Euro Jnl Phil Sci
DOI 10.1007/s13194-014-0102-9

ORIGINAL PAPER IN PHILOSOPHY OF SCIENCE

A practical philosophy of complex climate modelling

Gavin A. Schmidt · Steven Sherwood

Received: 25 February 2014 / Accepted: 12 November 2014

Abstract We give an overview of the practice of developing and using complex climate models, as seen from experiences in a major climate modelling center and through participation in the Coupled Model Intercomparison Project (CMIP). We discuss the construction and calibration of models; their evaluation, especially through use of out-of-sample tests; and their exploitation in multi-model ensembles to identify biases and make predictions. We stress that adequacy or utility of climate models is best assessed via their skill against more naïve predictions. The framework we use for making inferences about reality using simulations is naturally Bayesian (in an informal sense), and has many points of contact with more familiar examples of scientific epistemology. While the use of complex simulations in science is a development that changes much in how science is done in practice, we argue that the concepts being applied fit very much into traditional practices of the scientific method, albeit those more often associated with laboratory work.

Keywords Climate models · Complex simulation · Model skill

1 Introduction

The 2013 Nobel Prize for Chemistry was given to three pioneers of complex multi-scale chemistry simulations. The chair of the Prize Committee, in explaining the award stated that “Chemistry is an experimental science” but that “Theory is the new

G. A. Schmidt (✉)
NASA Goddard Institute for Space Studies, New York, NY, USA
e-mail: gavin.a.schmidt@nasa.gov

S. Sherwood
Climate Change Research Centre, University of New South Wales, Sydney, Australia
e-mail: s.sherwood@unsw.edu.au

Published online: 09 December 2014

experimentation" (Lidin 2013). This claim of novelty in the use of complex simulation in science is a common one even among Nobel prize winners. For instance, von Hayek (1974) contrasted the new need for complex simulation in economics to discover its emergent properties with the "simplicity" (in his view) of the physical sciences. In the broader literature, describing simulation as the "third pillar" of scientific enquiry (alongside theory and experimentation) is a commonplace (e.g. President's Information Technology Advisory Committee (PITAC) 2005):

> Computational science now constitutes what many call the third pillar of the scientific enterprise, a peer alongside theory and physical experimentation.

There are also many counter-claims that "Science only has two legs" (e.g. Vardi 2010):

> What has changed is the scale of computation. While once carried out by hand, computation has required over time more advanced machinery. Doing theory today requires highly sophisticated computational-science techniques carried out on cutting-edge high-performance computers.
>
> So science is still carried out as an ongoing interplay between theory and experimentation. The complexity of both, however, has increased to such a degree that they cannot be carried out without computation. There is no need, therefore, to attach new legs to science. It is doing fine with two legs. At the same time, computational thinking (a phrase coined by Jeannette Wing) thoroughly pervades both legs. Computation is the universal enabler of science, supporting both theory and experimentation. Today the two legs of science are thoroughly computational!

Others have described complex simulations as only engaging "the same old stew" of standard epistemological questions (Frigg and Reiss 2009). In response, Humphreys (2009) outlined four specific issues that, in his view, make complex simulations stand out from other kinds of science: 1) the difficulty in understanding why a simulation produces the results that it does from the theory it encodes ("epistemic opacity"), 2) the intertwining of semantic and syntactic variations in applying the model to its target system, 3) the time iterative nature of most simulations, and 4) the unavoidable difference between what is doable in principle, and what is doable in practice. Humphreys claims that each of these four issues are grounded in the idea that simulations require an epistemology in which human actions are no longer central. Specifically, he feels that simulations create a hybrid situation where human and computers share epistemic authority.

We aim to illuminate these issues with an examination of the practice of climate modelling. Climate is an important example of a complex system where large-scale simulations are used to assess the implications of a set of known or plausible relationships governing the system (Heymann 2010; Lloyd 2010; Parker 2013a). Such simulations can reveal a range of emergent properties, responses of the system to altered initial or boundary conditions, or characteristics of the internal dynamics of the system itself. As described eloquently in Edwards (2010), simulation has become pervasive in the technical literature and public discussion of climate science. The reasons for this prominence are both practical and rational. First, climate models, while imperfect, work well in many respects (that is to say, they provide useful skill over and above simpler methods for making predictions). Second, many vital and interesting questions are intractable without recourse to comprehensive simulations. Despite this, many outside (and even some within) the climate science community are unaware of how climate models are built, used, and evaluated,[1] and there is a lot of naïve commentary on their utility for predicting future or attributing past climate change. This is unfortunate, as these matters must be understood in order to resolve practical questions such as to how climate models should be used to learn about the real world (e.g. Frame et al. 2007). This issue is only becoming more acute as more models are developed and more simulations are being performed (Taylor et al. 2012).

In this paper, we will explore specifically to what extent complex simulation in climate science is a new "pillar" of inquiry as opposed to an expansion or cross-fertilization of existing notions of theory and experiment (Winsberg 2003; Lloyd 2010) from the point of view of model developers and users of the simulations. We also discuss classical scientific notions such as falsifiability, confirmation and reproducibility in the climate modelling context. These concepts have been widely challenged in more general terms, but the practice of climate modelling perhaps provides some novel examples of why naïve application of these notions is problematic while more nuanced variations are still applicable.

At least one reason we need models is very clear (Knutson and Tuleya 2005):

> If we had observations of the future, we obviously would trust them more than models, but unfortunately observations of the future are not available at this time.

Given this requirement, there is a need for methodologies to assess the utility and credibility of model projections and determine how to use them appropriately in making predictions (Lloyd 2010; Katzav et al. 2012; Katzav 2014). Additionally, scientific work, for example the development of heuristic models or understanding of the climate system, also frequently benefits from the predictions by complex models of variables that cannot be observed due to technical limitations. Finally, models can be used to address counterfactual questions, such as what today's climate would have been like without human interference; such uses enable conclusions to be drawn about the causes (or change in probability) of events that could not be inferred from observations alone.

In the following we will restrict discussion to the class of model most often described as a "General Circulation Model," or GCM. The methods by which climate models (GCMs) are developed, evaluated and used influence the methodology by which useful predictions can be made; these methods are briefly described in Section 2. Section 3 describes the process of evaluating a single model, and Section 4, the added utility of multi-model ensembles. The final section attempts to relate some issues in the actual practice of climate modelling to current discussions found in the philosophy of science literature.

[1] We use the term 'evaluated' in this paper (for assessing the adequacy of the model for a certain task) over 'validated' since the latter term has connotations of 'truth', which is of limited relevance here.

Following conventions in the climate sciences, we use the term "model" to refer to the actual computer code or, equivalently, the (discretised) equations and assumptions it encodes; "simulation" to refer to a single run of the code (within a particular computational environment, and with specific initial and boundary conditions); and "projection" to indicate a prediction of a future trajectory of the climate that is contingent upon a specific scenario of future boundary conditions. A model simulation consists of running the model itself combined with some initial conditions and a set of drivers (boundary conditions or externally specified fields) that are either transient (changing in time) or static (yielding a "time-slice experiment"). These initial conditions and drivers have the role of auxiliary hypotheses in the Duhem/Quine sense (Stanford 2013) since they are themselves uncertain approximations to the real world conditions. Evaluation of a simulation (or often, an ensemble of simulations) is therefore a test of all the components (model physics, parameterisations, and auxiliaries) at once (Lenhard and Winsberg 2010).

