge, London and New York, 1983.

Popper, K. R., Conjectures and refutations: the growth of scientific knowledge, Routledge, London and New York, 2002.

Popper, K. R., The logic of scientific discovery, Routledge, London and New York, 2005.

Psillos, S., Scientific realism: how science tracks truth, Routledge, London and New York, 1999.

Raisanen, J. and T. N. Palmer, A probability and decision-model analysis of a multimodel ensemble of climate change simulations, *J. Climate*, *14*(15), 3212-3226, 2001.

Randall, D. A., R. A. Wood, S. Bony, R. Coleman, T. Fichefet, J. Fyfe, V. Kattsof, A. Pitman, J. Shukla, J. Srinivasan, R. J. Stouffer, A. Sumi and K. E. Taylor, Climate Models and Their Evaluation, in *Climate Change* 2007: *The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. M. Averyt, M. Tignor and H. L. Miller, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Randall, D., M. Khairoutdinov, A. Arakawa and W. Grabowski, Breaking the Cloud Parameterization Deadlock, *Bull. Amer. Meteor. Soc.*, *84*(11), 1547-1564, 2003.

Rial, J. A., R. A. Pielke, M. Beniston, M. Claussen, J. Canadell, P. Cox, H. Held, N. De Noblet-Ducoudre, R. Prinn, J. F. Reynolds and J. D. Salas, Nonlinearities, feedbacks and critical thresholds within the Earth's climate system, *Climatic Change*, *65*(1-2), 11-38, 2004.

Rougier, J., Probabilistic inference for future climate using an ensemble of climate model evaluations, *Climatic Change*, *81*(3), 247-264, 2007.

Rowbottom, D., Popper's Critical Rationalism: A Philosophical Investigation, Routledge, New York, 2011.

Saatsi, J., On the Pessimistic Induction and Two Fallacies, *Philosophy of Science*, *72*(5), 1088-1098, 2005.

Salmon, W. C., Rational Prediction, *British Journal for the Philosophy of Science*, *32*(2), 115-125, 1981.

Sanderson, B. M., R. Knutti, T. Aina, C. Christensen, N. Faull, D. J. Frame, W. J. Ingram, C. Piani, D. A. Stainforth, D. A. Stone and M. R. Allen, Constraints on Model Response to Greenhouse Gas Forcing and the Role of Subgrid-Scale Processes, *J. Climate*, *21*(11), 2384-2400, 2008.

Schmidt, G. A., Enhancing the relevance of palaeoclimate model/data comparisons for assessments of future climate change, *Journal of Quaternary Science*, *25*(1), 79-87, 2010.

Schmittner, A., M. Latif and B. Schneider, Model projections of the North Atlantic thermohaline circulation for the 21st century assessed by observations, *Geophysical Research Letters*, *32*(23), L23710, 2005.

Schwartz, S., R. J. Charlson and H. Rodhe, Quantifying climate change - too rosy a picture?, *Nature Reports Climate Change*, *2*, 23-24, 2007.

Semenov, V. A., M. Latif, D. Dommenget, N. S. Keenlyside, A. Strehz, T. Martin and W. Park, The Impact of North AtlanticΓÇôArctic Multidecadal Variability on Northern Hemisphere Surface Air Temperature, *J. Climate*, *23*(21), 5668-5677, 2010.

Shackley, S., Epistemic lifestyles in climate change modeling, in Changing the atmosphere, edited by C. A. Miller and Edwards P.N., pp. 109-133, The MIT Press, Cambridge Mass. and London, 1997.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter and D. Paolino, Climate model fidelity and projections of climate change, *Geophys. Res. Lett., 33*(7), L07702, 2006.

Siddall, M., A. be-Ouchi, M. Andersen, F. Antonioli, J. Bamber, E. Bard, J. Clark, P. Clark, P. Deschamps, A. Dutton, M. Elliot, C. Gallup, N. Gomez, J. Gregory, P. Huybers, K. Kawarnura, M. Kelly, K. Lambeck, T. Lowell, J. Milrovica, B. Otto-Bliesner, D. Richards, J. Stanford, C. Stirling, T. Stocker, A. Thomas, B. Thompson, T. Tornqvist, N. V. Riveiros, C. Waelbroeck, Y. Yokoyama and S. Y. Yu, The sea-level conundrum: case studies from palaeo-archives, *Journal of Quaternary Science, 25*(1), 19-25, 2010.

