Building confidence in climate model projections: an analysis of inferences from fit

Christoph Baumberger,1* Reto Knutti2 and Gertrude Hirsch Hadorn1

Edited by Eduardo Zorita, Domain Editor, and Mike Hulme, Editor-in-Chief

Climate model projections are used to inform policy decisions and constitute a major focus of climate research. Confidence in climate projections relies on the adequacy of climate models for those projections. The question of how to argue for the adequacy of models for climate projections has not gotten sufficient attention in the climate modeling community. The most common way to evaluate a climate model is to assess in a quantitative way degrees of ‘model fit’; that is, how well model results fit observation-based data (empirical accuracy) and agree with other models or model versions (robustness). However, such assessments are largely silent about what those degrees of fit imply for a model’s adequacy for projecting future climate. We provide a conceptual framework for discussing the evaluation of the adequacy of models for climate projections. Drawing on literature from philosophy of science and climate science, we discuss the potential and limits of inferences from model fit. We suggest that support of a model by background knowledge is an additional consideration that can be appealed to in arguments for a model’s adequacy for long-term projections, and that this should explicitly be spelled out. Empirical accuracy, robustness and support by background knowledge neither individually nor collectively constitute sufficient conditions in a strict sense for a model’s adequacy for long-term projections. However, they provide reasons that can be strengthened by additional information and thus contribute to a complex non-deductive argument for the adequacy of a climate model or a family of models for long-term climate projections. © 2017 The Authors. WIREs Climate Change published by Wiley Periodicals, Inc.

How to cite this article:
WIREs Clim Change 2017, 8:e454. doi: 10.1002/wcc.454

INTRODUCTION

There is now broad scientific consensus that the Earth’s climate has changed significantly over the last century, that much of the observed large-scale warming can be attributed to greenhouse gas emissions associated with human activities, and that these trends will continue in the near future.1 Less clear is how exactly the climate will change in the more distant future on the global and on regional scales if greenhouse gas emissions increase, stabilize or decrease in particular ways. Climate predictions that are conditional on forcing scenarios are called climate projections. They tell us how one or more climate characteristics would evolve if greenhouse gas concentrations and other external forcings were to follow specified pathways in the future. Projections are currently a major focus of climate research, which is undertaken (and funded) in part with the

*Correspondence to: christoph.baumberger@usys.ethz.ch
1Institute for Environmental Decisions, ETH Zurich, Zurich, Switzerland
2Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
Conflict of interest: The authors have declared no conflicts of interest for this article.
aim of providing input to policy decisions for mitigation and adaptation to anthropogenic climate change. To project climate change, climate scientists inevitably have to rely on complex numerical climate models. But to what extent can we trust model-based climate projections?

Climate model projections cannot be directly evaluated. Most of them have forcing scenarios that are never realized; and in contrast to weather forecasts, they are generally too long term to allow repeated direct testing against observational data. Climate projections need therefore to be evaluated indirectly by assessing whether the models that generate them are adequate for this purpose. The question how we can argue for the claim that a model is adequate for making climate projections of the desired kind has not gotten sufficient attention in the climate modeling community.

The most common way to evaluate a climate model (or at least the one most commonly reported) is to assess in a quantitative way degrees of ‘model fit’; that is, how well model results fit observation-based data and how well they agree with results of other models or model versions within model inter-comparison projects. However, such assessments of the empirical accuracy and robustness of model results are largely silent about what those instances of fit imply for the trustworthiness of model applications. Less than two of the roughly ninety pages of the Intergovernmental Panel on Climate Change (IPCC) chapter on model evaluation of the most recent report from Working Group I^2 deal with the implications of model fit for climate projections. It is often assumed that the fact that state-of-the-art climate models reproduce many important features of current and past climate reasonably well warrants increased confidence in the model’s suitability for quantitative projections, particularly at continental scales and above; and the robustness of climate model projections is typically seen to warrant a further increase in confidence in the projected outcome.2–4 However, the arguments for these assumptions are hardly ever made explicit, and successful instances of model fit are often uncritically interpreted as confirming the models as such. But so-called model performance metrics are simply numbers that quantify the agreement between model results and observation-based data, and it is an open question what they imply for the model’s adequacy for a purpose.5,6 It seems fair to say that it is the transition from statements about quantitative measures of model fit (‘model performance metrics’) to hypotheses about a model’s adequacy for projecting future climate change (‘model quality metrics’) where climate science struggles.7–9 Or in simple words, which things need to be ‘right’ in a model, and how ‘right,’ so that we rely on the model to predict a certain quantity of interest?

In this article, we provide a conceptual framework for discussing the evaluation of a model’s adequacy for long-term climate projections. To develop this framework, we firstly argue that model evaluation must be specific to the purpose the model is intended to serve. We then discuss the potential and limits of inferences from model fit; that is, from empirical accuracy and from robustness of model results. Moreover, we suggest that support of a model by background knowledge is an additional consideration that we can appeal to in arguments for a model’s adequacy for climate projections. Empirical accuracy, robustness and support by background knowledge neither individually nor collectively constitute sufficient conditions in a strict logical sense for a model’s adequacy for long-term projections. However, they provide reasons that can be strengthened by additional information and thus contribute to a non-deductively strong argument for the adequacy of a climate model for long-term climate projections. Finally, we summarize the suggested framework (see Box 3 at the end of the article) and discuss how it can be applied to purposes other than long-term projections and areas outside climate. (Box 1 provides some terminological clarifications.)