2 Climate model development

We distinguish three classes of components of a climate model: well-accepted first principles (conservation of mass, energy etc.), approximations to well-understood physics, and empirical, phenomenological, "parameterisations" of unresolved processes. The need for approximations and parameterisations arises from limits on computer resources, and the impossibility of capturing all scales and phenomena of interest (from the planetary to that of an individual photon) on a quasi-uniform grid of points. The essential difference between these two classes is that, for approximations, an exact (or very accurate) theory is available, but would be far too computationally expensive to fully incorporate into the climate model; examples include atmospheric radiative transfer and the dynamical equations for fluid flow. For empirical parameterisations, needed in particular for turbulent processes, no adequate general theory exists.

The importance of parameterisations in climate modelling arises from the evident sensitivity of model behaviour at larger scales to the way that sub-gridscale phenomena (such as cloud droplet formation or surface turbulence) are represented. Parameterisations are also sometimes needed to represent the net effect of a process intentionally omitted from a model—for example the input of stratospheric water vapour via the oxidation of methane in models that don't include atmospheric chemistry. The use of parameterisations implies that climate models are not approximations to an accurately known but analytically intractable "theory of climate," but rather are distinct, alternative, composite hypotheses about climate that depend fundamentally on the specific resolution (the spatial or temporal scales at which proceses are truncated) and scope.

Model development consists mainly of improving the fidelity and/or computational efficiency of approximations, increasing the scope of models and processes they incorporate, and seeking more successful parameterisations (or indeed, replacing parameterisations with explicit representations). Atmospheric convection and cloud-related parameterisations continue to command scientists' attention and all extant examples still have substantial room for improvement (Knutti et al. 2013). Recent increases in climate model scope include the explicit prediction of aerosols and atmospheric chemistry, dynamic vegetation, and more complete representations of the carbon cycle. Model processes are sometimes made more elaborate so as to include previously ignored effects, such as mesoscale circulations (those larger than a cloud but smaller than a synoptic weather system or model grid cell) or microphysical effects of aerosols on cumulus clouds. The suitability of approximations (e.g. radiative transfer) can be tested "off-line" in situations where (nearly) exact solutions are available. Unfortunately this type of test is generally impossible or inconclusive for parameterisations, for which the principal test is typically the emergent performance of the climate model in which they are embedded.

An often overlooked aspect of model development is the work needed to produce inputs or boundary conditions—such as aerosol emissions, land characteristics and orography, and solar variations—and to produce datasets for diagnostic evaluations. These are constantly updated (as more data comes in, errors are corrected, and assumptions employed in their construction reevaluated). Simulation errors are sometimes due to poorly specified inputs or boundary conditions rather than errors in the model physics itself, and this may be especially relevant in paleo-climate simulations.

Once put together, a climate model typically has a handful of loosely-constrained parameters that can in practice be used to calibrate a few key emergent properties of the resulting simulations. In principle there may be a large number of such parameters that could potentially be tuned if one wanted to compare a very large ensemble of simulations (e.g. Stainforth et al. 2005), but this cumbersome exercise is rarely done operationally. The tuning or calibration effort seeks to minimise errors in key properties which would usually include the top-of-the-atmosphere radiative balance, mean surface temperature, and/or mean zonal wind speeds in the main atmospheric jets (Schmidt et al. 2014b; Mauritsen et al. 2012). In our experience however tuning parameters provide remarkably little leverage in improving overall model skill once a reasonable part of parameter space has been identified. Improvements in one field are usually accompanied by degradation in others, and the final choice of parameter involves judgments about the relative importance of different aspects of the simulations (for example, Australia uses a version of the UK Met Office atmosphere model but has made small modifications to mitigate problems in the Tropics and southern hemisphere that affect Australian forecasts, at the possible expense of performance in the UK Bi et al. 2013). Many recognizable aspects of model simulations remain similar across model generations, indicating that these aspects are robust to changes in the model including how it is calibrated (Masson and Knutti 2011). Specifically, many of the most persistent model biases are surprisingly resistent to tuning.

The decisions on what to tune, and especially what to tune for, involve value judgments (see also K. Intemann, "Values in Climate Models: The Good, The Bad, and the Ugly", submitted). This is notably more acute for complex simulations than it would be for a numerical calculation of the consequences of a well-specified theory, which has far fewer degrees of freedom in how such a calculation should be done.

There are claims that "social" values associated with the inductive risks associated with errors in climate model applications must play a role in model development (Winsberg 2012; Douglas 2000). However, this has been persuasively challenged by

Parker (2013b), who correctly points out that just because judgments are made does not mean they are related to specific outcomes in applications or the likelihoods of under or over-estimating sensitivities. Other authors (Betz 2013) have gone to the opposite extreme in claiming that model construction can be ’value free’ at least in a limited sense. Our experience is that values that are not purely epistemological do play a role in model development, but that they are not the social values associated with risk preferences that Winsberg or Douglas discuss. Instead these values are usually either aesthetic (e.g. elegance) or practical (tractability/ease of implementation, alignment with group research priorities).

Winsberg and Douglas suggest that, if given a choice between two different parameterisations with similar overall contributions to model skill, a modeller might choose one that yields predictions that might err in a specific way, for instance, producing under- or over-estimates of some feature matching their prior preference for avoiding false negatives over false positives. However, there are several reasons why such suggestions can be dismissed. First, it would require enormous study and expense to work out how to configure a model to produce preferred predictions (or predictions with a preferred sign of error), with little evident benefit to the modelling group, and enormous risks to their credibility if it became known that they had done so (which would be hard to avoid given the expense). The status of climate simulations in the public policy environment is already highly contested, and scientific credibility of the climate model development process is a frequent topic of debate. Second, models are now used to predict so many things under a wide range of applications that it is hard to imagine a modelling group or individual modeller settling on one particular outcome (or bias) to aim for, given the (indeterminate) impact this would inevitably have on other predictions. Third and most importantly, this activity would interfere and compete with the far stronger imperative to achieve the best skill possible against observable metrics in today’s (or past) climate, based on modelling decisions that are scientifically and practically defensible. Specifically, the effort required to assess the varying sensitivity of the model to multiple drivers as a function of the two paramterisations would be much better spent varying the parameterisations more finely and assessing improvements to climatological skill.

There is, on the other hand, a danger in groups “overfitting” their models to known climate changes in historical times. Empirically, there is a suggestion from 20th Century simulations performed by different groups for CMIP3 that choices of imposed aerosol radiative effects compensated for their different climate sensitivities, minimizing the range of temperature trends (Kiehl 2007), though this pattern is less apparent in CMIP5 (Knutti 2008; Forster et al. 2013). Estimates of historical changes additionally are not guaranteed to be stable over time, and so tuning to their variability would imply fitting to some non-climatic artifact. An example might be the trends in ocean heat content first assessed by Levitus et al. (2000), but signficantly revised in more recent compilations after bias corrections (Church et al. 2011). Nonetheless, the minor latitude that modellers have in selecting boundary conditions requires some consideration and may need to be taken into account to avoid overweighting success of the “model package”.

Tuning is perhaps best examined through an example. Biases in sea ice simulations are often very different in the Arctic and the Antarctic regions. This occurs because of the different environments (closed basin vs. open ocean, convergent flow vs. divergent flow, dominance of snow accumulation vs. basal ice formation, etc.) that make certain aspects of the simulation differently sensitive to the ocean heat flux or atmospheric circulation errors in each hemisphere. Despite the temptation to independently tune each hemisphere (by having regionally dependent parameters), this is not generally done since enforcing the global coherence of physical processes is a strong imperative. Parameters are instead chosen to do the best collective job across both hemispheres. Furthermore, there is a physical understanding and empirical evidence that the sensitivity of the Arctic sea ice is a strong function of the control run climatology (that is to say that a simulation that starts off with less ice than average will be more sensitive to perturbations than one with more). This implies that attempts to tune for a specific response may well negatively affect the control climate, which may lead to the simulation being too unrealistic to be weighted strongly in any projection (Massonnet et al. 2012).