Smith, L. A., Predictability past, predictability present, in Predictability of Weather and Climate, edited by T. Palmer and R. Hagedorn, pp. 217-250, Cambridge University Press, Cambridge, 2006.

Snyder, C. W., The value of paleoclimate research in our changing climate, *Climatic Change*, *100*(3-4), 407-418, 2010.

Sober, E., Bayesianism — its Scope and Limits, in Bayes' Theorem, edited by R. Swinburne, pp. 21-38, Oxford University Press, Oxford, 2002.

Spicer, R. A., A. Ahlberg, A. B. Herman, C. C. Hofmann, M. Raikevich, P. J. Valdes and P. J. Markwick, The Late Cretaceous continental interior of Siberia: A challenge for climate models, *Earth and Planetary Science Letters*, *267*(1-2), 228-235, 2008.

Stainforth, D. A., T. Aina, C. Christensen, M. Collins, N. Faull, D. J. Frame, J. A. Kettleborough, S. Knight, A. Martin, J. M. Murphy, C. Piani, D. Sexton, L. A. Smith, R. A. Spicer, A. J. Thorpe and M. R. Allen, Uncertainty in predictions of the climate response to rising levels of greenhouse gases, *Nature, 433*(7024), 403-406, 2005.

Stainforth, D. A., M. R. Allen, E. R. Tredger and L. A. Smith, Confidence, uncertainty and decision-support relevance in climate predictions, *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences, 365*(1857), 2145-2161, 2007a.

Stainforth, D. A., T. E. Downing, R. Washington, A. Lopez and M. New, Issues in the interpretation of climate model ensembles to inform decisions, *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences, 365*(1857), 2163-2177, 2007b.

Stanford, P. K., Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives, Oxford University Press, New York, 2006.

Stott, P. A., J. F. B. Mitchell, M. R. Allen, T. L. Delworth, J. M. Gregory, G. A. Meehl and B. D. Santer, Observational Constraints on Past Attributable

Warming and Predictions of Future Global Warming, *J. Climate*, *19*(13), 3055-3069, 2006.

Strevens, M., Probability and chance, in Macmillan Encyclopedia of Philosophy, vol. 8, edited by D. M. Borchert, pp. 24-40, MacMillan Reference USA. Thomson Gale., New York, 2006a.

Strevens, M., The Bayesian Approach to the Philosophy of Science, in Encyclopedia of Philosophy, edited by D. M. Borchert, Macmillan Reference, Detroit, 2006b.

Swanson, K. L., G. Sugihara and A. A. Tsonis, Long-term natural variability and 20th century climate change, *Proceedings of the National Academy of Sciences, 106*(38), 16120-16123, 2009.

Tebaldi, C. and R. Knutti, The use of the multi-model ensemble in probabilistic climate projections, *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences, 365*(1857), 2053-2075, 2007.

Tebaldi, C., R. L. Smith, D. Nychka and L. O. Mearns, Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles, *J. Climate, 18*(10), 1524-1540, 2005.

Thagard, P. R., Best Explanation - Criteria for Theory Choice, *Journal of Philosophy, 75*(2), 76-92, 1978.

Valdes, P., Built for stability, *Nature Geosci, 4*(7), 414-416, 2011.

Vineberg, S., Dutch Book Arguments (on line), *the Stanford Encyclopedia of Philosophy, http://plato. stanford. edu/archives/sum2011/entries/dutch-book/*, 2011.

Weisberg, M., Robustness analysis, *Philosophy of Science, 73*(5), 730-742, 2006.

Wigley, T. M. L., Climate-Change - Outlook Becoming Hazier, *Nature, 369*(6483), 709-710, 1994.

Winsberg, E., Models of success versus the success of models: Reliability without truth, *Synthese, 152*(1), 1-19, 2006.

Worrall, J., Error, tests and theory confirmation, in Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science, edited by D. G. Mayo and A. Spanos, pp. 125-154, Cambridge University Press, New York, 2010.