**BOX 1**

**EVALUATION, VALIDATION, CONFIRMATION, VERIFICATION**

We use ‘evaluation’ as an umbrella category for all kinds of epistemic assessments of models, model parts and model or simulation results.

- The kind of model evaluation we are primarily interested in is the assessment of whether a model (or a family of models) is adequate for a particular purpose. The purpose we focus on are climate projections; model evaluation is the basis for assessing which degree of confidence (or belief) in these projections is warranted.

- ‘Model validation’ is sometimes used in the way we use ‘model evaluation.’ Winsberg,^10 for example, defines model validation as the process of determining whether a model is a good enough representation of its target system for the purpose of the model. Whereas
climate modelers prefer ‘evaluation,’ hydrologists often use ‘validation.’

- ‘Model confirmation’ is often restricted to the assessment of a model’s empirical accuracy and thus to the question of how well a model is supported by data (Ref 12, p. 967). If confirmation is understood in this narrow sense, then one of our key claims is that model evaluation involves more than model confirmation, namely considerations of robustness and of coherence with background knowledge.

- In general philosophy of science, ‘verification’ of a hypothesis is the demonstration of its truth; the opposite is falsification. In climate science, modeling communities and the philosophy of computer simulation, ‘verification’ is typically understood as the determination of whether the output of the simulation approximates the true solutions to the differential equations of the original model. While verification in the first sense is out of reach for climate models and fundamentally any model that describes an open system, verification in the second sense is not our focus.

Evaluation, validation and confirmation come in degrees and can be expressed in different ways, for example as:

- **Comparative evaluation:** an increase in confidence in a model projection or a model’s adequacy for such a projection is (not) warranted.

- **Absolute evaluation:** the warranted confidence in a model projection or a model’s adequacy for such a projection is (not) above some threshold, or (not) sufficient, for example, for using the projection for a particular decision.

**Adequacy for Purpose**

**The Need for Purpose-specific Evaluation**

Climate models are representations of aspects of the climate system. They are based on fundamental physical laws (e.g., fluid dynamics and conservation laws), which are expressed as mathematical equations. Since these equations are analytically intractable to solve, high performance computers are used to approximate solutions numerically on a spatial and temporal grid. To do so, the equations must be discretized and turned into a computer code. Some processes (such as cloud formation) are important for the model results but cannot be explicitly represented, since they either occur on a smaller scale than can be resolved in the model, or are too complex to be modeled in detail, or are not well-enough understood physically. They are included via a parameterization that estimates the net effect of such unresolved processes from large-scale quantities available. For example, the cloud cover and its effect on radiation can be estimated from atmospheric humidity, vertical stability, temperature and particles available for condensation, but without actually simulating how clouds form. The climate model consists of the computer code or the discretized equations and assumptions it encodes. A model simulation consists of running the model within a particular computational environment and with specific initial and boundary conditions (Ref 4, p. 152). Initial conditions specify the state of the climate system at the beginning of the simulation. Boundary conditions describe the external forcings and factors that affect the climate system but are not directly simulated by the model. They are the drivers of climate change and include, for example, concentrations of greenhouse gases in the atmosphere at a given time, the amount of aerosols and the amount of solar radiation received by the earth.

As any numerical model of a complex system, climate models represent their target in an idealized way. An idealization is a deliberate simplification of something complicated in order to make it more comprehensible and tractable. Two types of idealizations have received much attention in philosophical debates: so-called Aristotelian and Galilean idealizations. Aristotelian idealization amounts to abstracting from all properties that are believed not to make a difference to the essential character of the system (e.g., describing planets as objects only having shape and mass). Galilean idealizations involve deliberate distortions into models in order to make them mathematically and computationally tractable (e.g., describing objects as point masses moving on frictionless planes).

Both types of idealizations are ubiquitous in climate modeling. They are involved in the construction of conceptual models that represent the relevant components of the climate system and their interaction in a qualitative way, as well as in the quantitative formulation of model equations, their discretization and their implementation into a computer. Even complex global circulation models (GCMs) that represent a wide range of atmospheric, oceanic, sea ice and land processes leave out many features of the climate system (e.g., biogeochemical, biological or ecological processes) and distort in different ways processes and aspects represented to make them mathematically...
and computationally tractable. GCMs include simplifying assumptions at the level of basic physical principles, for example, the continuum assumption that is necessary to apply the most common mathematical equations of fluid dynamics. This assumption ignores the fact that physical matter is made up of individual particles that collide with one another, and assumes that properties such as pressure and temperature vary continuously. Most obvious are idealizations at the level of parameterizations that account for the effect of processes that are too complex, too small, or not well-enough understood to be modeled in detail. In some cases, a theory is available but it would be too computationally expensive to fully incorporate it into the model (e.g., atmospheric radiative transfer); in other cases, no adequate general theory exists and the parameterizations need to be based on empirical measurements (Ref 4, p. 152).

