The epistemological status of general circulation models

Craig Loehle[1]

Received: 16 December 2016 / Accepted: 1 May 2017 / Published online: 10 May 2017

Abstract Forecasts of both likely anthropogenic effects on climate and consequent effects on nature and society are based on large, complex software tools called general circulation models (GCMs). Forecasts generated by GCMs have been used extensively in policy decisions related to climate change. However, the relation between underlying physical theories and results produced by GCMs is unclear. In the case of GCMs, many discretizations and approximations are made, and simulating Earth system processes is far from simple and currently leads to some results with unknown energy balance implications. Statistical testing of GCM forecasts for degree of agreement with data would facilitate assessment of fitness for use. If model results need to be put on an anomaly basis due to model bias, then both visual and quantitative measures of model fit depend strongly on the reference period used for normalization, making testing problematic. Epistemology is here applied to problems of statistical inference during testing, the relationship between the underlying physics and the models, the epistemic meaning of ensemble statistics, problems of spatial and temporal scale, the existence or not of an unforced null for climate fluctuations, the meaning of existing uncertainty estimates, and other issues. Rigorous reasoning entails carefully quantifying levels of uncertainty.

Keywords Climate change · General circulation models · Model testing · Epistemology

1 Introduction

General circulation models (GCMs) attempt to embody the current understanding of climate dynamics via process equations and numerically solve these equations to simulate climate with various scenarios of human influences (Taylor et al. 2012). These models are complex and have been evolving since the 1960s (Manabe and Wetherald 1967). The output of GCMs is given a central place in formulating public energy policy. The basis for this central policy position is that the models are based on physics (IPCC 2013), with high confidence (>95%) given to many attribution and forecast results (IPCC 2013, SPM). IPCC also reports that GCMs do a good job of matching historical data and that without including greenhouse gases the match is not good (IPCC 2013, Fig. SPM.6).

There is a vast literature that compares GCM outputs to various climate features (see following sections). Such tests are complicated by the stochastic nature of both climate and the models. GCM vs. data comparisons are judged to be poor, adequate, good, or excellent, depending on the variable and the study (McWilliams 2007). This ambiguity results from a multiplicity of criteria of model goodness as well as varying results.

Evaluating knowledge claims (of which there are several) based on GCMs can be aided by a consideration of epistemology (see Williams 2001 for an overview), which is the logical framework for evaluating how we know and what is knowable. With an epistemological analysis, we can assess the status of a theory/model in terms of its logical basis, reliability, and rigor. With this framework we can

Electronic supplementary material The online version of this article (doi:10.1007/s00382-017-3717-7) contains supplementary material, which is available to authorized users.

✉ Craig Loehle
cloehle@ncasi.org

[1] National Council for Air and Stream Improvement, Inc. (NCASI), 1258 Windemere Avenue, Naperville, IL 60540, USA

evaluate both the tests of model goodness and the consistency of results derived from GCMs with known physics. I first illustrate these issues from several areas of science and then return to the question of the epistemological status of climate models.

2 Models and epistemology

Science is the process of formally discovering regularities in nature. An explanation of or formal model for a regularity in nature is called a theory (or law if it is well-supported). Newton's law of gravity is a classic and simple example. In this case, the obedience of objects to this law at human scales is apparently exact. Such highly accurate theories are commonly treated as explanatory.

A "hypothesis" is a term used in two different senses (Loehle 1983). Empirical relationships (e.g., drug trials) can form a statistical hypothesis but are not a theory until they are based on falsifiable mechanistic models or explanations. A scientific hypothesis, in contrast, is a proposed explanation for some relationship or process and can be at various levels of abstraction. Specific predictions derived from theory are the only aspects of a theory that are testable, not the theory as a whole. The rejection of a statistical (empirical) hypothesis provides useful information but does not necessarily carry theoretical content. The rejection of a test of a scientific hypothesis (if rigorous) should lead to the hypothesis (or theory) being revised, refined, or rejected (Loehle 1983).

The ideal case of testable theories can be found in classical physics. Newton's and Maxwell's laws make very specific predictions as well as forbidding certain things from happening. These laws were convincingly demonstrated by experiments, but note that even here confounding factors such as friction must be controlled in order to test them. In these cases, the standard of theory validity is very high. Experimental data often match theory almost perfectly and events such as the return of a comet can be predicted decades in advance. The apparent perfection of these laws has perhaps led to a belief that they are "true" in the absolute, logical sense, but even gravity has some unexplained features.

Valid and useful theories, however, do not spring into life fully formed and perfect, nor are they always as accurate as Maxwell's equations. When Alfred Wegener (trans. 1966) proposed the theory of continental drift in 1912, it cannot in any sense be said that his theory was mature. A mechanism for continental movement was lacking (and it seemed impossible to many that continents could move), as was sufficient supporting data. As data were gathered, particularly on sea floor spreading and the process of subduction, a coherent picture came into existence of plate movements, the rise of mountain ranges, the origin of volcanoes, and the reason for the location of earthquake zones. However, after a century of maturation of this theory, it remains a qualitative theory because while it can explain the general locations of earthquake and volcanic zones, it cannot predict the size, precise location, or timing of either earthquakes or volcanic eruptions due to the heterogeneity of the Earth's crust and the impossibility of obtaining detailed data. Thus, even a mechanistic and well-tested theory need not be able to make precise predictions, perhaps ever. As a theory matures, it hopefully becomes more precise, but this is not guaranteed (Loehle 1983).

It is important to distinguish scientific knowledge from everyday concepts of physics (see diSessa 1993). At an early age, children figure out that objects continue to exist when hidden and cannot be in two places at once. We know that certain things either occur (e.g., going to the store) or not. We understand that certain things happen with probability (e.g., drawing an ace of spades). Such concepts are captured formally by predicate logic and probability, respectively. Much of epistemology concerns these types of knowledge (Williams 2001). Unfortunately, scientific "proof" does not follow the predicate logic model. There is an asymmetry noted by Popper (1959, 1963) in his famous Principle of Demarcation: it is possible to reliably disprove a theory, but a theory can never be proven. Instead, successive successful tests of a theory only increase our confidence in it. This does not mean that we know nothing, as knowledge relativists might assert, but rather that scientific knowledge is provisional, bounded (gravity is not clearly explicable at the atomic level), and a matter of degree (Loehle 2011). In some cases this knowledge can encompass many significant digits, but in others, it may be more qualitative.

Critically, testing an evolving theory does not and should not follow the simple hypothesis testing model used in empirical experimentation. When testing a medicine vs. a placebo, a simple better or worse or a "how much" answer often results from statistical tests. When testing a theory, there are multiple aspects of the theory that may each receive partial support at a particular time, and alternate explanations that may need to be ruled out (Reiss 2015). A network of confirmation, mathematics, and causal explanation supports belief in a theory at any moment, not a simple yes/no. As a theory becomes more mature and more rigorously tested, we ascend the scale of epistemic certainty (left side of Fig. 1). There is an asymmetry, however, from proving a theory to using it for some calculation. The tests that lead to acceptance of a theory as "true" are often done under carefully controlled and ideal conditions, such as a vacuum. In any calculation based on a theory we may instead be using it under non-ideal conditions. For example, a falling feather behaves differently in a vacuum