Wu, Z., N. Huang, J. Wallace, B. Smoliak and X. Chen, On the time-varying trend in global-mean surface temperature, *Climate Dynamics*, 1-15, 2011.

Wunsch, C., Towards understanding the Paleocean, *Quaternary Science Reviews, 29*(17-18), 1960-1967, 2010.

Wyatt, M., S. Kravtsov and A. Tsonis, Atlantic Multidecadal Oscillation and Northern Hemisphere's climate variability, *Climate Dynamics*, 1-21, 2011.

Captions

Fig. 1 Temperature changes relative to the corresponding average for 1901-1950 (°C) from decade to decade from 1906 to 2005 over the entire globe, global land area and the global ocean. The black line indicates observed temperature change, while the colored bands show the combined range covered by 90% of recent model simulations. Red indicates simulations that include natural and human factors, while blue indicates simulations that include only natural factors. Dashed black lines indicate decades and continental regions for which there are substantially fewer observations. Adapted from Hegerl et al., FAQ9.2, Fig. 1 (2007, p. 703).

Fig. 2 Projected 21st century global mean temperatures changes for various greenhouse gas emission scenarios. Solid lines are multi-model global averages of surface warming for scenarios A2, A1B and B1, shown as continuations of the 20th-century simulations. These projections also take into account emissions of short-lived GHGs and aerosols. The pink line is not a scenario, but is for Atmosphere-Ocean General Circulation Model (AOGCM) simulations where atmospheric concentrations are held constant at year 2000 values. The bars at the right of the figure indicate the best estimate (solid line within each bar) and the likely range assessed for the six SRES marker scenarios at 2090-2099. All temperatures are relative to the period 1980-1999. Adapted from the Synthesis Report for IPCC AR4, Fig. 3.2 (2007, p. 7).

Fig. 3 Evolution of the MOC at 30°N in simulations with the suite of comprehensive coupled climate models from 1850 to 2100 using 20th Century Climate in Coupled Models (20C3M) simulations for 1850 to 1999 and the SRES A1B emissions scenario for 1999 to 2100. Some of the models continue the integration to year 2200 with the forcing held constant at the values of year 2100. Observationally based estimates of late-20th century MOC are shown as vertical bars on the left. Adapted from Meehl et al., Fig. 10.15 (2007b, p. 773), who build on Schmittner et al. (2005).

Fig. 4. Root-mean-square (RMS) error of 1980–99 surface temperature (averaged over space, relative to the 40-year reanalysis of the European Centre of Medium range Weather Forecast) shown as a function of the number of models included in the model average. Panel (a) shows the December-January-February period (DJF), panel (b) the June-July-August (JJA) period. Red dashed lines indicate the range covered by randomly sampling the models for the subset; the red solid line indicates the average. The RMS error converges to a constant value that is more than half of the initial value for one model. The black dashed line is a theoretical RMS. If the model biases were independent, then the RMS error for a large sample of models should decrease with the square root of the number of models (dotted). The blue line results if the models are sorted by how well they agree with DJF and JJA observations combined, and it indicates that the average of a few good models outperforms an average of more models with poorer performance. Adapted from Knutti et al., Figs 3(c) and 3(d) (2010, p. 2744).

Fig. 5 Constraints on the radiative forcing from the observed atmospheric and oceanic warming. Probability density functions (PDF) for the total (anthropogenic and natural) radiative forcing (a–c) and the indirect aerosol forcing (d–f) in the year

2000 are based on 25,000 Monte Carlo simulations. The initially assumed PDFs are given in a and d. The requirement that the model matches the temperature observations strongly narrows the PDFs (b and e). If in addition the climate sensitivity is restricted to the range adopted by the IPCC (1.5–4.5 K), the PDFs in c and f are obtained. Adapted from Knutti et al., Fig. 2 (2002, p. 720).

Fig. 6. The response to parameter perturbations: the frequency distribution of simulated climate sensitivity using all model versions (black), all model versions except those with perturbations to the cloud-to-rain conversion threshold (red), and all model versions except those with perturbations to the entrainment coefficient (blue). Adapted from Stainforth et al, Fig. 2(a) (2005, p. 404).