Aristotelian idealizations lead to incomplete representations, Galilean idealizations to models that partly misrepresent their target system. Galilean idealization is often seen as merely pragmatic and bound up with the expectation that advances in computational power, physical understanding and mathematical techniques should lead to de-idealization and more accurate models (Ref 18, p. 261; Ref 20, p. 641). There is indeed a tendency to work towards less idealized models by removing simplifying assumptions, adding back complexities and replacing parameterizations with explicit representations. However, such efforts come up against their boundaries since computer resources, physical understanding and mathematical techniques will always be limited. Even within these limits, models may get too expensive to operate, too tedious to maintain and too complicated to handle and to understand. While some idealizations are simply the best science can currently do, others are preferable to more accurate representations of their target, for example, because they enhance our understanding of the target by highlighting those factors that make a difference to the essential character of the system.21–23

The aim in climate modeling—as in modeling in general—is thus not (and cannot be) to arrive at a complete representation of the climate system that is correct in all details. The aim is rather to construct models that represent processes of the climate system in ways that make the model adequate for specific purposes.9,11,24 These purposes involve answering a limited range of questions about the target system, for example, under a given forcing scenario, would Earth’s global mean surface temperature (GMST) in 2100 be more than 2°C warmer than it was in a pre-industrial state? To what extent is the global warming of the last fifty years due to human causes? Does climate change affect extreme summer precipitation over the Alps?25 Whether idealizations are adequate depends on the question about the target system we want to answer. The evaluation of climate models needs thus to be specific to the purposes for which they are used because these purposes determine which processes and aspects a model has to represent, on what spatial and temporal scale, and how accurately.

How to Conceive of Adequacy for Purpose
The idea to assess a model’s adequacy for a given purpose has been advocated for models in general.26,27 In connection with climate models, the idea is often mentioned (e.g., Ref 16); it has been defended by Parker9 and Knutti11 and criticized by Katzav.28 But what does it mean that a model is adequate for a purpose?

The purpose of a model is to provide information about the target system that allows model users to formulate hypotheses that satisfactorily answer their questions about the target system. These hypotheses are of different types (e.g., projection versus explanation), concern different climate variables (e.g., temperature versus precipitation) or events (e.g., storms versus sea-level rise) on different temporal (short- versus long-term) and spatial scales (global versus regional), are more or less specific (e.g., trends versus absolute numbers) and allow for different error margins.

A particular climate model can be adequate for some purposes, but not for others. For instance, a model may adequately predict within a specified error margin the GMST increase by 2100 relative to certain initial conditions under a certain forcing scenario, but it may not be adequate for predictions of changes in precipitation patterns in the Mediterranean area between 2050 and 2100. There is widespread consensus that projections are more reliable for temperature than for precipitation and most other quantities, for longer time averages, larger spatial averages, low specificity, and, all other things being equal and for absolute changes, for shorter lead-times (Ref 12, p. 970). The mentioned criteria for specifying a purpose are crucial to determine whether a model that is adequate for a particular purpose is also adequate for another purpose. This will be the case if the purposes are similar in the relevant respects.

A climate model can be adequate for a range of purposes, but often different purposes require different types of models. A simple energy balance model
provides a good qualitative understanding of the greenhouse effect and a reasonable estimate of GMST trends; simple models are also used to explore scenarios and for probabilistic projections; Earth models of intermediate complexity are often used for paleoclimate simulation that extend over thousands of years; high resolution GCMs and Earth system models (ESMs) are used to simulate climate change over the 20th and 21st century; and regional climate models with an even higher resolution of 10 to 20 km (compared to the typical 100 to 300 km resolution of GCMs and ESMs) are used for climate projections on a regional or even local scale. Moreover, even for a particular purpose, different models can be adequate.

A model is adequate for a purpose in the relevant sense if it conveys information about the target system that allows model users to formulate hypotheses that correctly answer their questions about the target, not by chance or luck, but because the model has properties that make it suitable for the purpose at issue (Ref 9, p 236, fn. 6). But what are those properties?

To be adequate for long-term projections, a climate model must represent sufficiently well (precise, detailed) all physical, chemical and biological processes of the climate system that are relevant for the evolution of the climate variables of interest at space and time scales of interest. This means that idealizations must be such that they do not substantially affect the results (relative to a perfect complete model) for the purpose at issue. We do not require that the model actually resolves all relevant climate processes, but if it incorporates unresolved processes via parameterization, it becomes more questionable whether the model represents the processes accurately enough beyond the range where data is available. Whether processes can be explicitly resolved or need to be parameterized depends (among other things) on the spatial resolution, which needs to be appropriate for the projection in question.

However, a model may represent all climate processes underlying the projection of interest fairly well at an appropriate spatial resolution, but may still not be adequate to be used for this projection. For instance, the model may be too computationally costly to be run on available computers and does thus not allow the model users to formulate hypotheses that answer their question about the target system. Thus, besides properties of models that are important for epistemic reasons, there are also properties that meet practical requirements. When constructing a model, we therefore inevitably face trade-offs in the performance or quality of different aspects or properties of the model (Ref 37, p. 797). The relevance of each of these properties for the purpose of the model provides a reasonable criterion for how to decide on trade-offs.

In applied contexts, the hypotheses investigated serve to inform policy decisions and need thus provide the information needed by society in order to properly inform decision-making. This requires that idealizations in model construction account for social aims and values. For instance, since considerations of justice count in decision-making, we need information about the spatial and temporal distribution of effects, rather than about aggregated effects. Criteria that guide decisions on idealizations related to the various steps in model construction need to be made explicit. Otherwise, there is a threat of intransparent influence of political interests. But also when researchers take their decisions based on criteria such as how they can handle each of the alternatives or how familiar they are with each of the options, it is necessary to explicitly consider whether these decisions are appropriate for the purpose in question.

In the remaining parts, we focus on the epistemic requirement that a model needs to represent the relevant processes sufficiently well rather than on practical and societal requirements. The next two sections examine how assessments of model fit can be used to evaluate the adequacy of a climate model for long-term climate projections.