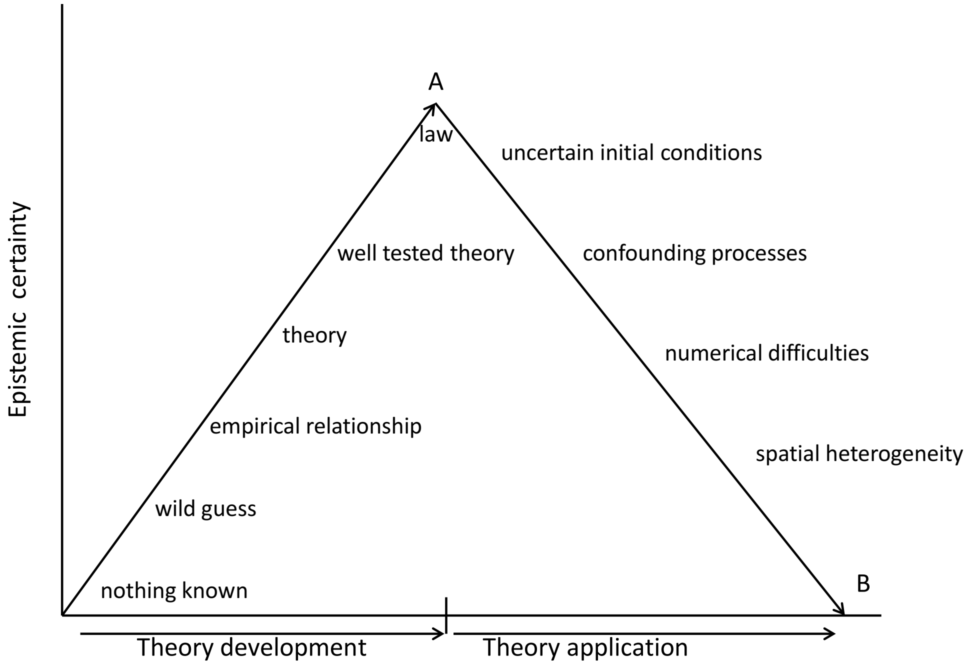

Fig. 1 The epistemic triangle. As theory is developed, epistemic certainty increases for ideal conditions (**A**). However, for applications an accumulation of unresolvable complications reduces certainty, even to zero (**B**); for example, for predicting the flight of a paper airplane or the fall of a feather due to turbulence

compared to in air. The bridge from idealized physics to real world applications is the set of approximations, simplifications, discretizations, empirical relationships, estimated initial conditions, and numerical methods used to create a calculation tool (Loehle 1983) that can be used to compute some result. These bridge relationships are what prevent a calculation tool from being a perfect representation of the underlying physical (or other) theory. If these confounding factors are sufficiently difficult to quantify and model, we may descend all the way down to the far right of Fig. 1, where we cannot make any predictions (e.g., for the path of a dropped feather) at B. The correctness of a calculation tool is thus an empirical question of how accurate or useful it is, rather than a question of true or false as we take it to be for theories/laws.

What then of "facts"? In everyday speech we often make statements about the existence of objects such as, "My office has a computer." Such existence statements can validly be called facts and are subject to yes/no evaluation. When we try to be more specific about these "facts", however, trouble arises. Any description (e.g., temperature, size) or classification (e.g., type of cloud) necessarily involves quantification or discretization, respectively, which can never be perfect. Thus, existence statements can be evaluated in a binary manner, but any description or prediction must be evaluated in terms of accuracy/precision.

When people speak of scientific facts, they are generally making a shorthand reference to some body of knowledge which they are claiming is valid or true. For example, someone may say "Evolution is a fact" by which they mean "Life evolved rather than being created as described in Scripture." However, the existence of a body of knowledge addressing evolution does not mean that all questions about this topic have been resolved. Likewise, plate tectonics as a fact does not enable us to make specific statements about particular volcanoes. When a knowledge claim is made at too high a level of abstraction, it is not epistemologically properly formed. For example, saying "physics is true" is meaningless.

Thus, statements that a scientific theory is a fact are denotations for bodies of more or less reliable knowledge. Such denotation may be useful for everyday conversation and general reasoning but is uselessly vague if we need specific information. The mere existence of a theory (as fact or truth) does not necessitate either precision of knowledge or predictability of events. Nor does it mean we know initial or boundary conditions well enough to use the theory or that we are applying the theory properly in any specific case. A putted golf ball follows Newton's laws of motion but the details of the green's surface may be unknowable, so it is not always possible to predict the ball's path.

Epistemology, then, allows us to make certain statements about theories and knowledge claims. A scientific statement can be an empirical relationship, for which we have no real explanation. It can be a wild guess, which can be more or less epistemically grounded (i.e., having valid reasons that it might be true). If a guess becomes better supported, it may become a provisional theory. An example would be evolution as framed by Darwin; not everything was explained and support was weak at the time. Highly developed theories are often called laws (e.g., Newton's laws of motion). This sequence represents a hierarchy of epistemological certainty, which never reaches 100% (peak of Fig. 1) because ultimate causes and fundamental forces are never perfectly definable. When we seek to apply laws or theories, new complications arise and we may descend from high epistemic certainty (right side of Fig. 1). Even in applying Newton's laws to a simple system, complications such as static electricity, air currents, friction, elastic rebound, and magnetic fields must be controlled or accounted for, and we may lack knowledge of how to do so in any particular case. In spatially extensive systems, new complications arise due to our inability to obtain initial conditions and the difficulties of solving spatially explicit equations. It can be difficult to quantify how much uncertainty these factors add to the result of a computation, but we are rarely in the same domain of high epistemic certainty that pertains to a law of physics tested under ideal conditions.

3 Basis of climate models in physics

What then is the epistemological status of GCMs in terms of their basis in physics? GCMs are a mix of simulated processes that are viewed as well-understood physics (e.g., radiative transfer) and those that are poorly understood (e.g., cloud microphysics, IPCC 2013, p. 599). To what extent can we trace the algorithms used directly back to known physics? To what extent does the basis in physics prove their truth value, explanatory power, or reliability? As we have seen above, theories in physics that approximate our common notions of "truth" are, at least in idealized settings (e.g., frictionless vacuums), able to make very precise real-world predictions. Can GCMs approximate such clean physical theories as Newton's laws of motion in a vacuum? If so, then a great deal of confidence in their results is warranted. However, even for a simple problem like tossing a die or flipping a coin, sensitivity to initial conditions means that the outcome cannot be predicted even though based on known physics. In the case of climate models, Rougier and Goldstein (2014) state that the laws of the Earth's climate system are not all known and are not explicitly solvable at sufficient resolution. Katzav et al. (2012) note that model completeness and structural stability are unknown. This is particularly true for the Navier–Stokes (N–S) equations for fluid dynamics, for which no analytic solutions are known. This inability to explicitly solve the equations is why numerical simulation is used. However, the proper simulation of the equations of fluid dynamics is far from straightforward (Thuburn 2008). A particular problem is that while the proper solution of these equations requires conservation of mass, energy, momentum, and other properties in a continuous fashion (at infinitely many scales) because they are partial differential equations, the models are discrete. Processes such as dissipation of energy and the propagation of vortices occur below the grid scale and no theory exists to guarantee that the gridded model handles them properly (McWilliams 2007; Marston et al. 2016). Simulated processes within a grid may not propagate smoothly to neighboring cells, creating the potential for ringing, the accumulation of numerical solution errors with time, or result in errors in winds or proper modeling of phenomena such as the Quasi-Biennial Oscillation (Thuburn 2008). These issues have not been adequately resolved (e.g., Katzav et al. 2012) and, in fact, the solution of N–S equations remains a Millennium problem (see http://www.claymath.org/millennium-problems/navier-stokes-equation). Thus, the models may violate conservation laws and exhibit numerical solution artifacts. Stevens and Bony (2013a) showed, for example, that even in an idealized model of a water planet with prescribed surface temperatures, the spatial responses of clouds and precipitation to warming are quite different depending on the model (SI Fig. 1). This illustrates that agreement has not been reached on how to represent or compute these processes on a grid. Zhou et al. (2015) document errors in how solar radiation is zonally averaged in some models. Staniforth and Thuburn (2012) document that all existing grid numerical solution schemes have known problems including grid imprinting and the excitation of computational modes. The inadequacy of current gridding schemes is shown by the fact that a higher resolution model often produces many differences compared to current models (Sakamoto et al. 2012). Improved numerical methods continue to be introduced to resolve the known problems with solving N–S PDEs (e.g., Marston et al. 2016). In addition, sub-grid parameterizations exist in all models (McWilliams 2007; Katzav et al. 2012; Hourdin et al. 2017) increasing uncertainty. McWilliams (2007) notes that small structural (equation form) differences in sub-grid parameterizations can lead to different dynamical attractors in such fluid dynamics systems.