Nonetheless, judgments must be made in the development process. Because of the limited personnel and time available, different groups tend to prioritize different aspects of the model simulations. Thus a modelling group with a strong interest in Arctic sea ice might set parameters to get the best possible estimate of the seasonal sea ice extent, while another might instead try to maximise the fidelity of the El-Niño/Southern Oscillation (ENSO) and its regional impacts. Given the expense and difficulty in running multiple transient simulations, it is rare for any tuning to be done to match trends or transient responses to external stimuli (like volcanic eruptions for instance). Indeed, some modelling groups (Schmidt et al. 2014b) eschew tuning to trends altogether in order to retain the possibility of using trends as an evaluation of model skill.

Arctic sea ice trends provide an instructive example. The hindcast estimates of recent trends were much improved in CMIP5 compared to CMIP3 (Stroeve et al. 2012). This is very likely because the observation/model mismatch in trends in CMIP3 (Stroeve et al. 2007) lead developers to re-examine the physics and code related to Arctic sea ice to identify missing processes or numerical problems (for instance, as described in Schmidt et al. 2014b). An alternate suggestion that model groups specifically tuned for trends in Arctic sea ice at the expense of global mean temperatures (Swanson 2013) is not in accord with the practice of any of the modelling groups with which we are familiar, and would be unlikely to work as discussed above.

3 Evaluation

Models are evaluated by several different types of practitioner, each focusing on somewhat different simulation characteristics. For a model developer, evaluation is important for assessing targets and strategies for model improvement. Thus, a characteristic that happens to be strongly controlled by a specific, uncertain model process may be particularly valuable regardless of its direct societal relevance (see Kim et al. 2012 for an example associated with tropical intraseasonal variability). For a model user, on the other hand, evaluation is most often concerned with determining model adequacy for a particular scientific question, for example prediction in a particular region, or of a particular phenomenon. Such users may therefore care whether the model's present-day climatology and variability are accurate in that region, or for that phenomenon, on the basis of an assumption that if behaviour there is poorly simulated today then simulations of local change are also less trustworthy. Users are most often focused on assessing emergent properties that have some real-world consequence or impact.

Common general evaluations often use well-characterised global climatologies (relatively long term averages) for temperature, precipitation, humidity etc. Reichler and Kim (2008) and Knutti et al. (2013). These are all emergent properties—they depend on the interaction of (in principle) all parts of the model and auxiliary hypotheses. Thus while model-observation mismatches can be clear (for instance ocean temperatures too warm off the coast of Peru), and the reasons understood at some level (e.g., insufficient marine stratus cloud in coastal upwelling zones), it is often not clear how to change the model to improve the situation. Although these errors may downgrade our assessed reliability of the model's predicted change in the region, we don't really know ahead of time whether, or how, whatever is causing the biases in the region will also affect the response to a change in boundary conditions or forcings. In practice, this needs to be examined by looking at multiple simulations (across multiple models or versions) to see whether there is any substantial relationship between the bias and the sensitivity of any particular feature (for instance as in Hall and Qu (2006)).

Poor skill in any of these metrics does not lead modellers to abandon the models, but rather to search for missing physics (such as the role of heterogenous chemistry on polar stratospheric clouds), improvements to misspecified parameterisations, and—perhaps most importantly—particular aspects of the complex system for which our current understanding does appear sufficient to make confident explanations and predictions.

Several factors further complicate model evaluation. One is that the most important variables in a model are often not well observed. For example, models have long been thought to overpredict global mean precipitation but newer observational estimates are much closer to that simulated (Stephens et al. 2012). When comparing satellite records to simulations, it is often necessary to embed code to simulate what a satellite would directly measure according to the model, rather than trying to infer climate variables from satellite radiances (Cesana and Chepfer 2012; Webb et al. 2001; Bodas-Salcedo et al. 2011). A similar situation holds for paleo-climate proxies such as water isotopes (Schmidt et al. 2007) or tree rings (Evans et al. 2013).

A second complicating factor is the chaotic nature of the real climate system. In all climate models the specific transient evolution of the simulation is sensitively dependent to tiny variations in the initial conditions (Deser et al. 2012). An ensemble of simulations from a model is typically necessary to determine whether an observation is consistent with the model spread or not, and even then, a singular result could always be related to an extreme (but unsampled) outlier. A free-running coupled climate model simulation initialised with 1850 conditions should not be thought deficient for failing to produce a record magnitude El Niño event in 1997/98, even though the model-observation mismatch might be large. If an ensemble of observed cases is also available (say for weather forecasting), a more precise evaluation of the coherence of the two ensembles becomes possible. Unfortunately for some of the most interesting predictions, such as centennial-scale trends, only a single realization of the real world may have been adequately observed.

It is incumbent upon those who develop models to know how they have (and have not) been tuned, in order to avoid inappropriate conclusions from successful tests, though the literature has historically been a little opaque on this topic. Though perhaps an obvious point, characteristics (or metrics) that are used to explicitly tune a model or its inputs should not also be used to evaluating the model - this would be a form of ’double counting’. A recent paper argued the opposite, that in fact some kinds of ’double counting’ are both permissible and practiced (Steele and Werndl 2013). On closer inspection though, both examples of ’double counting’ addressed in that paper are simple versions of parameter tuning or model selection with no evaluation beyond the fitting procedure. The authors describe this as ’relative confirmation’ (among models), but in our opinion that is irrelevant to assessments of model predictive skill which is the point that we are concerned with here. Specifically, Steel and Werndl ignore the fact that results that are predicted “out-of-sample” demonstrate more useful skill than results that are tuned for (or accommodated).

In some circumstances, the inability of a model to match an observed feature, *despite* extensive efforts to tune for it, does imply that the model is deficient and this can often be useful information limiting inferences to the real world, or providing targets for model development.

The most important measure of model skill is of course its ability to predict previously unmeasured (or unnoticed) phenomena or connections in ways that are more accurate than some simpler heuristic. Many examples exist, from straightforward predictions (ahead of time) of the likely impact of the Pinatubo eruption (Hansen et al. 1992), the skillful projection of the last three decades of warming (Hansen et al. 1988; Hargreaves 2010) and correctly predicting the resolution of disagreements between different sources of observation data e.g., between ocean and land temperature reconstructions in the last glacial period (Rind and Peteet 1985), or the satellite and surface temperature records in the 1990s (Mears et al. 2003; Thorne et al. 2011). Against this must be balanced predictions that did not match subsequent observations—for instance the underestimate of the rate of Arctic sea ice loss in CMIP3 (Stroeve et al. 2007).

In some cases, research groups using individual models have made surprising predictions, for example that global warming would not diminish Antarctic sea ice in the short term (Manabe et al. 1992), or that global-mean surface temperatures would cool temporarily during the last decade despite continued heat buildup in the system (Keenlyside et al. 2008). The first of these surprising predictions has been borne out, while the second was an over-prediction of what turned out to be a reduction in the mean surface warming rate rather than a reversal. These were not robust predictions across multiple models and it remains unclear as to whether these predictions were based on the 'right' reasons, so it cannot be claimed that the community at large foresaw these things, but they show the ability of models to explore unexpected possibilities.