Figures

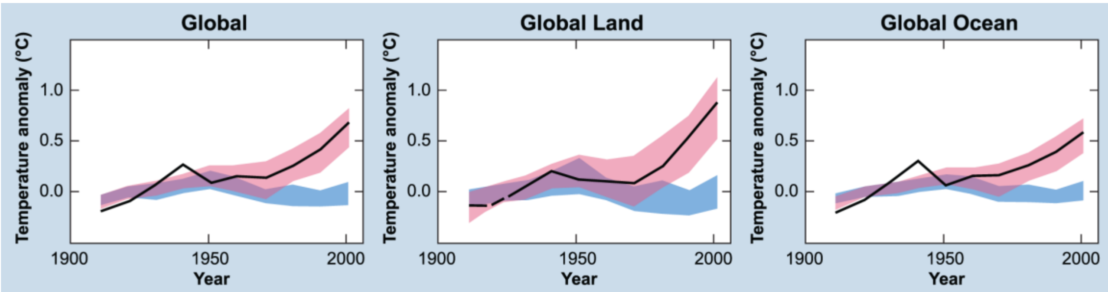

Fig. 1

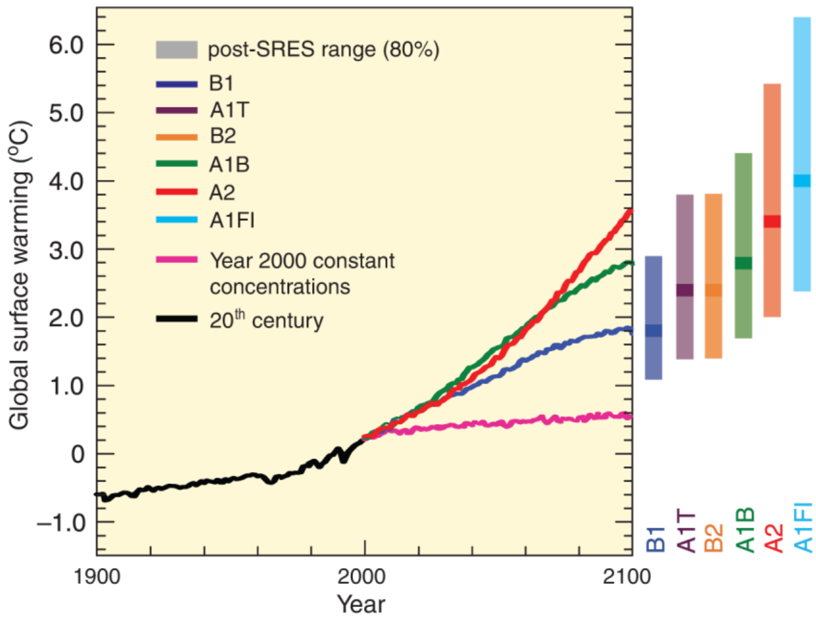

Fig. 2

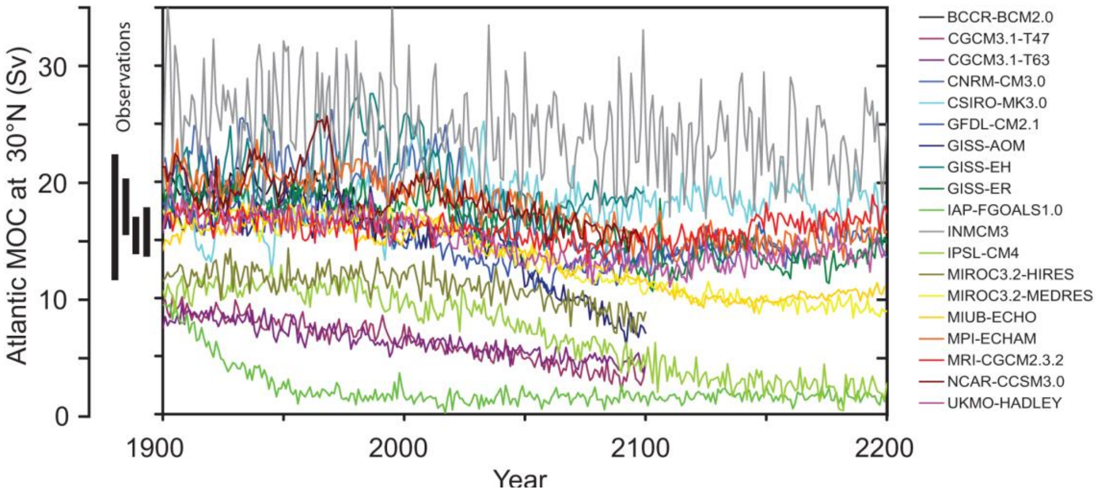

Fig. 3

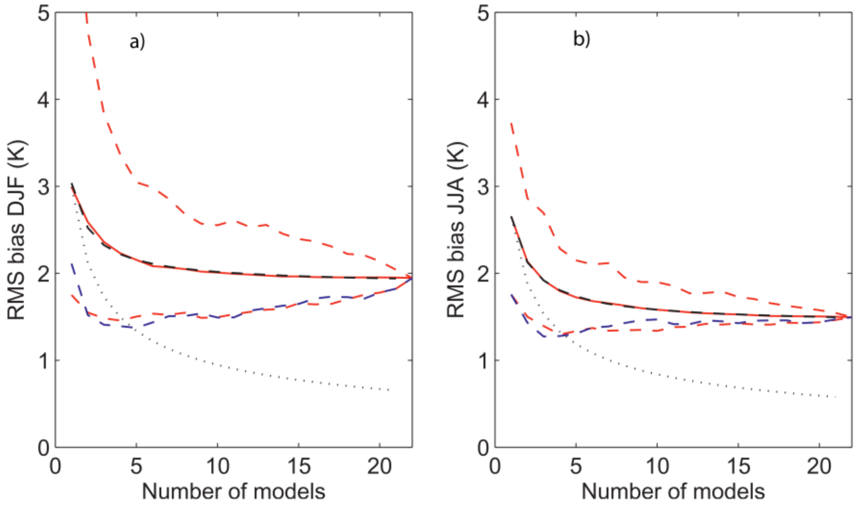