There is considerable support for arguments that key feedback processes in the Earth climate system operate in a bottom-up manner and below the grid-scale used by GCMs. Stephens et al. (2015), for example, note that albedo values for the two hemispheres are nearly identical in spite of very different land/ocean configurations and note annual albedo buffering as well, suggesting the operation of negative feedback processes not captured by GCMs. A series of papers (Stevens and Bony 2013a, b; Xiao et al. 2014; Bony et al. 2015; Mauritsen and Stevens 2015) show that key cloud and energy dissipation processes are affected by turbulence and thunderstorm aggregation effects at the sub-grid scale such that net cloud feedbacks in GCMs may be quite wrong (see also Lacagnina and Selten 2014). A link between cloud feedbacks and ENSO has been proposed, with results from data and models not in agreement (Sun et al. 2009). It has recently been shown how the spatial pattern of warm and cool pools in the Pacific can alter large-scale cloud cover enough to alter global temperatures (Mauritsen 2016; Zhou et al. 2016). It has further been argued that the diagnosis of feedbacks is far from simple (Spencer and Braswell 2011).

The deficiencies in the solution to the N–S equations also ramify through other aspects of Earth system simulations besides sub-grid parameterizations. Proper simulation of ocean circulation is critical to predicting ocean heat uptake and latitudinal heat distribution and radiation to space as well as the dynamics of phenomena such as ENSO, which at present can be qualitatively simulated but not in terms of the timing or magnitude of events (McWilliams 2007). The upwelling and turnover of moist tropical air at the Intertropical Convergence Zone is fundamentally a fluid dynamics phenomenon that is currently not handled properly by GCMs, as are large convective systems, the Walker circulation, and other aspects of the redistribution and dissipation of heat by the global heat engine that are not properly simulated (see Zhou and Xie 2015). Thus, the inability to handle an N–S system adequately may affect the simulated net energy balance of the Earth as well as spatial patterns of climate.

What about the principle of demarcation of popper? Do the GCMs as embodiments of theory make strong predictions that would qualify as a rigorous test of correctness in spite of numerical difficulties? An example of a strong prediction made by climate theory is the tropical tropospheric hot spot, prominently featured in the IPCC Fourth Assessment Report. This prediction has not yet been verified (e.g., McKitrick et al. 2010; Po-Chedley and Fu 2012) even though theory suggests it should be evident by now. However, we cannot say it has been disproven due to data uncertainties. The divergence of global surface temperature in models vs. data post-2000 (e.g., Stott et al. 2013; Outten et al. 2015) and the related pause in warming (IPCC 2013, p. 870; Thorne et al. 2015; Trenberth 2015) indicate that forecasts produced by GCMs are not entirely consistent with climate theory. On the other hand, other authors looking at past predictions of global temperatures (e.g., Hargreaves 2010; Frame and Stone 2013) report that the first IPCC assessment predictions have held up well, though these forecasts were based on both models and forcing data that differ from those currently used, and they used results ending 4–6 years ago. Stouffer and Manabe (2017) compared spatial pattern projections of warming made in 1989. They found good qualitative agreement in some but not all regions, but it is difficult to assess the significance of a qualitative comparison.

A valid out-of-sample test of GCMs would be the ability to match ancient climates that were not used to build the models. Tests of GCMs for paleo-climates of the Holocene (Bakker and Renssen 2014; Harrison et al. 2014; Liu et al. 2014), last glacial period (Harrison et al. 2014), multiple interglacials (Bakker et al. 2014), and the Miocene (Steppuhn et al. 2007) have not shown very good agreement, though the role of paleo-climate and forcing test data uncertainty is difficult to separate from model failures. The ambiguity of these tests, while not adding to confidence in the models, also does not allow them to be rejected. These and similar tests do, however, enable us to say that this type of out-of-sample confirmation of model validity has not occurred.

Let us consider the most fundamental physics of climate models: the radiative properties of CO_2 in the atmosphere. While there is indeed a basic theory for this process, there are many radiative transfer software tools (Oreopoulos and Mlawer 2010) because calculation of radiative transfer on a globe with a heterogeneous atmosphere is a difficult numeric problem, unlike the acceleration of a falling body in a vacuum. The spectrum is evaluated at different resolutions using various geometric assumptions and methods in each of these tools. More seriously, Oreopoulos and Mlawer (2010) document that (1) the basic theory itself continues to evolve; (2) the algorithms used in GCMs are much simplified due to computational considerations; and (3) different GCMs do not use the same radiative transfer algorithms. It is thus clear that even here there is a gap between basic theory and what is computed, with unclear consequences.

Likewise, each GCM makes different assumptions about forcing histories, clouds, land surfaces, spatial gridding, etc., and uses different numerical methods for solution. Estimated forcings changed considerably between the IPCC AR4 and AR5 reports, and the effect of aerosols is still being revised (e.g., Stevens 2015) with major differences in representation between models (Wilcox et al. 2013). Parameterizations (i.e., empirical relationships) are used for processes that take place below the grid resolution, such as cloud behaviors and precipitation (McWilliams 2007). These empirical relationships have free parameters that must be tuned (Lahsen 2005; McWilliams 2007; Mauritsen et al. 2012; Schmidt and Sherwood 2015; Hargreaves 2010; Hourdin et al. 2017) and these tunings can be arbitrary (e.g., Soon et al. 2001, their Fig. 4). Errors in these approximations are difficult to quantify, but certainly take the models far from the domain of pure representation of ideal laws of physics such as black-body radiation from a uniform surface of known temperature, as also argued by Katzav et al. (2012). Arguments can also be made that significant physical processes are left out of the models, such as effects of the Earth's electric field (Andersson et al. 2014).

Thus, these models are not "a theory" such as the law of gravity. The many processes incorporated into the computer software come from many different disciplines.

Many relationships in them are empirical, and some, such as cloud behaviors, are approximations of unknown validity. GCMs are thus calculation tools based on physics, as also argued by Rougier and Goldstein (2014). In some cases, the physics used in different GCMs even represents competing physical theories for particular processes (Schmidt and Sherwood 2015). In addition, the verisimilitudes of the gridding and numerical solutions of fluid dynamics are themselves open to question (Thuburn 2008). Until recently, for example, flux adjustments were necessary to overcome numerical solution deficiencies (Lahsen 2005).