**FIT TO DATA: EMPIRICAL ACCURACY**

The most straightforward approach to evaluate models is to assess their empirical accuracy; that is, the degree to which model results fit observed or observation-based data. Simulation results are routinely compared by means of statistical measures with the current mean state in different variables (e.g., the monthly rainfall at each location), the variability in different variables (e.g., the magnitude and timescale of the El Niño Southern Oscillation in the tropical Pacific), recent observed trends and patterns in many variables (e.g., the decline of Arctic sea ice), the response to specific perturbations (like large volcanic eruptions), or with more distant past climate states (e.g., the climate response to solar variations in the Holocene, or the climate of the last ice age, or periods even further back).  

**Reliability of Data and Fit to Data**

Climate scientists extensively deal with the questions of how well climate model results fit to data and how
reliable observation-based data are. Let us review some of the associated difficulties. To begin with, the fit of model results to data is far from perfect. Each of today’s climate models is known to give many results that do not come close to matching data within observational uncertainty and variability (i.e., the differences one would expect from random unpredictable variations like weather). Moreover, the fit is not equally good for all variables and scales. It is, for example, less good for precipitation than for temperature, and for regional than for global quantities.42–44

Furthermore, the various ways of obtaining data in the climate sciences raises a host of questions about the reliability of the data. Particularly important ones concern the theory-ladenness of observation and data, which comes in different forms (Ref 45, p. 956). First, since instruments are used to obtain data, the quality of the data depends on the reliability of the instruments and the theory behind them. The data can thus only be trusted if the working of the instruments has been independently tested and confirmed, which is indeed the case for many relevant observations,46 but is, for example, much harder for satellites that once deployed are inaccessible. Second, since the unprocessed raw data received directly from the measurement instruments often contain errors, gaps and are incomplete in many ways, models are applied to filter and to correct raw data, infer quantities (e.g., temperature and vegetation cover from irradiance at particular wavelengths) and extend them to gridded and reanalysis data sets that include values for chosen variables on regular spatial grid and at regular time intervals. That these general problems raise complex questions in the climate case is shown, for example, by the controversy about warming trends in upper troposphere from satellite and radiosonde measurements47–50 (for an overview see Ref 51). Model filtered data52,53 can be trusted to the extent to which the models used to correct and extend the data have been independently tested and are confirmed. Third, since often no direct measurements, for example of surface temperature in the distant past, are available, scientists gather proxy data of natural (or human) systems that are affected by temperature, such as ocean sediments, tree rings and ice cores, or tax records from grape harvest. The quality of the data depends on the availability of the proxy and on the reliability of the statistical methods used to process the raw proxy data and turn it into the data one is interested in, for example to turn a tree ring width or density into a temperature record. Finally, for many quantities there are only spatially sparse or short periods of observations, or none at all, because the quantity was or is too hard or too expensive to measure (e.g., the conditions at the ground of an ice sheet three kilometers thick), or because nobody cared about systematically measuring it at the time (e.g., the properties of seawater in the deep ocean before about 1950).

Despite these difficulties, it is safe to say that with each generation, models continue to better represent many aspects of the mean climate state and variability that we can observe,54,55 and the amount and quality of data improves as well. But how can assessments of the empirical accuracy of models be used to argue for their adequacy for a particular purpose?

Determining Observational Consequences
Drawing inferences from empirical accuracy of model results to the adequacy of a model for a purpose is less straightforward than it first might seem. This is so even if for the moment we ignore problems in observations. The reason is that a model need not be empirically accurate in every respect and degree to be adequate for a specific purpose. Instances of fit (or misfit), for example, do not support (or undermine) a model’s adequacy for projecting the value of some variable if they concern a quantity that is unrelated to the variable of interest.

The challenge then is to determine which instances of fit do support and which instances of misfit do undermine an adequacy-for-purpose hypothesis. This involves the challenge of determining what the model is likely to indicate about the observable properties of the target system if the model is adequate for the purpose of interest. What we are looking for, then, are conditional claims of the following form: If a model is adequate for purpose P, then it reliably indicates the values of variables V for time T with accuracy α. In order to evaluate the model’s adequacy for P, we need to check how well what is actually observed fits with the observational consequences that are likely to follow if the model is adequate for P.

Determining the observational consequences that are likely to follow if a model is adequate for projecting the values of some variable X is straightforward for the quantity of interest for past and present. It is often reasonable to claim that if a model is adequate for projecting within a specified error margin the values of X for some period in the far future (e.g., GMST for each year of the 2050s within 0.3°C under a given forcing scenario), then the model reliably indicates values of X for recent past and present with at least that accuracy. Since climate quantities
are often interdependent, an accurate simulation of the values of X will require an appropriate simulation of related quantities upon which X depends. Thus, if a model is adequate for projecting within a specified error margin the values of X (e.g., GMST) for some period, then the model reliably indicates values of Y (e.g., GM precipitation) for some period within a certain error margin. Arguing for such a conditional claim requires a great deal of analysis and physical understanding. In order to determine how accurately the model is likely to indicate the values of Y for some period if it is adequate for projecting the values of X for some period, we need to consider how particular climate processes are represented in the model and how those representations deviate from our background beliefs about how the processes actually work and depend on each other. In many cases, it may remain unclear what we should expect with regard to model performance in simulating the values of other variables if the model is adequate for the projections of interest (Ref 9, p. 241). Moreover, arguments for such conditionals may face a circularity problem. In order to determine the observational consequences with respect to different climate quantities that are likely to follow if a model is adequate for some projection, we need to rely on climate models to know how different climate quantities and processes depend on each other. In relying on the models, we would be assuming that they are largely correct and thus adequate for whatever purpose they are intended to serve (Ref 28, p. 232–233).