The use of the models as explanatory tools for observed phenomena (whether it is the variance of tropical Pacific temperatures, jet stream variability or oscillations in the ocean circulation, attribution of historical changes) is more common than explicit prediction, though there are often implicit predictions associated with these results.

Up until now we have mainly discussed the use and evaluation of single models and much of the work of specific model groups is devoted to this. However, it is the emergence of a rich and complex coordinated set of multi-model simulations that greatly expands the scope for model evaluation and inferences, and we now turn our attention to this multi-model ensemble.

4 The multi-model ensemble

Over the last two decades, the development of large-scale model intercomparison projects involving nearly all climate modelling groups has generated an ever more comprehensive database for model output, which has become the dominant source of data for scientific uses and model evaluation and assessment. Starting with the Atmospheric Model Intercomparison Project (AMIP) (Gates et al. 1999), the program has expanded enormously to the current Coupled Model Intercomparison Project, Phase 5 (CMIP5) (Taylor et al. 2012). The scope of the models, the number of simulations per model, the number of requested diagnostics, the sampling frequency (e.g., daily) of simulations, and the typical model grid size have all combined to increase the size of the archive by orders of magnitude. While CMIP3 (initiated in 2004) had an archive of around 50 Terabytes, CMIP5 is expected to produce 3-10 Petabytes of distributed data from over 30 specific experiments, using over 60 models, from 29 model groups from 14 countries.

This multi-model ensemble (MME) of opportunity provides an immense resource for climate model evaluation. There are opportunities to assess the structural uncertainty of model predictions, to identify interesting and potentially informative patterns of behaviour across models, to tie future projections to past skill in out-of-sample paleo-climate tests with the same models, and to assess the impact of specific model characteristics (e.g., scope, grid size) on specific aspects of behaviour and fidelity. Many hundreds of papers looking at the CMIP5 results have already been published.

However, the MME was not designed with a particular focus, and consequently the variations in structure are somewhat haphazard (following the preferences of individual groups rather than any pre-determined plan) (Knutti et al. 2010). The histogram of any particular prediction or result from this set of models should therefore not be interpreted as any kind of probability density function (van Oldenborgh et al. 2013; Allen et al. 2013), although its breadth offers qualitative guidance if interpreted carefully. Curiously, the multi-model mean is often a more skillful predictor on average across a suite of quantities than is any individual model (Reichler and Kim 2008), though the effect soon saturates (Tebaldi and Knutti 2007). Systematic biases across all models are clearly present that cannot be removed via averaging (a good example is the prevalence of the biases in the Inter-tropical convergence zone (ITCZ), where two bands of equatorial rainfall are often simulated instead of a single one — the so-called 'double ITCZ' problem Hwang and Frierson 2013).

A more natural framing in our view for using the MME to make inferences about the real world or refine predictions is a Bayesian one (Jaynes 2003), where prior expectations are informally adjusted by model results after accounting for their skill, scope, or other biases.[2] Poorly resolved features in the models for which there is little demonstrated skill will not shift the posterior probabilities much (if at all), while well-modelled, skillful elements can affect the final predictions more heavily (Rougier 2007). In practice this is a challenge: what should be used for the prior expectation? What diagnostics are appropriate to use in any weighting scheme? How can one do the complex integrations over parameter space formally? (See Frame et al. (2007) for some discussion of the practicalities).

More generally, in response to these challenges and as outlined by Betz (2013), the increasingly dominant way to convey this information (for instance in the Fifth Assessment report from the Intergovernmental Panel on Climate Change (IPCC AR5) Stocker et al. 2013) has been to supplement key conclusions or predictions with a confidence statement reflecting the expert judgment of how strong is the evidence that underlies it (Cubasch et al. 2013). Simulation results, for instance, related to tropical cyclone intensity changes, are given as numerical indications, but with low levels of confidence, reflecting the lack of supporting evidence for skill in modelling changes for this metric in this class of model. In our view, this is effectively an 'informal' application of the concepts of Bayesian updating but with no actual calculation of the likelihoods or posterior probability.[3]

One helpful development for building confidence is the inclusion of paleo-climate simulations in the CMIP5 ensemble (Schmidt et al. 2014a). This is the first time that a coherent set of 'out-of-sample' simulations have been included in CMIP. The three targeted time periods—Last Glacial Maximum, mid-Holocene and last millennium—can safely be described as 'out-of-sample' because (due to lack of time and/or interest) we do not believe any models used have been tuned to get better matches to paleo data, making these truly independent test cases. Furthermore, global and regional climate changes during the first two of these periods were significantly larger than those during the modern instrumental period, and some of them are commensurate with predicted changes for the rest of the 21st Century. One therefore might expect skill in reproducing these changes to be at least as relevant to projections than that for the smaller changes seen in the modern record. This must be set against the greater uncertainty of paleoclimate states, but the signal-to-noise ratio can be as high as in modern data (Köhler et al. 2010).

[2] Alternative framings are discussed in detail in Katzav (2014).

[3] Katzav (2014) outlines 5 'views' of climate model assessement, and this approach is akin to a combination of his description of the adequacy-for-process and 'conservative' views.

There are at least two classes of inference that can be made via these out-of-sample tests: verification of robust prediction characteristics (such as the ocean to land temperature change ratio, the connection between the cross equatorial gradient in ocean temperature and the position of the inter-tropical convergence zone), and discrimination among non-robust projection characteristics that appear to be related to testable paleo-climate predictions. Both examples are explored in Schmidt et al. (2014a).

IPCC AR5 presented a preliminary assessment of the CMIP5 simulations, and future projections from these models (Flato et al. 2013; Collins et al. 2013). While assessments presented there were typically based on comparing an observation to the mean of all available models, two novelties in the presentation of the results are worth noting. First, there was an extensive discussion of how to present the spread among the simulations; multiple methods were suggested to highlight the different possible combinations of large/small mean signals and high/low model agreement. Second, there was a (hesitant) move towards weighting model projections in certain cases. The most prominent example is in the discussion of Arctic sea ice extent, where a crude weighting based on having a modern day seasonal cycle 'close' to observed was used to exclude outliers (Massonnet et al. 2012). This was justified through the acknowledged sensitivity of the Arctic system to the base climate state; however, it does not account for the possible differences in tuning among groups given the known target. In almost all other projections 'model democracy' reigns, i.e., each simulation or ensemble average is weighted equally. This is unlikely to be optimal, but deciding on an objective basis how to weight models is problematic and is only starting to be fully explored (Knutti et al. 2010). Assessments currently therefore tend to reflect that lack of methodological consensus, though this might change in the future.

In general we are more confident in predictions that are more consistent across models. More precisely, we are more confident that the results are a consequence of the common underlying assumptions embedded within each model. For those where there is both consistency among models and observational support from past changes (e.g., land temperature warming, increases in high latitude precipitation, Arctic amplification of temperature changes), the result can be viewed with "high confidence," and we would also conclude that the processes that determine this result are encapsulated in our models.

For many important questions however, for example how atmospheric humidity or cloud cover respond to a given change in global temperature, no past observational tests are available that could directly confirm model predictions. One important way for scientists to develop confidence in such situations is to construct a simpler model for the predicted phenomenon, which shows that it follows from a few well-known physical principles rather than depending on poorly understood processes. Indeed, much of the value of complex models in science (as opposed to practical application) involves their use to test more heuristic models, physical understanding, or assumptions that have been made in inferring general results from specific observations; it is these approaches from which genuine scientific confidence emerges (Held 2005). Indeed much of our understanding of global climate, including the basics of what controls global temperature, was worked out long before GCMs were developed.