If GCMs cannot be viewed as precise representations of theory based on the derivation of some components from well-supported physics (per above), what epistemological status do they have? One approach to assessing their truth value is to argue, not forward from the underlying physics, but back from the quality of their outputs. It can be successfully argued that they do embody aspects of current understanding of the Earth climate system or they would not work at all. Katzav (2014) and Schmidt and Sherwood (2015), for example, argue that this knowledge embodiment is indicated by the superiority of current models compared to a naïve model or compared to previous generation climate models. Smith (2002), Hargreaves and Annan (2014), and Oreskes et al. (1994) suggest that the models are a useful analogy or heuristic. McWilliams (2007) argues that because of irreducible uncertainty in model outputs due to chaotic dynamics, GCMs should be judged based on plausibility rather than whether they are correct or best. He argues that the models "yield space–time patterns reminiscent of nature ... thus passing a meaningful kind of Turing test between the artificial and the actual." The IPCC (2013, p. 145) states that these models can be viewed as tools for learning about the climate system. Many outputs (particularly temperature) show good agreement between models, indicating some sort of truth value to the models (Räisänen 2007). However, inter-model agreement can arise from common assumptions, shared algorithms, and similar data used for tuning. Parker (2011) argues that agreement of predictions across models, while providing some supporting evidence, is not sufficient to establish any epistemic certainty in their truth value. For these reasons, efforts to confirm (verify) climate models (e.g., Lloyd 2010, discussion in Katzav et al. 2012) are missing the point. While these models can be plausible, pass a Turing test of sorts, and agree with each other, the problems of irreducible dynamics and numeric uncertainty (e.g., McWilliams 2007) and other issues mean that the theoretical underpinning of the models cannot be assumed to imply validity for making useful predictions. This raises the question of their usefulness as predictive tools, discussed next.

4 Climate models as calculation tools

Because GCMs are continuously evolving and some aspects may lack a rigorous and close link to the underlying physics, they are unfalsifiable by Popper's criteria (see Curry and Webster 2011), and must be judged as calculation tools. It is thus necessary to test the models in some way before using them.

Testing complex simulation models is difficult. The large number of tuned (estimated from data) parameters in these models (Murphy et al. 2004; Hargreaves 2010; Schmidt and Sherwood 2015; Hourdin et al. 2017) suggests that model parametric uncertainty could be high but this has been insufficiently evaluated to date (Guttorp 2014). There are potential structural (equation form), parameter, and data error issues (Loehle 1987, 1988; Hourdin et al. 2017) that have been little explored. There are many specific types of sensitivity and error analyses that can be conducted (e.g., Falloon et al. 2014; Guttorp 2014; Rougier and Goldstein 2014) to evaluate the reliability of model outputs, but these methods have almost never been applied to GCMs because of their large computational burden (Falloon et al. 2014). Allen and Ingram (2002) and McWilliams (2007) argue that ensembles of opportunity (a collection of models) do not adequately sample model uncertainty and recommend a full uncertainty (initial condition, parametric, equation functional form, numerical method, etc.) analysis in order to bound possible forecasts, an analysis which has still not been performed for GCMs. Thus, critical information for decision makers on model uncertainty is not available for GCMs.

Models of turbulent dynamics exhibit sensitivity to initial conditions (Frigg et al. 2013; Collins 2002). Given a structurally perfect model (i.e., all equations and parameters are correct; numerical methods work correctly), the effect of initial condition uncertainty can be estimated by making multiple runs with perturbed initial conditions, giving a probability distribution for the outputs. This assumes that the errors in initial conditions can be characterized and that a sufficient number of runs can be made, neither of which is usually true in the case of climate models (McWilliams 2007). In a unique case study, Deser et al. (2016) perturbed a base run with machine error-level noise (i.e., round-off error) applied to the initial temperature field. They found very large differences in winter 50 year trends for regions of North America across 30 runs of several °C. They found that an ensemble approach could separate the internal variability vs. the forced signal to give better agreement with historical data. However, this is based on an infinitesimal initial condition perturbation. True initial condition uncertainties are many orders of magnitude greater. More significantly, if there are any structural errors (wrong equation form to represent a process), this stochastic perturbation of initial conditions can be not only uninformative, but misleading (Smith 2002; Frigg et al. 2014; Hourdin et al. 2017).

For certain parameters (e.g., aerosol forcing, IPCC 2013, Fig. 7.19), the uncertainty is large. Schwartz (2004) argued that uncertainty in the amount of aerosols and their effect would need to be reduced threefold to properly identify radiative forcing due to anthropogenic effects. It is clear that the physics of cloud formation is still insufficiently understood to allow clouds to be properly simulated. Perturbed physics analyses (Collins et al. 2011) attempt to evaluate the magnitude of parametric uncertainty by perturbing parameter values but this again assumes that no structural errors exist. In addition, far too few runs have been made even for a proper parametric sensitivity analysis in most cases. Hourdin et al. (2017), Katzav et al. (2012), Mauritsen et al. (2012), Soon et al. (2001), and Kiehl (2007) all found that multiple tunings of the models can produce similar outputs (i.e., the models are poorly constrained), which suggests that tuning is not mechanistically sound. Finally, the pool of multiple climate models may not sample the uncertainty due to structural error (see Tebaldi and Knutti 2007; Hargreaves 2010; Collins et al. 2011; Frigg et al. 2013). However, GCMs are ensembles of opportunity and share data, code, and assumptions (Parker 2011; Katzav et al. 2012; Katzav 2014). Different methods for weighting and combining ensemble members can give very different outcomes for ensemble means or distribution statistics (Tebaldi and Knutti 2007). Furthermore, unlike initial condition error or parametric error which can in many cases be reasonably characterized, structural error (wrong equation form, missing processes, numerical computation error; see Loehle 1987) is not characterizable by a distribution (e.g., Gaussian) and is not finitely delineable (McWilliams 2007). For example, McNeall et al. (2016) document that for the land surface forest model component of the reduced resolution climate model FAMOUS, parameters fit to data for the Amazon forest yield a model that does not work properly elsewhere or when other forests are used for fitting, indicating a structural error. For this reason, an ensemble of runs from different models cannot be viewed as sampling a meaningful model space and neither the ensemble distribution nor the mean of the ensemble can be assumed to have any epistemic meaning or truth value (Winter and Nychka 2010; Curry and Webster 2011; but see; Gleckler et al. 2008). What can be shown from these types of comparisons of outputs is that the currently knowable uncertainty is large (Curry and Webster 2011) and may not encompass the true values (McWilliams 2007; Frigg et al. 2014).

Complex computational tools with multiple outputs cannot be evaluated based on a single output. For example, the match of model global mean temperature history with data could be achieved with regional temperature values that are incorrect everywhere (e.g., Arctic too cold but tropics too warm). As noted by Shepherd (2014) and Räisänen (2007), the verisimilitude of precipitation regimes by the GCMs is poor and unrelated to the agreement of models on temperature. Thus, broad, long-term temperature history verisimilitude does not necessarily imply realism of precipitation or smaller-scale features of climate, nor does it mean that response to increased forcing will be correct. Rougier and Goldstein (2014) suggest that proper acceptance testing of these models should include a decision to not make a forecast for any model or model-specific output that cannot meet reasonable accuracy limits compared to historical climatologies. Such is standard practice in engineering but there is no counterpart in climate science (Guillemot 2010).

It may be more informative to examine GCM outputs more narrowly rather than as a whole to see what can be predicted with sufficient accuracy. The IPCC (2013) graphs GCM outputs of global mean temperature since 1850 on an anomaly basis (as departures from the mean), but if plotted on an absolute temperature basis, the time series differ by up to 4 °C (SI Fig. 2). A similar result (up to 4 °C offsets) was found for the continental US (Anagnostopoulos et al. 2010). This is not a trivial difference because long-wave radiation from an object by the Stefan–Boltzmann relation is proportional to the fourth power of the surface absolute temperature (Anagnostopoulos et al. 2010). If models differ in mean temperature by this much, are they handling the basic physics in the same ways or implementing the physics with correct algorithms? This raises epistemic questions about the forecasts produced by GCMs. Hawkins and Sutton (2016) note that it has been argued that if the response to increased forcing is linear, then the absolute temperature does not matter much for estimating a response to increased forcing. However, if there is strong positive feedback, then response to increased forcing is greater at higher temperatures (Bloch-Johnson et al. 2015; Gregory et al. 2015). If, in contrast, negative feedback acts to dampen CO_2 forcing (e.g., Spencer and Braswell 2011), this would also depend on actual temperature. In either case, absolute temperature would matter (i.e., the response is nonlinear) and the use of anomalies cannot be justified. Anomalies, sometimes called “bias-correction”, are also used for comparing other climate outputs. However, crops, biodiversity, sea level, and ice sheets all respond to actual precipitation and temperatures, and thus the different models would forecast very different impacts even if their anomaly trends matched, as noted by Hawkins and Sutton (2016). The net effect of bias correction or use of anomalies is to obscure the epistemological status of the models by reducing the spread of the model outputs with respect to each other and making disagreements with data difficult to determine.