From Deductive to Non-Deductive Inferences

The conditional claims mentioned in the previous section roughly say that reliably indicating the values of some variables for past and present is a necessary condition for a model to be adequate for projecting X for the far future. If this necessary condition is not fulfilled, it deductively follows that the model is not adequate for projecting X (see Box 2); hence, the adequacy-for-purpose hypothesis is falsified. For example, if we learn that our climate model only rarely indicates GMST of the recent past within 0.3°C of those derived from observation-based data, we can infer that the model is not adequate for projecting GMST values for the 2050s within 0.3°C. However, if the necessary condition is fulfilled, it does not follow that the model is adequate for projecting X. For this, the condition would need to be sufficient. As we will argue below, reliably reproducing past and present climate is not a sufficient condition for a model to be adequate for projecting future climate. Hence, we can use conditionals of the mentioned types to construct a deductively sound argument against the hypothesis that a model is adequate for the projection at issue, but the suggested arguments in support of such a hypothesis are either deductively invalid or contain a false premise.

Reliably indicating past and present climate cannot ensure that a model is adequate for long-term projections, but under certain conditions, it warrants an increased confidence in those projections. Claims about the empirical accuracy of model results should therefore be understood as premises of a non-deductive argument for the model's adequacy for projections of the desired kind. Non-deductive strength is non-monotonic, that is, adding premises can yield a stronger or weaker argument (see Box 2).

**BOX 2**

**DEDUCTIVE VALIDITY AND NON-DEDUCTIVE STRENGTH**

An argument is **deductively valid** if the truth of its premises guarantees the truth of its conclusion.

- Example: If a model is adequate for projecting X for the far future, then the model reliably indicates X for past and present. Model M does not reliably indicate X for past and present. Hence, M is not adequate for projecting X for the far future.

- Deductive validity is (a) incompatible with the conclusion being false while all premises are true, (b) an all-or-nothing matter, and (c) monotonic, that is, adding premises never turns a valid into an invalid argument.

An argument is **non-deductively (e.g., inductively) strong** if its premises provide a good reason for the truth of its conclusion.

- Example: Model M reliably indicates X and climate quantities upon which X depends for past and present. So probably, M is adequate for projecting X for the near future.

- Non-deductive (e.g., inductive) strength is (a) compatible with the conclusion being false even if all premises are true, (b) a matter of degree, and (c) non-monotonic, that is, adding premises can yield a stronger or weaker argument.

A **deductively valid argument with true premises** is called *sound*; a non-deductively strong argument with true premises is usually called *cogent*.
Hence, the evaluation of such arguments needs to assess whether all relevant information has been taken into account and the conditions for increased confidence are met. Whether this is indeed the case, is often hard to decide. This makes the evaluation of non-deductive arguments difficult.

**Non-deductive Inferences from Empirical Accuracy**

Non-deductive arguments from fit of model results with data about past and present climate to a model’s adequacy for long-term projections are confronted with a number of worries.

Besides issues related to the reliability of data and the fit to data, there is the general problem of induction going back to the philosopher David Hume. Hume’s problem was roughly the worry whether nature is uniform and things continue to go on in the same ways (Ref 57, 1.3.6). The specific worry in connection with arguments for a model’s adequacy for long-term climate projections is whether the model’s success in representing past and present climate provides good reasons to assume that it is adequate for projecting future climate. Here are two instances of this worry.

The first concerns calibration or tuning. The values of the parameters involved in parameterizations are often poorly constrained by an understanding of the underlying processes and need to be calibrated against observational data. Model calibration is unavoidable in climate modeling, and routinely done, but has so far been rarely discussed and documented systematically.\(^5\)\(^8\)\(^9\)\(^4\)\(^5\)\(^6\) Model calibration consists in choosing a parameter configuration so that the model results better fit to data about past and present climate.\(^6\)\(^0\) The worry is that if the fit to data is due to calibration, then it does not provide a strong reason for the adequacy of the model for long-term projections.\(^3\)\(^4\)\(^9\)\(^5\)\(^6\)\(^1\) One reason is that parameters are often not tuned to their ‘correct’ values; calibration allows to compensate for structural errors by introducing compensating biases (e.g., in climate sensitivity and radiative forcing) during the calibration process.\(^3\)\(^9\)\(^6\)\(^2\) The calibration of a model may thus guarantee success with respect to past and present climate irrespective of whether the model correctly accounts for those underlying processes that are relevant to the long-term evolution of the climate system. A second reason why model fit that is due to tuning does not provide a strong reason for a model’s adequacy for projections is that the choice of parameters or model structure may be inconclusive given the data used in calibration. At least in some cases, there are different sets of parameter values that result in equally good fits with data.\(^5\)\(^3\)

Different models that agree in their performance as far as the dataset in calibration is concerned can disagree with respect to out-of-sample applications and thus with respect to long-term projections. For example, calibrating a model to short GMST trends provides only weak constraints for projections of future climate.\(^2\)\(^4\)\(^6\)\(^4\) Both reasons substantiate the worry that the performance of a model with respect to the future might not be similar to the performance with respect to the data to which the model is tuned. It is thus unclear whether model success can be extrapolated from past and present to the future. A priori we should not expect that it can, yet this is often done implicitly.