In any case “high confidence” is not a guarantee that a conclusion or projection is correct. The problem of “unknown unknowns” (Rumsfeld 2002) is clearly recognised since models are incomplete and may share common errors or important omissions — the polar ozone hole is a classic example of this in the Earth Science field (Farman et al. 1985). There is therefore a widespread expectation of ’inevitable surprises’ from very large climate changes should they occur (National Research Council Committee on Abrupt Climate Change 2002).

At the opposite extreme, when projections from different models disagree substantially, we infer that the result is either sensitive to one of the many non-robust model elements or is inherently unpredictable. A suitably designed ensemble of runs with a single model can distinguish between these reasons: for example, the simulated rainfall trend in a specific continental location over a twenty year period in a single ensemble member is not predictable even though we have perfect knowledge of the model that generated it (Deser et al. 2012). Given that the real world is more complex, we infer that the equivalent real-world phenomenon is not likely to be predictable either. Regardless of the reason, projections are of limited use in such cases. In cases where model projection divergence can be explained via differences in scope or model skill, the level of confidence in a weighted projection that took that into account might be higher than for the multi-model mean.

Finally, an increasingly important issue is that it is currently impractical to calculate all of the process-based diagnostics that could be envisaged to assess models. These include multi-variate characterisations of storms, overshooting deep convection or ocean-ice interactions in the North Atlantic. This effort is precluded by the very size of the CMIP5 MME dataset and the lack of data analysis infrastructure that can analyse something this large in flexible ways. Thus while a great deal of desired information is available in principle, it is in practice inaccessible, and hence unknown. This is in part a resource allocation issue, with more resources devoted to gathering observational data and generating model results than bringing the two together.

5 Practical climate model philosophies

We have discussed how climate models are typically built, evaluated, and used for prediction, as well as how they can serve as an important tool in developing deeper scientific understanding. It is clear that climate models are ’messy’. That is, while they represent a quantification of a complex basic theory combined with necessary engineering ’kludges’ to make the models work (Lenhard and Winsberg 2010), they also resemble a laboratory apparatus that requires calibration and replication to produce useful results and deeper understanding.

We began this article referring to a number of philosophical questions related to the nature of complex simulations, to which we now return. We deal first with the classical scientific issues of falsifiability and confirmation.

No-one using a climate model should have any illusion that such a model can be “true” in any pure sense. All climate models are indeed wrong (Box 1979) and yet models can be useful in the sense that their predictions are often demonstrably more skillful than any simpler alternative. If however we assume that a strict Popperian would discard a ’falsified’ model, the fact that models will always disagree with observations if examined closely enough, would leave no valid scientific approach available, regardless of skill, since *all* theories of the climate system would be falsified by this standard. A naïve positivism, such as could be attributed to Feynman (1965):

> ...we compare the result of the computation to nature, with experiment or experience, compare it directly with observation, to see if it works. If it disagrees with experiment, it is wrong.

is thus over-simplistic for the study of very complex systems, which would lead (for our strict Popperian) to an immediate and permanent dead end. The concept of falsification can however be rescued by applying it solely to empirical predictions that are demonstrated to be unskillful (i.e. no better than a naive baseline prediction), rather than to models as a whole (Lloyd 1987). Similarly it follows that any specific climate model as a whole cannot be ’confirmed’ (according to a black-and-white definition). Rather individual empirical predictions can be (within some uncertainty). We stress that the concept of skill relative to a naïve baseline is a much more useful frame in which to place model predictions, though this is analogous to confirmation if it is understood to be a matter of degree (i.e. evidence increasing the trustworthiness of a prediction).

Next we look at reproducibility, which is widely taken to be a hallmark of any robust scientific claim. A theoretical prediction should clearly be reproducible at whatever precision is required, since it should be exactly specified by the theory assuming it is calculated correctly (which is of course worth confirming by independent checks). Reproducibility for an experimental result is of a different kind: reproduction is not expected to be exact, given differences in apparatus, operators, environmental conditions etc., but should occur within some estimated level of precision if the measurements are capturing some underlying process in the real world. An individual climate model simulation is almost always ’bit-reproducible’ i.e. given the exact code, compiler, mathematical library and pseudo-random number seed, the exact same numerical series of results can be reproduced. However, in practice the simulations provided to the CMIP5 archive cannot be exactly replicated at other institutions, or even at the same institution after an upgrade of equipment or software. Repeat simulations should certainly be consistent, however, with respect to any important characteristics from which inferences about the real world might be drawn. Modelling centres typically provide benchmark tests with their models that can be compared to others to ensure that their simulations are reproducing—in the laboratory sense—those obtained by the developers of the model (occasionally this fails owing, for example, to errors or bugs whose impacts depend on the computing environment). The reproducibility of inferences drawn from climate models therefore has more in common with laboratory experimentation than with the exact repeatability of a theoretical calculation. On this issue we differ slightly from Frigg and Reiss (2009) and feel justified in describing simulation as being “in-between” theory and experimentation (i.e. having characteristics of both).

We turn now to Humphreys' claims for the philosophical novelty of simulations (Humphreys 2009). His claims are based on an overall principle that science with large simulations "uses methods that push humans away from the epistimological centre" and cedes some epistemological authority to the simulation (and its emergent properties) as opposed to the scientists. However, in our view, no result, or emergent feature, is accepted without human evaluation and assessment which, as we described above, can encompass may kinds of tests and comparisons. In each of the stages discussed above—development, evaluation and inference from a single model or multimodel ensemble— humans are at the center of the epistemological decisions, in similar ways as they would be in simply observing or experimenting on a natural complex system. Expert judgments on these results are made based on exisiting knowledge from observations or more fundamental understanding. Indeed, we argue that the complex systems created in climate modelling are easier to control and understand (due to the larger range of model manipulations that are possible) than natural complex systems. Thus the claim that this necessarily leads to a new situation appears to us to be weak.

However, some of Humphreys' specific points are of interest. His first claim is that results from complex simulations are "epistemologically opaque" — specifically that we don't immediately know how any (emergent) result arose — unlike for some more tractable theoretical results. We agree that this is often true for climate simulations, as discussed above, but there is an analogous opacity in laboratory experiments involving complex systems for which no comprehensive theory exists. For instance, the Briggs-Rauscher chemical reaction (Briggs and Rauscher 1973) shows emergent oscillatory behaviour. The initial results (first observed in classroom laboratory demonstration) were epistemologically opaque, but easily reproducible. In both computational and analogous laboratory cases one can in principle (and sometimes in practice) break down the reasons why any change has occurred by using simpler (heuristic) models and performing additional tests to increase understanding, indeed, this is often easier in a computational experiment. In the case of the Briggs-Rauscher reaction, it took nearly a decade for a comprehensive mechanism for the emergent oscillations to be deduced (Noyes and Furrow 1982). Thus while epistemological opacity may be a novel aspect of theory application, it does not seem to us to be a novel issue within a broader view of scientific epistemology.