The use of bias correction can cause other difficulties with testing. Consider the case of comparing global temperature histories to model outputs. If data are in actual °C or are shifted to a common baseline over some period, the correlation statistic is not affected because the constant term drops out of the computation. For other measures, however, the baseline can have an effect. For example, the R^2 statistic for model goodness of fit will be different for actual vs. anomaly series, and can actually be negative for unshifted series (i.e., the fit to data is worse than to a simple mean of the data). Hawkins and Sutton (2016) note that normalization (baseline shifting) of a climate series is based on a reference period, typically 30 years, but it can be the entire period of record. Both data and model output are shifted up or down so that their respective means over the reference period are zero. When comparing multiple runs of a single model or of multiple models vs. data, they will all agree most closely during the reference period. This means that the visual impression of model fit or the timing of model good or bad performance can depend completely on the reference period chosen (see Hawkins and Sutton 2016 for examples). This impacts, for example, the question of whether models are currently running hotter than the data. The closer the chosen reference period is to the

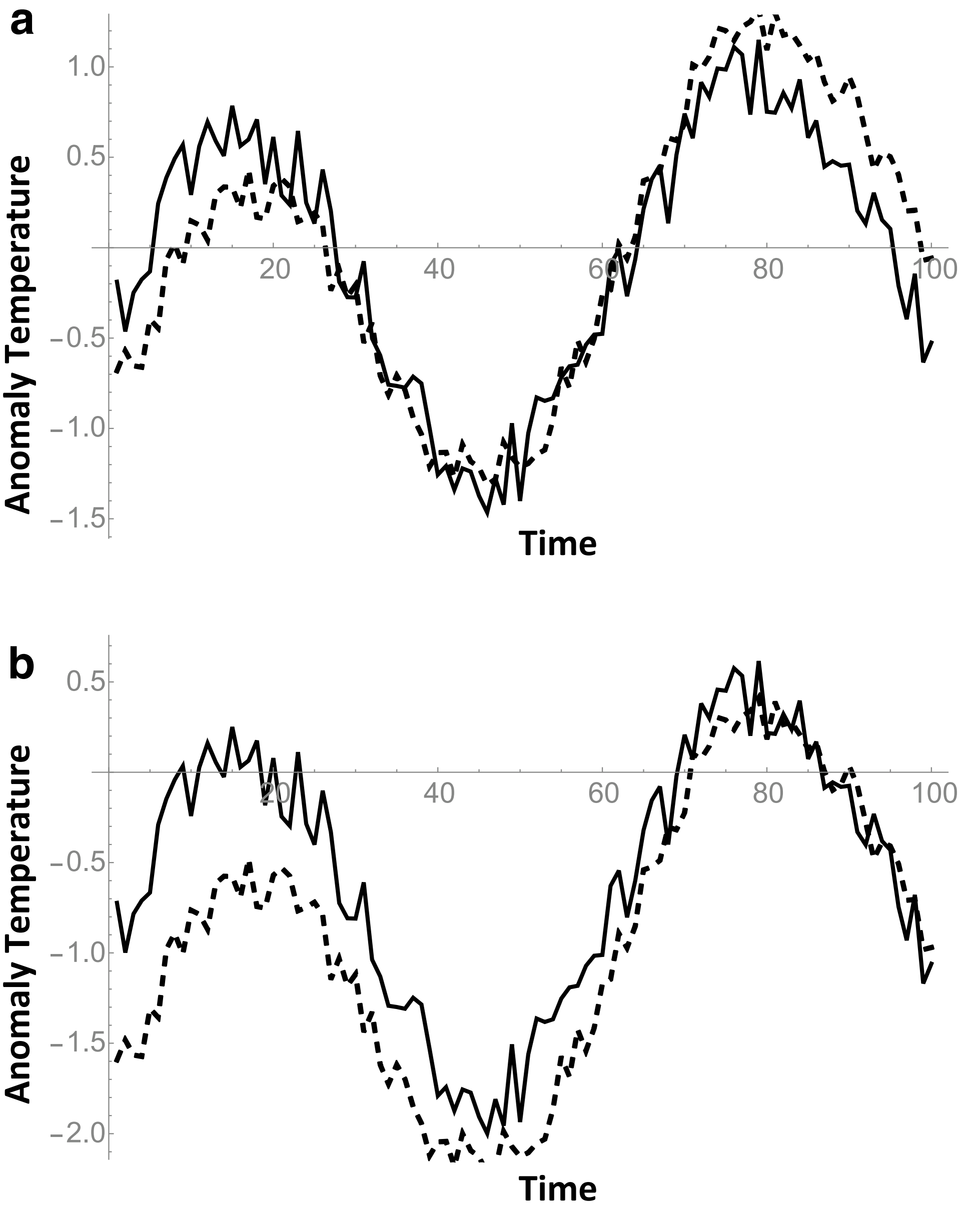

Fig. 2 Effect of reference period choice on visual and numeric goodness-of-fit. A 100 year arbitrary time series was generated with a slight upward trend plus sinusoidal signal and noise (*solid line*). A model was generated with different noise and a steeper rise (*dashed line*). **a** Adjusted to 100 year reference period; $R^2=0.79$. **b** Adjusted to most recent 30 year reference period; $R^2=0.54$

present, the greater the apparent agreement between the models and data in recent years. For fit statistics such as R^2, the choice of reference period can also affect the result and thus the implied model fit. For example, in Fig. 2 an artificial example is shown. In Fig. 2a, the data and model are both shifted to the 100 year reference period (mean 0). The fit appears visually to be quite good, and $R^2 = 0.79$. However, in Fig. 2b the most recent 30 years is used as the reference period. Now the model appears to fit worse in the past and better (almost perfectly) in recent decades, but now $R^2 = 0.54$, a considerable degradation. This raises an epistemic dilemma. If correlation is used as a measure of common trend and pattern (e.g., ups and downs of temperature), this does not account for the bias (offset) in model outputs. If models and data are put on an anomaly basis, this assumes for temperature and precipitation that actual values don't matter, only the trend, but this is still open to debate. Furthermore, the reference period chosen affects both the visual impression of model goodness-of-fit (for both ensemble spread and pattern of fit over time) and all fit statistics except simple correlation. Issues such as this have implications for epistemic certainty.

Comparisons of trends may also be affected by the time segments chosen for analysis. A trend starting in 1980, for example, could be confounded by internal Earth system cycles like the PDO or AMO (Loehle 2014, 2015). For a non-experimental system, the fact that choices of time period for analysis affect results and may be influenced by confounding raises a unique type of epistemic uncertainty.

Assuming that the choice of time period for analysis is valid, some statistical challenges remain. In typical statistical analyses, we may wish to test a hypothesis that some treatment is different from zero or that two treatments differ from each other. One- or two-tailed t-tests provide a simple example. The null hypothesis is that the two treatments do not differ and we examine whether the null should be rejected based on results of our statistical test. In climate science, in contrast, we often wish to test whether two things do not differ (i.e., that the model and data match). Loehle (1997), Robinson and Froese (2004) and Robinson et al. (2005) argue that the proper approach is to frame the null as model failure and attempt to reject it. The statistical power of the data (sample size and variance) then become critical along with the precision with which we wish to compare model and data. In experimental statistics, power analysis is used to specify how many samples would be needed to obtain a given level of precision in tests. Criteria should be such that a rejection of the null implies some useful degree of precision vs. data. See also Meehl (1997) for a discussion of prediction precision and confidence intervals on results in preference to simple hypothesis tests.