The familiar strategy to avoid this problem is to split data and use one part of a dataset to calibrate a model and the other part to test it.\(^3\) This has triggered a debate about whether data used in calibrating a model can nonetheless also be used in evaluating the model; that is, whether double-counting is legitimate or whether data used for evaluation should not be used in calibration and need in this sense be use-novel (see Ref 65, for an overview). Climate scientists often declare double-counting as illegitimate.\(^7\) This is an overstatement if it implies that the accommodation of data through calibration provides no epistemic support for the model. Successful calibration can confirm a model at least to some extent because it is far from trivial that a model can be successfully calibrated. The reason is that climate models can be evaluated and calibrated on a large number of variables and scales, but calibration involves usually only a limited number of parameters (Ref 24, p. 174). Philosophers, on the other hand, have argued that from a Bayesian perspective, there is no difference between calibrating and confirming and thus no problem with using the same data to calibrate and confirm a model. For the Bayesian, calibration ‘is simply the common practice of testing hypotheses against evidence’ (Ref 66, p. 615). Frisch\(^2\)\(^4\) has shown that this claim does not follow from Bayesian formalism alone, and that in case of complex climate models, fit to data not used in calibration is a superior test of the adequacy of a model for long-term projections than fit to data used to calibrate the model.\(^8\) Thus, to assess the strength of a non-deductive argument from model fit to the adequacy of the model for long-term projections, we need information about the extent to which the fit in question was dependent on tuning and whether tuned elements on which the fit depends have been tested out-of-sample (Ref 9, p. 245). The strength of
such an argument depends also on how independent the data used to calibrate the model are from the data used to test it, and on the extent to which data not used in explicit calibration guided model construction in some other way, for example, by influencing choices in the model structure and design.

The second worry is a more fundamental one. Even if climate model results fit to data that were not used in calibration, this may not provide strong reasons for a model’s adequacy for long-term projections. Long-term projections for high-forcing scenarios lie outside the scope of boundary conditions previously observed in the instrumental record. At least prima facie, there is no reason to assume that successful performance under current boundary conditions is a good guide to successful performance under future boundary conditions that describe high-forcing scenarios. We are confident that the physical principles on which climate models are based can be extrapolated beyond the range where they are evaluated. However, this is less clear for parameterizations that are empirically derived from observations mostly covering the narrow climate regime of the past century, and for the interaction between parameterizations and physical principles. For long-term projections, additional processes and feedbacks (e.g., methane emissions from thawing permafrost) may become relevant and take the system out-of-sample with respect to existing observations. If a model does not account for these processes and feedbacks, it could fit almost perfectly even to data about past and present climate not used for calibration but still be biased for projections. Success with respect to past and present climate alone is thus no assurance that the model will also be successful in projecting future climate (Ref 2, p. 828). Some climate scientists conclude from this that it is hard to tell how relevant past data are or that they are not relevant at all for evaluating a model’s adequacy for climate projections (Ref 68, p. 2146). This conclusion may be too hasty, but the considerations behind it show that to further strengthen an argument from model fit to the adequacy of the model for long-term projections, we need independent reason to assume that the model captures the relevant climate processes and feedbacks.

Worries related to calibration and missing feedbacks can also be mitigated by testing model results against data about paleoclimate epochs. Paleoclimate states provide partly independent information not used in model development, and they were driven by forcings quite different from those of modern climate. However, the boundary conditions and the data are limited and derived from proxy data, which introduces large uncertainties, and those types of data also get increasingly used in the model development and evaluation process, which weakens the argument of an independent test.

**FIT TO RESULTS OF OTHER MODELS: ROBUSTNESS**

Climate model projections are too long term to allow repeated direct comparison with data, but they can be compared with projections of other models or model versions. This is what climate modelers extensively do in ensemble studies because of uncertainties in how to represent the climate system such that models lead to accurate projections of future climate.

In an ensemble study, each of several climate models or model versions is run with the same or similar initial and boundary conditions. There are two main types of such studies (Ref 67, p. 582). Perturbed physics (or parameter) ensemble studies employ different versions of the same model that differ in the values of their uncertain parameters, that is, are effectively parameter sensitivity tests. In this way, the ensemble explores how climate projections are impacted by the uncertainty about the values that should be assigned to model parameters. Multimodel ensemble studies employ several models that differ in a number of ways, for example, in number and complexity of processes included, parameterizations, spatiotemporal resolution, numerical methods and computing platforms. In this way, the ensemble explores how climate projections are impacted by structural and parametric uncertainty; that is, uncertainty about the form that modeling equations should take and how they should be solved computationally. The most ambitious multimodel ensemble study to date is Coupled Model Intercomparison Project 5 (CMIP 5) which has collected results from about 60 models from nearly thirty modelling centres around the world. Both types of ensemble studies often include a limited investigation of the impacts of initial condition uncertainty as well by running multiple cases for the same experiments with different initial conditions.

Ensembles help to deal with uncertainties either by producing robust projections or by providing estimates of uncertainty about future climate change. A model projection is robust if all or most models in the ensemble agree regarding the projection. If all models in an ensemble show more than a 4°C increase in GMST by 2100 when run under a certain forcing scenario, this projection is robust. In what follows, we focus on multimodel ensemble studies but similar arguments can be made for perturbed
Inference to the Likely Truth of a Projection

An inference from robustness of projections to their likely truth is legitimate if we have reasons to assume that it is likely that at least one model in the multimodel ensemble correctly projects the quantity of interest within the specified error margin.