Fig. 4

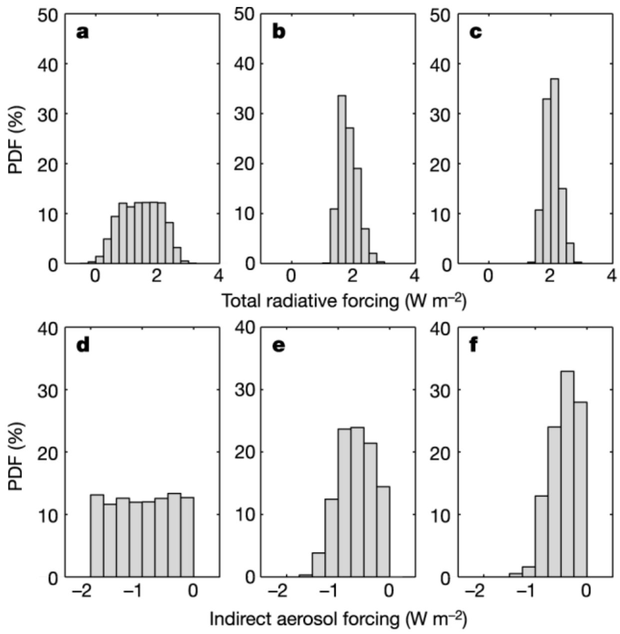

Fig. 5

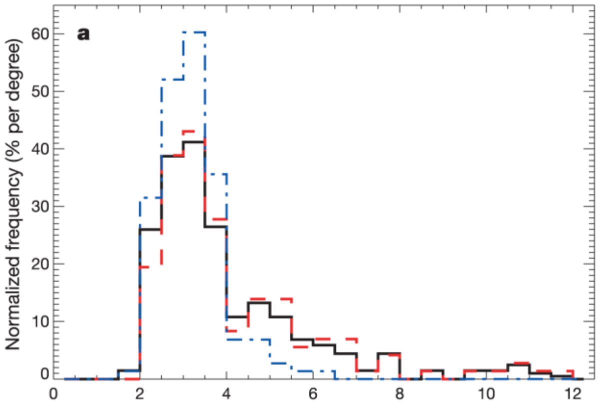
a
Normalized frequency (% per degree)
0
10
20
30
40
50
60
0
2
4
6
8
10
12

Fig. 6