The concept of a null expectation is relevant to evaluation of time series and trends. Highly nonlinear dynamic systems are likely to oscillate (McWilliams 2007). It has been shown that historically the Earth's climate has fluctuated at all temporal scales (Lovejoy 2015). In fact, mechanisms are known by which internal oscillations can arise, be maintained, and affect global temperatures (Mauritsen 2016; Zhou et al. 2016). As such, their dynamics may be bounded but may lack an "equilibrium" and may thus only be characterized by an invariant measure (e.g., an orbit) that gives a distribution of possible states. The sunspot cycle, driven by a heated fluid (the sun), is an example; the pattern is bounded but has so far (in historical records) never repeated exactly. In the Earth system there is evidence for endogenous ocean circulation oscillations (e.g., Wang et al. 2015), which might be emergent properties of chaotic dynamics on bounded geographic features such as ocean basins. The fact that past climates have always fluctuated (McWilliams 2007) prevents us from ruling out endogenous oscillations of potentially large magnitude and over long time periods (e.g., centuries). That is, the null model for temperature trends cannot be assumed *a priori* to be strict stability. In fact, a toy model has been developed that demonstrates this point. Koutsoyiannis (2006) developed a model with a positive and negative feedback term, each based on the chaotic tent map. This deterministic model was shown to be able to match integrated (smoothed) data for multiple long timeseries of river flow and temperature, including long periods of rise or fall, as well as the scaling exponent. The ups and downs at all scales were present solely as a deterministic function of the chaotic model. This means that chaotic dynamics could be a sufficient null model for climate, as could quasiperiodic external (e.g., solar, cosmic ray, gravitational) forcings. It is not necessary for natural fluctuations to account for all of the recent warming to be a plausible factor. Instead, even a partial effect will reduce the estimate of climate sensitivity (e.g., Loehle 2015). The importance of this alternative null for testing climate models involves the extent to which the test is strong or weak (senso Meehl 1997). If no alternate explanation exists for warming post-1950, then the match of models is a strong test, which is what is assumed. But if internal oscillations can produce such a pattern of temperature, then it is not a strong test.

5 Conclusions

What, then, of the knowledge question posed by GCMs? As parameterized simulators that generate climate behavior, these tools must fundamentally be judged statistically, quantitatively. Qualitative assessments do not answer the key policy-relevant questions of how much warming, when, and where. Held (2005) argues that achieving improved knowledge of the climate requires the development of simplified, idealized “worlds” (e.g., see SI Fig. 1) to enable an exploration of the processes of large-scale turbulence, heat transfer to the poles, ocean circulation, and particularly how large climate features such as ENSO can persist. Without this exploration of mechanisms, Held argues, it is not possible to explain why different GCMs produce different outputs, why they differ from data, and how they can be improved. This is because the complexity of the models results in epistemic opacity. Proper explanations of the behavior of complex hierarchical systems such as the climate must usually be multilevel and account for factors such as ocean currents, continents, and clouds. Improved understanding achieved in this way could lead to better sub-grid parameterizations. An example is the recent work by Moncrieff et al. (2017) which derives a multi-scale approach to understanding of organized tropical convection that can be used to develop sub-grid parameterizations.

According to Fogelin (1994), making a knowledge claim requires both epistemic responsibility and adequate grounding (or justification), which requires proper reasoning and an adequate basis in data, facts, and theory. Fogelin (1994) also argues that potentially misleading information, such as confounding by uncontrolled factors or unmeasured processes, must be considered epistemically and reduces certainty in conclusions. In the climate change arena, confounding could result from getting the right answer (realistic looking output) for the wrong reason. We can identify several candidates for such confounding. First, if assumed aerosol concentrations and forcings are too high for the past 80 years or so, then if the model response has been tuned to match historical temperatures (see Schwartz 2004; Tebaldi and Knutti 2007; Hourdin et al. 2017) this will yield a high estimate of climate sensitivity and thus of future warming. New lower estimates of aerosol forcing (Stevens 2015) highlight the problem. A second cause of confounding could arise due to internal Earth system fluctuations. The ENSO system is a short-cycle example, but longer cycles plausibly exist (e.g., the Pacific Decadal Oscillation, Atlantic Multidecadal Oscillation) which could account for part of late twentieth Century warming, in which case a lower climate sensitivity is implied (see Loehle 2014, 2015 and references therein). Third, the models could be tuned to match historical data, including choice of aerosol history (see Kiehl 2007), solar forcing history, sea temperature record, assumptions about ocean turnover, and so on (Knutti et al. 2002), in which case their fit to this data is not unambiguous evidence of model validity. Hourdin et al. (2017) and McWilliams (2007) note that tuning of GCMs does in fact take place and that it may be impossible to avoid using knowledge of twentieth Century warming histories during the tuning process. In fact, they note that some modeling teams use temperature trends explicitly for tuning.

In these three cases, the models may match twentieth Century temperatures for the wrong reasons (Tebaldi and Knutti 2007; Hourdin et al. 2017). If so, the epistemically justified approach is to quantify the level of uncertainty associated with knowledge/reliability claims or to rigorously show that such potentially confounding factors are not in fact affecting one’s results. Assuming model correctness in order to test for confounding presents the risk of circular reasoning according to Tebaldi and Knutti (2007). In the face of non-trivial counterfactuals (such as known numerical solution problems or unresolved confounding), one should report the uncertainty (Curry and Webster 2011) and note its implications for knowledge claims (Williams 2001).

The challenge of epistemic responsibility is even greater for knowledge claims based on GCM forecasts of sub-global scale changes, which is the scale where impact assessments necessarily are conducted. Not only is it known that GCMs fail to properly simulate smaller-scale features such as the QBO or the ITCZ, but GCMs disagree with each other at regional scales, making forecasts about regional impacts arbitrary (see Anagnostopoulos et al. 2010; Kundzewicz and Stakhiv 2010; Dawson et al. 2012; Chen and Frauenfeld 2014; Hall 2014; Deser et al. 2016). More detailed regional forecasts are made by using the coarser-scale GCM output as boundary conditions, but this dynamical downscaling process itself does not appear to be reliable (e.g., Evans and McCabe 2013; Hall 2014). However, regional forecasts are rarely evaluated critically (Hall 2014). While the reliability of regional forecasts cannot be precisely determined, good practice should at least include using ensembles (not just the mean of an ensemble) to give some idea of uncertainty. Bias adjustments (e.g., Ho et al. 2012) may also be needed to properly utilize regional or local model outputs for impact studies.

If climate models are only “similar to” the real Earth system and act more as an analogy (Oreskes et al. 1994) or as exploratory tools, then they are most useful as a basis for qualitative predictions such as that some warming is likely. If the models can make some predictions (e.g., global temperature) with acceptable precision, it is important to determine which variables can be so predicted. If models exhibit a common bias, perhaps this bias can be accounted for in making policy decisions. Explanations for model performance differences should be pursued, especially the wide range of future trajectories. Given the complexity of the Earth climate system, the foundational basis for the knowledge claims made based on GCMs deserves greater attention. Epistemology, properly applied, can help clarify what we know, how we know it, and the limits of rigorous reasoning that can be justified.

Climate change poses a wicked policy problem. There is a high risk both from action and inaction. This paper does not lead to any particular policy conclusion. Rather, it focuses on the methods that lead to rigorous reasoning. Policy decisions necessarily also involve perceptions of risk, tolerance of risk, cultural values, economics, and other factors beyond the scope of this analysis. However, any policy can only benefit from a better understanding of how climate models are constructed, their physical basis, how they can be tested, and how to assess their outputs.