A premise to this effect could be justified in two different ways (Ref 67, p. 584–589). One is to cite the success of the models in simulating past and present climate to support the claim that it is likely that at least one simulation in the ensemble correctly projects the quantity of interest within the specified error margin. Considerations in the last section pointed out the limits of such an argument. A second way to justify the required premise refers to the construction of the models rather than to their performance. It argues that the multimodel ensemble samples enough of the current uncertainty about how to represent the climate system for the projection at issue that it is likely that at least one simulation correctly projects the quantity of interest within the specified error margin. The problem with this line of argument is that today’s multimodel ensembles group together existing models and are thus ‘ensembles of opportunity,’ not designed to span an uncertainty range (Ref 62, p. 2653). One of the main sources of uncertainty are parameterizations of subgrid-processes such as cloud formation. Each state-of-the-art climate model includes some representation of clouds, but ensemble studies do not attempt to ensure that the ensemble as a whole adequately samples (or spans) current uncertainty about how clouds should be represented; the same holds for other subgrid-processes (Ref 67, p. 585). Moreover, it is unclear how such a sampling could be achieved. In case of parameter uncertainty, the space of possibilities in which plausible alternatives are to be identified is clear since it is the space of numerical values (although it is computationally intractable due to its dimensionality), but in the case of structural uncertainty as it is addressed in multimodel ensemble studies, the space of possibilities is indeterminate since it ranges over model structures (Ref 74, p. 216). One may argue that in the presence of limited understanding and potential unknown unknowns it is fundamentally impossible to sample the uncertainty in how to build and calibrate a model since we do not really know what the uncertainty is.

Inference to the Warranted Confidence in a Projection

Similar difficulties beset sampling-based arguments from the robustness of model projections to the warranted confidence (or degree of belief) in these projections. Such arguments combine robustness considerations with additional criteria of adequacy. Suppose that $S$ is the set of all theoretically possible models that meet sufficiently well basic criteria such that each model in $S$ has a significant chance of being adequate for projection $P$ within some specified margin of error. For example, they simulate relevant aspects of past and present climate sufficiently well, include particular physical assumptions and have an appropriate spatiotemporal resolution. The models in $S$ are currently considered the best theoretically possible models for $P$. In the absence of overriding evidence, the warranted confidence in $P$ can then be identified with the fraction $f$ of models in $S$ whose simulations agree with respect to $P$ within the specified error margin. Now, if the models in an ensemble constitute a random sample from $S$, then the fraction of models in the ensemble whose simulations agree with respect to $P$ within the specified error margin provides a good estimator of $f$ and thus of the warranted degree of confidence in $P$ (cf. Ref 67, p. 593–595).

The problem with an argument along this line is that today’s multimodel ensembles are not random samples from the set of all theoretically possible models that meet basic criteria of adequacy. It is unclear what the space of possible models that meet the required criteria actually is, and climate scientists do not select today’s models from this space by randomized procedures. As ensembles of opportunity, today’s ensembles are not the kind of sample from which statisticians would usefully estimate uncertainty since their ‘sampling is neither systematic nor random’ (Ref 75, p. 2068). Currently available multimodel ensembles such as CMIP5 are not designed to systematically explore the space of models that meet the required criteria, and the statistical interpretation of the ensemble is unclear.76 All we have is a very limited space of practically possible models, and this space involves near duplicates since models are used several times with minor modifications only. Even if these duplications were eliminated and the remaining model space randomly sampled, there would still be structural dependencies between the models.54,77–79

The models are of course based on the same physical understanding and use the same basic equations, but they also partly use the same parameterizations, make similar simplifications, and use the same computational methods; in many cases, they even share...
large fractions of code. As a result, the models inevitably share common errors (e.g., in the simulation of the Inter-tropical Convergence Zone ITCZ\textsuperscript{80}). Moreover, some climate processes that will significantly influence future climate change are not represented in any of today’s models; some of these processes are recognized (e.g., the effect of methane hydrates), and perhaps some not. Both raises the worry that simulations from today’s climate models might not so frequently agree with respect to a projection, even though it is false (or biased),\textsuperscript{6} because most models share similar deficiencies. Furthermore, the interdependency of models within today’s small ensemble studies makes it likely that the models do not differ enough to provide a representative sample of the set of all theoretically possible models that meet the basic criteria of adequacy (whatever they exactly are). Given current uncertainty about how to represent the climate system adequately, the set of possible models that meet the basic criteria of adequacy is likely to include models that differ significantly from today’s models. If today’s models differ from one another much less than random samples from S would, they are biased estimators of the fraction of models in S whose simulations agree with respect to P within some error margin, and thus of the warranted degree of confidence in the projection at issue (Ref 67, p. 594). The IPCC (Ref 1, ch. 12) acknowledges this and downgrades the probabilities based on the frequency of ensemble results.\textsuperscript{6}

The robustness of model projections cannot be directly translated into probabilities without making strong assumptions about the ensemble, about dependence and criteria for adequate models.\textsuperscript{84} But to the extent that the models of an ensemble are independent and differ thus in ways that are relevant for the projections at issue, the robustness warrants an increased confidence in the projections and can thus figure as an additional premise in a non-deductive argument for a model’s adequacy for those projections. Such an argument combines premises about the robustness of model results with premises about their empirical accuracy; that is, premises about the model’s success in reproducing (use-novel) data of past and present climate with premises about the agreement of the projections with projections of other models of an ensemble that are equally successful in reproducing (use-novel) data of past and present climate. The strength of such a non-deductive argument depends on whether all relevant information has been taken into account. To assess the strength of the argument, we need therefore more information about the extent to which the models in the ensemble are independent from each other and differ thus in relevant ways, for example, regarding equations, parameterizations, parameter values, resolution, boundaries, and numerical coding.\textsuperscript{85}

**Inference to the Correctness of Causal Assumptions**

Even if a sampling-based argument from robustness were sound, it would not prove that the models generate the robust projections because the models capture the relevant climate processes sufficiently well for this purpose. However, robustness of model results (combined with their empirical accuracy) is often seen as making it likely, or at least increasing our confidence, that the processes that determine these results are encapsulated sufficiently well in the models (Ref 3, p. 979–980; Ref 4, p. 160).