His second claim relates to the semantics of theories. Given a general theory that needs to be applied to a specific situation, realistic simulations requires various approximations, discretisations, and parameterisations of unresolved processes (as discussed above). Humphreys notes that implementing these aspects requires attention to syntax, since some implementations will be more amenable to solution than others. However, we suggest that any specific climate model (including the relative approximations) actually functions as an individual theory of climate, leaving differences in possible syntax as irrelevant to the predictions of the theory. Differences in syntax in simulations that exhibit chaos (specifically extreme sensitivity to initial conditions) will likely produce different specific trajectories through phase space, but a difference in empirically skillful predictions is unlikely. This is because no prediction unique to any specific trajectory is going to be robust. Rather, robust predictions require either an ensemble spread or simulations long enough to average sufficiently over the chaotic dynamics. As outlined above, the presence of chaos radically affects how one uses a model, constrains model-observation comparisons, and demands the use of probabilistic or ensemble-based approaches. This phenomenology is indeed a novelty, as was recognised 50 years ago (Lorenz 1963), but is again not limited to complex simulations of the sort we are concerned with here.

Humphreys, Frigg and Reiss make a point to contrast situations where there is a skillful analytical solution to a problem and where the only solution arises via simulation. In contrast, we do not see this as fundamental. It is easy to envisage a multi-variate, multi-dimensional analytical solution that is so complex that it is impossible to fully visualize or understand without computational help, just as it is easy to find examples of simulated solutions to 'intractable' problems that exhibit very simple emergent behaviour. The advantages to an analytic solution decrease rapidly with increasing complexity.

More generally, it is worth noting that only highly idealised applications of theories of the real world admit analytic solutions. The pendulum examples used by Humphreys are clear idealisations of the real world. For any real pendulum (single, double or a multiple upside-down oscillating one), the addition of factors such as friction, air resistance, finite size, and turbulent air flow will impact their trajectory and, if included, will make their solution intractable and only accessible via simulation. Therefore focusing on the difference between tractable and intractable idealisations seems orthogonal to the issue of how one should confront general theories with real world data.

The time-dependence of dynamic solutions is the third issue highlighted by Humphreys. He, along with Parker (2013a), sees this as an essential component of complex simulation. However, while current large-scale climate models (without exception) use time-stepping to explore the sensitivities of the modelled climate system, it is conceivable that at some point (as with slightly simpler systems (Broer et al. 2002)), it may be possible to define the climate attractor(s) and their sensitivity to forcings directly without recourse to a time-stepped solution while using the same model code. For instance, both Sherwood (1999) and Schmidt and Mysak (1996) used Newton-Rapheson methods to identify equilibria in reduced complexity climate models that could optionally be run also in time-stepping mode. Time in such analyses is implicit, rather than explicit, but the issues raised above would all still hold.

Humphreys' final point revolves around the idea that it is no longer possible, when discussing complex simulations, to separate theoretical and practical limitations. For climate models, for instance, one cannot assume that time steps or spatial grid sizes can become as small as desired because of the impractically rapid increases in computation time as scales become smaller. Instead these limitations are an inherent part of the whole exercise: they account for why a climate model has the form it does (by defining the scale at which physics gets truncated and parameterisations are necessary). We agree that this is a key observation, but practical considerations also limit experimental science: measurements either in the laboratory or in the field cannot be made with infinite precision, and often must measure something other than what would be ideal (for example we detect planets around other stars not by observing them directly, which would be far more challenging, but by observing their impact on larger bodies). So again, while this is may be novel in terms of theory application, a broader view indicates that similar concerns have always existed in experimental science.

To conclude, while we do think that the rise of complex simulation in climate and other fields has important practical consequences, we are not convinced that it requires any wholesale adjustment to philosophical understandings of the scientific method. Rather, as should be clear from the previous sections, while the practice of climate model development and use is different from many other fields in science, it acts to help clarify previous ideas, rather than undermine them. We agree with many previous authors that experimentation with climate models resembles laboratory science more than calculations using known theories, and in that sense transcends a binary theory/experiment divide. But we find that the limitations in using climate models to describe and predict the real world simply make more obvious the equivalent limitations that any models of any real world systems have. Thus they mainly serve to illuminate classic problems of scientific epistemology —across theory and experimentation—rather than create new ones.

As with all tools, there can be both appropriate and inappropriate applications of climate models. Generating predictions without taking into account uncertainties associated with different models, parameterisations, or initial conditions, is not particularly useful and cannot be used to assess skill. Relatively arbitrary weighting strategies that are missing a demonstration of relevance to any specific outcome are commonplace, as are mistaken interpretations of the MME histogram as a probability distribution. But just as an attempt to use a hammer to fix a watch does not invalidate the use of the same hammer to drive home a nail, the best practices in climate modelling are not invalidated by the least useful.

Overall, the paradigm of understanding emergent properties of the complex system via the bottom-up agglomeration and interaction of small scale processes has become dominant in climate science. While it is conceivable that a top-down principle could be found that provides better predictions and thus supplants climate modelling, no such principle has yet been discovered and we think it unlikely that one will emerge. Meanwhile, climate models continue to increase in skill and scope (Knutti et al. 2013) and the challenge to intelligently marshall that resource in the most effective and rigorous way possible to better understand the world continues to grow.

Acknowledgments This paper has benefited greatly from extensive discussions with Wendy Parker and Joel Katzav, two anonymous reviews and from conversations at a workshop on climate model philosophy in Eindhoven in November 2013.

References

Allen, M.R., Mitchell, J.F.B., Stott, P.A. (2013). Test of a decadal climate forecast. *Nature Geoscience*, *6*, 243–244. doi:10.1038/ngeo1788.

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science*, *3*, 207–220. doi:10.1007/s13194-012-0062-x.

Bi, D., Dix, M., Marsland, S., O'Farrell, S., Rashid, H., Uotila, P., Hirst, A., Kowalczyk, E., Golebiewski, M., Sullivan, A., Yan, H., Hannah, N., Franklin, C., Sun, Z., Vohralik, P., Watterson, I., Zhou, X., Fiedler, R., Collier, M., Ma, Y., Noonan, J., Stevens, L., Uhe, P., Zhu, H., Griffies, S., Hill, R., Harris, C., Puri, K. (2013). The ACCESS coupled model: description, control climate and evaluation. *Australian Meteorological and Oceanographic Journal*, *63*, 41–64.

Bodas-Salcedo, A., Webb, M.J., Bony, S., Chepfer, H., Dufresne, J.L., Klein, S.A., Zhang, Y., Marchand, R., Haynes, J.M., Pincus, R., John, V.O. (2011). COSP satellite simulation software for model assessment. *Bulletin of the American Meteorological Society*, *92*(8), 1023–1043. doi:10.1175/2011BAMS2856.1.

Box, G. (1979). Robustness in the Strategy of Scientific Model Building. MRC technical summary report, University of Wisconsin–Madison, Mathematics Research Center, Defense Technical Information Center.

Briggs, T.S., & Rauscher, W.C. (1973). An oscillating iodine clock. *Journal of Chemical Education*, *50*, 496.

Broer, H., Simó, C., Vitolo, R. (2002). Bifurcations and strange attractors in the Lorenz-84 climate model with seasonal forcing. *Nonlinearity*, *15*, 1205–1268. doi:10.1088/0951-7715/15/4/312.

Cesana, G., & Chepfer, H. (2012). How well do climate models simulate cloud vertical structure? A comparison between CALIPSO-GOCCP satellite observations and CMIP5 models. *Geophysical Research Letters*, *39*, L20803. doi:10.1029/2012GL053153.