Acknowledgements The author notes no financial or other conflicts of interest. No outside funding was used to perform this work. Thanks to M. Briggs, P. Frank, W. Kininmonth, W. Eschenbach, D. McLaughlin, W. Soon, and D. Stockwell for constructive comments.

References

Allen MR, Ingram WJ (2002) Constraints on future changes in climate and the hydrological cycle. Nature 419:224–232

Anagnostopoulos GG, Koutsoyiannis D, Christofides A, Efstratiadis A, Mamassis N (2010) A comparison of local and aggregated climate model outputs with observed data. Hydrol Sci J 55:1094–1110

Andersson ME, Verronen PT, Rodger CJ, Clilverd MA, Seppälä A (2014) Missing driver in the sun-earth connection from energetic electron precipitation impacts mesospheric zone. Nat Commun 5:5197

Bakker P, Renssen H (2014) Last interglacial model-data mismatch of thermal maximum temperatures partially explained. Clim Past 10:1633–1644

Bakker P, Masson-Delmotte V, Martrat B, Charbit S, Renssen H, Groeger M, Krebs-Kanzow U, Lohman G, Lunt DL, Pfeiffer M, Phipps SJ, Prange M, Ritz SP, Schulz M, Stenni B, Stone EJ, Varma V (2014) Temperature trends during the present and last interglacial periods—a multi-model-data comparison. Quat Sci Rev 99:224–243

Bloch-Johnson J, Pierrehumbert RT, Abbot DS (2015) Feedback temperature dependence determines the risk of high warming. Geophys Res Lett 42:4973–4980

Bony S, Stevens B, Frierson DMW, Jakob C, Kageyama M, Pincus R, Shepherd TG, Sherwood SC, Siebesma AP, Sobel AH, Watanabe M, Webb MJ (2015) Clouds, circulation and climate sensitivity. Nat Geosci 8:261–268

Chen L, Frauenfeld OW (2014) Comprehensive evaluation of precipitation simulations over China based on CMIP5 multimodel ensemble projections. J Geophys Res: Atmos 119:5767–5786

Collins M (2002) Climate predictability on interannual to decadal time scales: the initial value problem. Clim Dyn 19:671–692

Collins M, Booth BBB, Bhaskaran B, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2011) Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. Clim Dyn 36:1737–1766

Curry JA, Webster PJ (2011) Climate science and the uncertainty monster. Bull Am Meteorol Soc 92:1667–1682

Dawson A, Palmer TN, Corti S (2012) Simulating regime structures in weather and climate prediction models. Geophys Res Lett 39:L21805

Deser C, Terray L, Phillips AS (2016) Forced and internal components of winter air temperature trends over North America during the past 50 years: mechanisms and implications. J Clim 29:223–2258

diSessa AA (1993) Toward an epistemology of physics. Cogn Instr 10:105–225

Evans JP, McCabe MF (2013) Effect of model resolution of a regional climate model simulation over southeast Australia. Clim Res 56:131–145

Falloon P, Challinor A, Dessai S, Hoang L, Johnson J, Koehler A-K (2014) Ensembles and uncertainty in climate change impacts. Front Environ Sci 2:33

Fogelin RJ (1994) Pyrrhonian reflection on knowledge and justification. Oxford University Press, Oxford

Frame DJ, Stone DA (2013) Assessment of the first consensus prediction on climate change. Nat Clim Change 3:357–359

Frigg R, Smith LA, Stainforth DA (2013) The myopia of imperfect climate models: the case of UKCP09. Philos Sci 80:886–897

Frigg R, Bradley S, Du H, Smith LA (2014) Laplace's demon and the adventures of his apprentices. Philos Sci 81:31–59

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res 113:D06104

Gregory JM, Andrews T, Good P (2015) The inconstancy of the transient climate response parameter under increasing CO_2. Philos Trans R Soc A 373:20140417

Guillemot H (2010) Connections between simulations and observation in climate computer modeling. scientists' practices and 'bottom-up epistemology' lessons. Stud Hist Philos Mod Phys 41:242–252

Guttorp P (2014) Statistics and climate. Ann Rev Stat Appl 1:87–101

Hall A (2014) Projecting regional change. Science 346:1461–1462

Hargreaves JC (2010) Skill and uncertainty in climate models. Wiley Interdiscip Rev Clim Change 1:556–564

Hargreaves JC, Annan JD (2014) Can we trust climate models? WIREs Clim Change 5:435–440

Harrison SP, Bartlein PJ, Brewer S, Prentice IC, Boyd M, Hessler I, Holmgren K, Izumi K, Willis K (2014) Climate model benchmarking with glacial and mid-Holocene climates. Clim Dyn 43:671–688

Hawkins E, Sutton R (2016) Connecting climate model projections of global temperature change with the real world. Bull Am Meteorol Soc 2016:963–980

Held IM (2005) The gap between simulation and understanding in climate modeling. Bull Am Meteorol Soc 86:1609–1614

Ho CK, Stephenson DB, Collins M, Ferro, C.A.T., Brown SJ (2012) Calibration strategies—a source of additional uncertainty in climate change projections. Am Meteorol Soc 1:21–26

Hourdin F, Mauritsen T, Gettelman A, Golaz J-C, Balaji V, Duan Q, Folini D, Ji D, Klocke D, Qian Y, Rauser F, Rio C, Tomassini L, Watanabe M, Williamson D (2017) The art and science of climate model tuning. Bull Am Meteorol Soc 98:589–602

IPCC (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. In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) Cambridge University Press, Cambridge, pp 1535

Katzav J (2014) The epistemology of climate models and some of its implications for climate science and the philosophy of science. Stud Hist Philos Mod Phys 46:228–238

Katzav J, Dijkstra HA, de Laat ATJ (2012) Assessing climate model projections: state of the art and philosophical reflections. Stud Hist Philos Mod Phys 43:258–276

Kiehl J (2007) Twentieth century climate model response and climate sensitivity. Geophys Res Lett 34:L22710

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

Koutsoyiannis D (2006) A toy model of climatic variability with scaling behaviour. J Hydrol 322:25–48

Kundzewicz ZW, Stakhiv EZ (2010) Are climate models ‘ready for prime time’ in water resources management applications, or is more research needed? Hydrol Sci J 55:1085–1089

Lacagnina C, Selten F (2014) Evaluation of clouds and radiative fluxes in the EC-Earth general circulation model. Clim Dyn 43:2777–2796

Lahsen M (2005) Seductive simulations? Uncertainty distribution around climate models. Soc Stud Sci 35:895–922

Liu Z, Zhu J, Rosenthal Y, Zhang X, Otto-Gliesner BL, Timmermann A, Smith RS, Lohmann G, Zheng W, Timm OE (2014) The Holocene temperature conundrum. Proc Natl Acad Sci 11:E3501–E3505

Lloyd EA (2010) Confirmation and robustness of climate models. Philos Sci 77:971–984

Loehle C (1983) Evaluation of theories and calculation tools in ecology. Ecol Modell 19:239–247

Loehle C (1987) Errors of construction, evaluation, and inference: a classification of sources of error in ecological models. Ecol Modell 36:297–314

Loehle C (1988) Philosophical tools: potential contributions to ecology. Oikos 51:97–104

Loehle C (1997) A hypothesis testing framework for evaluating ecosystem model performance. Ecol Modell 97:153–165

Loehle C (2011) The logic of scientific discovery. Curr Trends Ecol 2:75–81

Loehle C (2014) A minimal model for estimating climate sensitivity. Ecol Modell 276:80–84

Loehle C (2015) Global temperature trends adjusted for unforced variability. Univ J Geosci 3:183–187

Lovejoy S (2015) A voyage through scales, a missing quadrillion and why the climate is not what you expect. Clim Dyn 44:3187–3210

Manabe S, Wetherald RT (1967) Thermal equilibrium of the atmosphere with a given distribution of relative humidity. J Atmos Sci 24:241–259

Marston JB, Chini GP, Tobias SM (2016) Generalized quasilinear approximation: application to zonal jets”. Phys Rev Lett 116:21450

Mauritsen T (2016) Clouds cooled the earth. Nat Geosci doi:10.1038/ngeo2838.