Church, J.A., White, N.J., Konikow, L.F., Domingues, C.M., Cogley, J.G., Rignot, E., Gregory, J.M., van den Broeke, M.R., Monaghan, A.J., Velicogna, I. (2011). Revisiting the Earth's sea-level and energy budgets from 1961 to 2008. *Geophysical Research Letters*, *38*, L18601. doi:10.1029/2011GL048794.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W.J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A.J., Wehner, M. (2013). Long-term climate change: Projections, commitments and irreversibility. In T.F. Stocker, D. Qin, G.K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P. Midgley (Eds.)*, Climate Change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Cubasch, U., Wuebbles, D., Chen, D., Facchini, M.C., Frame, D., Mahowald, N., Winther, J.G. (2013). Introduction. In T.F. Stocker, D. Qin, G.K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P. Midgley (Eds.)*,Climate Change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Deser, C., Knutti, R., Solomon, S., Phillips, A. (2012). Communication of the role of natural variability in future North American climate. *Nature Climate Change*, *2*, 775–779. doi:10.1038/nclimate1562.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, *67*, 559–579. doi:10.2307/188707.

Edwards, P.N. (2010). *A Vast Machine*. Cambridge, MA: MIT Press.

Evans, M.N., Tolwinski-Ward, S.E., Thompson, D.M., Anchukaitis, K.J. (2013). Applications of proxy system modeling in high resolution paleoclimatology. *Quaternary Science Reviews*, *76*, 16–28.

Farman, J.C., Gardiner, B.G., Shanklin, J.D. (1985). Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction. *Nature*, *315*, 207–210. doi:10.1038/315207a0.

Feynman, R.P. (1965). *The character of physical law: The 1964 Messenger Lectures*. Cambridge MA: MIT Press.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S.C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., Rummukaine, M. (2013). Evaluation of climate models. In T.F. Stocker, D. Qin, G.K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P. Midgley (Eds.)*, Climate Change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Forster, P.M., Andrews, T., Good, P., Gregory, J.M., Jackson, L.S., Zelinka, M. (2013). Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models. *Journal of Geophysical Research Atmospheres*, *118*, 1139–1150. doi:10.1002/jgrd.50174.

Frame, D.J., Faull, N.E., Joshi, M.M., Allen, M.R. (2007). Probabilistic climate forecasts and inductive problems. *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences*, *365*(1857), 1971–1992. doi:10.1098/rsta.2007.2069.

Frigg, R., & Reiss, J. (2009). The philosophy of simulation: hot new issues or same old stew? *Synthese*, *169*(3), 593–613. doi:10.1007/s11229-008-9438-z.

Gates, W.L., Boyle, J.S., Covey, C., Dease, C.G., Doutriaux, C.M., Drach, R.S., Fiorino, M., Gleckler, P.J., Hnilo, J.J., Marlais, S.M., Phillips, T.J., Potter, G.L., Santer, B.D., Sperber, K.R., Taylor, K.E., Williams, D.N. (1999). An overview of the results of the Atmospheric Model Intercomparison Project (AMIP1). *Bulletin of the American Meteorological Society*, *80*, 29–55.

Hall, A., & Qu, X. (2006). Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophysical Research Letters*, *33*. doi:10.1029/2005GL025,127,L03,502.

Hansen, J., Fung, I., Lacis, A., Rind, D., Lebedeff, S., Ruedy, R., Russell, G., Stone, P. (1988). Global climate changes as forecast by Goddard Institute for Space Studies three-dimensional model. *Journal of Geophysical Research*, *93*, 9341–9364.

Hansen, J., Lacis, A., Ruedy, R., Sato, M. (1992). Potential climate impact of Mount Pinatubo eruption. *Geophysical Research Letters*, *19*, 215–218.

Hargreaves, J.C. (2010). Skill and uncertainty in climate models. *Wiley Interdisciplinary Reviews: Climate Change*, *1*, 556–564.

von Hayek, F.A. (1974). Prize lecture: the pretence of knowledge. http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1974/hayek-lecture.html. Accessed 21 Oct 2013.

Held, I.M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, *86*, 1609–1614.

Heymann, M. (2010). Understanding and misunderstanding computer simulation: the case of atmospheric and climate science - an introduction. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *41*(3), 193–200. doi:10.1016/j.shpsb.2010.08.001.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, *169*(3), 615–626. doi:10.1007/s11229-008-9435-2.

Hwang, Y.T., & Frierson, D.M.W. (2013). Link between the double-Intertropical convergence zone problem and cloud biases over the Southern Ocean. *Proceedings of the National Academy of Sciences*, *110*, 4935–4940. doi:10.1073/pnas.1213302110.

Jaynes, E.T. (2003). *Probability theory: the logic of science*. Cambridge, 727 pp.

Katzav, J. (2014). The epistemology of climate models and some of its implications for climate science and the philosophy of science. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *46*, 228–238. doi:10.1016/j.shpsb.2014.03.001.

Katzav, J., Dijkstra, H.A., de Laat, A.T.J. (2012). Assessing climate model projections: state of the art and philosophical reflections. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *43*, 258–276. doi:10.1016/j.shpsb.2012.07.002.

Keenlyside, N.S., Latif, M., Jungclaus, J., Kornblueh, L., Roeckner, E. (2008). Advancing decadal-scale climate prediction in the north atlantic sector. *Nature*, *453*(7191), 84–88.

Kiehl, J.T. (2007). Twentieth century climate model response and climate sensitivity. *Geophysical Research Letters*, *34*, L22710. doi:10.1029/2007GL031383.

Kim, D., Sobel, A.H., Del Genio, A.D., Chen, Y.H., Camargo, S.J., Yao, M.S., Kelley, M., Nazarenko, L. (2012). The tropical subseasonal variability simulated in the NASA GISS general circulation model. *Journal of Climatology*, *25*, 4641–4659. doi:10.1175/JCLI-D-11-00447.1.

Knutson, T., & Tuleya, R.E. (2005). Reply. *Journal of Climate*, *18*, 5183–5187. doi:10.1175/JCLI3593.1.

Knutti, R. (2008). Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society A*, *366*, 4647–4664. doi:10.1098/rsta.2008.0169.

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

Knutti, R., Masson, D., Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, 40. doi:10.1002/grl.50256.

Köhler, P., Bintanja, R., Fischer, H., Joos, F., Knutti, R., Lohmann, G., Masson-Delmotte, V. (2010). What caused Earth’s temperature variations during the last 800,000 years? Data-based evidence on radiative forcing and constraints on climate sensitivity. *Quaternary Science Reviews*, *29*, 129–145. doi:10.1016/j.quascirev.2009.09.026.

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *41*, 253–262. doi:10.1016/j.shpsb.2010.07.001.

Levitus, S., Antonov, J.I., Boyer, T.P., Stephens, C. (2000). Warming of the world ocean. *Science*, *287*, 2225–2228. doi:10.1126/science.287.5461.2225.

Lidin, S. (2013). Interview: 2013 Nobel Prize in Chemistry announcement. http://www.youtube.com/watch?v=igfed5l66Qo. Accessed 21 Oct 2013.

Lloyd, E.A. (1987). Confirmation of ecological and evolutionary models. *Biology and Philosophy*, *2*, 277–293.

Lloyd, E.A. (2010). Confirmation and robustness of climate models. *Philosophy of Science*, *77*, 971–984.

Lorenz, E.N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*, 130–141. doi:10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2.

Manabe, S., Spelman, M.J., Stouffer, R.J. (1992). Transient responses of a coupled ocean-atmosphere model to gradual changes of atmospheric CO_2. Part II: Seasonal response. *Journal of Climate*, *5*, 105–126. doi:10.1175/1520-0442(1992)005⟨105:TROACO⟩2.0.CO;2.