Mauritsen T, Stevens B (2015) Missing iris effect as a possible cause of muted hydrological change and high climate sensitivity in models. Nat Geosci 8:346–351

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. J Adv Model Earth Sys 4:M00A01.

McKitrick R, McIntyre S, Herman C (2010) Panel and multivariate methods for tests of trend equivalence in climate data series. Atmos Sci Lett 11:270–277

McNeall D, Williams J, Booth B, Betts R, Challenor P, Wiltshire A, Sexton D (2016) The impact of structural error on parameter constraint in a climate model. Earth Syst Dyn. doi:10.5194/esd-2016-17.

McWilliams JC (2007) Irreducible imprecision in atmospheric and oceanic simulations. Proc Natl Acad Sci 104:8709–8713

Meehl PE (1997) The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In: Harlow LL, Mulaik SA, Steiger JH (eds) What if there were no significance tests? Erlbaum, Mahwah, pp 393–425

Moncrieff MW, Liu C, Bogenschutz P (2017) Simulation, modeling, and dynamically based parameterization of organized tropical convection for global climate models. J Atmos Sci 74:1363–1380

Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change situations. Nature 430:768–772

Oreopoulos L, Mlawer E (2010) The Continual Intercomparison of Radiation Codes (CIRC): assessing anew the quality of GCM radiation algorithms. Bull Am Meteorol Soc 91:305–310

Oreskes N, Shrader-Frechette K, Belitz K (1994) Verification, validation, and confirmation of numerical models in the earth sciences. Science 263:641–646

Outten S, Thorne P, Bethke I, Seland Ø (2015) Investigating the recent apparent hiatus in surface temperature increases: 1. Construction of two 30-member earth system model ensembles. J Geophys Res: Atmos 120:8575–8596

Parker WS (2011) When climate models agree: the significance of robust model predictions. Philos Sci 78:579–600

Po-Chedley S, Fu Q (2012) Discrepancies in tropical upper tropospheric warming between atmospheric circulation models and satellites. Environ Res Lett 7:044018

Popper KR (1959) The logic of scientific discovery. Hutchinson, London

Popper KR (1963) Conjectures and refutations: the growth of scientific knowledge. Harper & Row, New York

Räisänen J (2007) How reliable are climate models? Tellus 59A:2–29

Reiss J (2015) A pragmatist theory of evidence. Philos Sci 82:341–362

Robinson AP, Froese RE (2004) Model validation using equivalence tests. Ecol Modell 176:349–358

Robinson AP, Duursma RA, Marshall JD (2005) A regression-based equivalence test for model validation: shifting the burden of proof. Tree Physiol 25:903–913

Rougier J, Goldstein M (2014) Climate simulators and climate projections. Ann Rev Stat Appl 1:103–123

Sakamoto TT, Komuro Y, Nishimura T, Ishii M, Tatebe H, Shiogama H, Hasegawa A, Toyoda T, Mori M, Suzuki T, Imada Y, Nazawa T, Takata K, Mochizuki T, Ogochi K, Emori S, Hasumi H, Kimoto M (2012) MICRO4h—a new high resolution atmosphere-ocean coupled general circulation model. J Meteorol Soc Japan 90:325–359

Schmidt GA, Sherwood S (2015) A practical philosophy of complex climate modelling. Eur J Philos Sci 5:149–169

Schwartz SE (2004) Uncertainty requirements in radiative forcing of climate change. JAWMA 54:1351–1359

Shepherd TG (2014) Atmospheric circulation as a source of uncertainty in climate change projections. Nat Geosci 7:703–708

Smith LA (2002) What might we learn from climate forecasts? Proc Natl Acad Sci 99:2487–2492

Soon W, Baliunas S, Idso SB, Kondratyev KY, Posmentier ES (2001) Modeling climatic effects of anthropogenic carbon dioxide emissions: unknowns and uncertainties. Clim Res 18:259–275

Spencer RW, Braswell WD (2011) On the misdiagnosis of surface temperature feedbacks from variations in Earth’s radiant energy balance. Remote Sens 3:1603–1613

Staniforth A, Thuburn J (2012) Horizontal grids for global weather and climate prediction models: a review. Q J R Meteorol Soc 138:1–26

Stephens GL, O’Brien D, Webster PJ, Pilewski P, Kato S, Li J-l (2015) The albedo of Earth. Rev Geophys 53:141–163

Steppuhn A, Micheels A, Bruch AA, Uhl D, Utescher T, Mosbrugger V (2007) The sensitivity of ECHAM4/ML to a double CO_2 scenario for the late Miocene and the comparison to terrestrial proxy data”. Glob Planet Change 57:189–212

Stevens B (2015) Rethinking the lower bound on aerosol radiative forcing. J Clim 28:4794–4819

Stevens B, Bony S (2013a) What are climate models missing? Science 340:1053

Stevens B, Bony S (2013b) Water in the atmosphere. Phys Today 66:29–34

Stott P, Good P, Jones G, Gillett N, Hawkins E (2013) The upper end of climate model temperature projections is inconsistent with past warming. Environ Res Lett 8:014024

Stouffer RJ, Manabe S (2017) Assessing temperature pattern projections made in 1989. Nat Clim Change 7:163–165

Sun D-Z, Yu Y, Zhang T (2009) Tropical water vapor and cloud feedbacks in climate models: a further assessment using coupled simulations. J Clim 22:1287–1304

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498

Tebaldi C, Knutti R (2007) the use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc A 365:2053–2075

Thorne P, Outten S, Bethke I, Seland Ø (2015) Investigating the recent apparent hiatus in surface temperature increases: 2. Comparison of model ensembles to observational estimates. J Geophysl Res: Atmos 120:8597–8620

Thuburn J (2008) Some conservation issues for the dynamical cores of NWP and climate models. J Comput Phys 227:3715–3730

Trenberth KE (2015) Climate change: has there been a hiatus? Science 349:691–692

Wang Z, Zhang X, Guan Z, Sun B, Yang X, Liu C (2015) An atmospheric origin of the multi-decadal bipolar seesaw. Sci Rep 5:8909

Wegener A (1966) The origin of continents and oceans (Biram J, trans.). Courier Dover p 246.

Wilcox LJ, Highwood EJ, Dunstone NJ (2013) The influence of anthropogenic aerosol on multi-decadal variations of historical global climate. Environ Res Lett 8:024033

Williams M (2001) Problems of knowledge: a critical introduction to epistemology. Oxford University Press, Oxford

Winter CL, Nychka D (2010) Forecasting skill of model averages. Stoch Env Res Risk A 24:633–638

Xiao H, Gustafson WI Jr, Wang H (2014) Impact of subgrid-scale radiative heating variability on the stratocumulus-to-trade cumulus transition in climate models. J Geophys Res: Atmos 119:4192–4203

Zhou Z, Xie S (2015) Effects of climatological model biases on the projection of tropical climate change. J Clim 28:9909–9917

Zhou L, Zhang M, Bao Q, Liu Y (2015) On the incident solar radiation in CMIP5 models. Geophys Res Lett 42:1930–1935

Zhou C, Zelinka MD, Klein SA (2016) Impact of decadal cloud variations on the Earth's energy budget. Nat Geosci. doi:10.1038/ngeo2828