Masson, D., & Knutti, R. (2011). Climate model genealogy. *Geophysical Research Letters*, *38*, L08703. doi:10.1029/2011GL046864.

Massonnet, F., Fichefet, T., Goosse, H., Bitz, C.M., Philippon-Berthier, G., Holland, M., Barriat, P.Y. (2012). Constraining projections of summer Arctic sea ice. *Cryosphere*, *6*, 1383–1394. doi:10.5194/tc-6-1383-2012.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., Mikolajewicz, U., Notz, D., Pincus, R., Schmidt, H., Tomassini, L. (2012). Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, *4*, M00A01. doi:10.1029/2012MS000154.

Mears, C.A., Schabel, M., Wentz, F.J. (2003). A reanalysis of the MSU Channel 2 tropospheric temperature record. *Journal of Climatology*, *16*, 3650–3664.

National Research Council Committee on Abrupt Climate Change (2002). Abrupt Climate Change: Inevitable Surprises. The National Academies Press. http://www.nap.edu/openbook.php?record_id=10136.

Noyes, R.M., & Furrow, S.D. (1982). The oscillatory Briggs-Rauscher reaction. 3. A skeleton mechanism for oscillations. *Journal of the American Chemical Society*, *104*, 45–48. doi:10.1021/ja00365a011.

Parker, W. (2013a). Computer simulation. In S. Psillos & M. Curd (Eds.), The Routledge Companion to Philosophy of Science, 2nd edn. Routledge.

Parker, W. (2013b). Values and uncertainties in climate prediction, revisited. Studies in History and Philosophy of Science Part A pp. doi:10.1016/j.shpsa.2013.11.003.

President's Information Technology Advisory Committee (PITAC) (2005). Report to the President, 2005, Computational Science: Ensuring Americas Competitiveness. http://www.csci.psu.edu/docs/computational.pdf. Accessed 21 Oct 2013.

Reichler, T., & Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, *89*, 303–311.

Rind, D.H., & Peteet, D. (1985). Terrestrial conditions at the last glacial maximum and CLIMAP sea surface temperature estimates: Are they consistent? *Quaternary Research*, *24*, 1–22.

Rougier, J.C. (2007). Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, *81*, 247–264. doi:10.1007/s10584-006-9156-9.

Rumsfeld, D. (2002). Dept. of Defense, News briefing, February, 12. http://www.defense.gov/transcripts/transcript.aspx?transcriptid=2636. Accessed 5 Dec 2012.

Schmidt, G.A., & Mysak, L.A. (1996). The stability of a zonally averaged thermohaline circulation model. *Tellus*, *48A*, 158–178.

Schmidt, G.A., LeGrande, A., Hoffmann, G. (2007). Water isotope expressions of intrinsic and forced variability in a coupled ocean-atmosphere model. *Journal of Geophysical Research*, *112*, D10103. doi:10.1029/2006JD007781.

Schmidt, G.A., Annan, J.D., Bartlein, P.J., Cook, B.I., Guilyardi, E., Hargreaves, J.C., Harrison, S.P., Kageyama, M., LeGrande, A.N., Konecky, B., Lovejoy, S., Mann, M.E., Masson-Delmotte, V., Risi, C., Thompson, D., Timmermann, A., Tremblay, L.B., Yiou, P. (2014a). Using palaeoclimate comparisons to constrain future projections in CMIP5. *Climate of the Past*, *10*, 221–250. doi:10.5194/cp-10-221-2014.

Schmidt, G.A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G.L., Aleinov, I., Bauer, M., Bauer, S., Bhat, M.K., Bleck, R., Canuto, V., Chen, Y., Cheng, Y., Clune, T.L., Del Genio, A., de Fainchtein, R., Faluvegi, G., Hansen, J.E., Healy, R.J., Kiang, N.Y., Koch, D., Lacis, A.A., LeGrande, A.N., Lerner, J., Lo, K.K., Matthews, E.E., Menon, S., Miller, R.L., Oinas, V., Oloso, A.O., Perlwitz, J., Puma, M.J., Putman, W.M., Rind, D., Romanou, A., Sato, M., Shindell, D.T., Sun, S., Syed, R., Tausnev, N., Tsigaridis, K., Unger, N., Voulgarakis, A., Yao, M.S., Zhang, J. (2014b). Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *Journal of Advances in Modeling Earth Systems*, *6*, 141–184. doi:10.1002/2013MS000265.

Sherwood, S.C. (1999). Feedbacks in a simple prognostic tropical climate model. *Journal of the Atmospheric Science*, *56*, 2178–2200. doi:10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2.

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

Stanford, K. (2013). Underdetermination of scientific theory. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Winter 2013 Edition)*. http://plato.stanford.edu/archives/win2013/entries/scientific-underdetermination/.

Steele, K., & Werndl, C. (2013). Climate models, calibration, and confirmation. *British Journal for the Philosophy of Science*, *64*, 609–635. doi:10.1093/bjps/axs036.

Stephens, G.L., Li, J., Wild, M., Clayson, C.A., Loeb, N., Kato, S., L'Ecuyer, T., Stackhouse, P.W., Lebsock, M., Andrews, T. (2012). An update on Earth's energy balance in light of the latest global observations. *Nature Geoscience*, *5*, 691–696. doi:10.1038/ngeo1580.

Stocker, T., Dahe, Q., Plattner G.K. (Eds.) (2013). *Climate Change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Stroeve, J., Holland, M.M., Meier, W., Scambos, T., Serreze, M. (2007). Arctic sea ice decline: Faster than forecast. *Geophysical Research Letters*, *34*, L09501. doi:10.1029/2007GL029703.

Stroeve, J.C., Kattsov, V., Barrett, A., Serreze, M., Pavlova, T., Holland, M., Meier, W.N. (2012). Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. *Geophysical Research Letters*, *39*, L16502. doi:10.1029/2012GL052676.

Swanson, K.L. (2013). Emerging selection bias in large-scale climate change simulations. *Geophysical Research Letters*, *40*, 3184–3188. doi:10.1002/grl.50562.

Taylor, K.E., Stouffer, R., Meehl, G. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*, 485–498. doi:10.1175/BAMS-D-11-00094.1.

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections in probabilistic climate projections. *Society*, *365*(1857), 2053–2075.

Thorne, P.W., Lanzante, J.R., Peterson, T.C., Seidel, D.J., Shine, K.P. (2011). Tropospheric temperature trends: History of an ongoing controversy. *WIREs Climate Change*, *2*, 66–88. doi:10.1002/wcc.80.

van Oldenborgh, G.J., Doblas Reyes, F.J., Drijfhout, S.S., Hawkins, E. (2013). Reliability of regional climate model trends. *Environmental Research Letters*, *8*(1), 014,055. doi:10.1088/1748-9326/8/1/014055.

Vardi, M.Y. (2010). Science has only two legs. *Communications of the ACM*, *53*, 5. doi:10.1145/1810891.1810892.

Webb, M., Senior, C., Bony, S., Morcrette, J.J. (2001). Combining ERBE and ISCCP data to assess clouds in the Hadley Centre, ECMWF and LMD atmospheric climate models. *Climate Dynamics*, *17*, 905–922.

Winsberg, E. (2003). Simulated experiments: Methodology for a virtual world. *Philosophy of Science*, *70*(1), 105–125. doi:10.1086/367872.

Winsberg, E. (2012). Values and uncertainties in the predictions of Global climate models. *Kennedy Institute of Ethics Journal*, *22*, 111–137. doi:10.1353/ken.2012.0008.