A framework for arguments to this effect is provided by Weisberg’s (Ref 86; Ref 35, p. 156–70) conception of robustness analysis, which involves three steps. First, a set of models is examined in order to determine whether they all lead to the same result, the robust property. Second, the models are analyzed for a common structure that generates the robust property, and the two are linked together into a robust theorem: Ceteris paribus, if [common causal structure] obtains, then [robust property] will obtain. For climate models that approximately agree in their GMST outcomes, we roughly get: Ceteris paribus, if [greenhouse gases are causally related to the energy budget of the Earth] obtains, then [for increasing greenhouse gases, increasing GMST] will obtain (Ref 3, p. 980). In the third step, stability analysis of the robust theorem is conducted in order to determine the limits of robustness and spell out the ceteris paribus clause. It needs to be determined how frequently the common structure shows up within ‘a sufficiently heterogeneous set of models’ (Ref 86, p. 739). And it needs to be investigated what defeats the core structure giving rise to the robust property, which is done by varying the values for parameters or by adding or removing mechanistic features from the model in multimodel ensemble studies. According to this proposal, we can establish a robust theorem by running simulations of a target system using models that differ systematically, but share common causal assumptions. If the property shows up in most of the simulations, then the robust theorem is established. One can then confirm the common causal assumptions of the models by observing the relevant property (Ref 86, p. 739). Thus, if the climate models within a multimodel ensemble are sufficiently heterogeneous and differ systematically, but share common causal assumptions about the climate system (e.g., a
common causal structure of greenhouse gas causation), lead to the same robust property (e.g., GMST model outcomes for past and present approximately agree) and the robust property obtains in the target system (GMST has approximately increased as described by the models), then we can conclude that it is very likely that the common causal assumptions of the models are approximately true (i.e., greenhouse gas emissions are the relevant cause of global warming).

As already seen, the assumption that the models are sufficiently heterogeneous and differ systematically does not seem to be true for today’s multimodel ensembles. More importantly, even if we can argue along the suggested line for the attribution of global warming up to now, we cannot argue in this way for the claim that the models of an ensemble represent sufficiently well the climate processes that are relevant for long-term climate projections. The reason is that confirming the common causal assumptions of models requires observation-based data for the robust property, but we only have such data for robust results about past and present climate. Whether a model captures sufficiently well the climate processes that are relevant for the projections of interest needs rather to be shown with reference to background knowledge.

BEYOND FIT: BACKGROUND KNOWLEDGE

Climate scientists often stress that our confidence in climate model projections does not only come from the empirical accuracy and robustness of simulation results, but also from the foundation of the underlying models in accepted theoretical principles and the physical, chemical or biological understanding of the processes behind the results (Ref 2, p. 745).

Process understanding or, more generally, coherence with background knowledge is indeed a key consideration in determining the adequacy of a climate model for specific long-term projections. Implications of model fit for long-term projections are limited, but if we can argue independently from model fit that the relevant climate processes for a particular purpose are well observed and likely to be understood beyond the range observed, or even based on fundamental principles like conservation laws, and further that the model represents those processes sufficiently well, then a model’s success in representing aspects of past and present climate provides good reasons to assume that it is adequate for projecting the quantity of interest in a future climate. Background knowledge can be used to decide which processes a model needs to represent, and coherence with background theories and assumptions that embody a theoretical understanding of the processes at issue provides reasons that are independent from model fit. Claims about how well a climate model is supported by independently confirmed background theories and assumptions can thus figure as further premises in a non-deductive argument for the model’s adequacy for the purpose of projecting future climate. But the supporting role of background knowledge is limited by the need of empirical parameterizations and the epistemic opacity7 of complex models.

Problem of Empirical Parameterization

It greatly contributes to our confidence in model projections that climate models are based on fundamental physical principles (e.g., conservation of mass and energy) and that many of their equations are derived from well-confirmed physical theories (e.g., fluid dynamics). However, the extent to which theory can guide climate model evaluation (and construction) is limited. Besides well-accepted physical principles and approximations to well-understood physics (in cases where a physical theory is available but it would be computationally too costly to fully incorporate it into the model), climate models contain also empirical parameterizations of unresolved processes for which no general theory exists. Parameterization is even essentially involved in representations of processes for which we possess basic equations, since the discretization of the equations requires grid-scale dependent parameterization of unresolved subgrid processes (e.g., a parameterization of atmospheric convection). While physical principles can certainly be extrapolated beyond the range where the model is tested, this is less clear for parameterizations and their interactions with physical principles.

However, parameterizations like those for atmospheric convection can be independently tested by comparing them with specifically targeted observation-based data covering different regions and times (e.g., from aircraft campaigns and ground data measuring interesting weather situations), conducting experiments, varying parameter values, incorporating the parameterizations into other types of models (e.g., weather models which have daily verification) to test them in a different context and on a different time scale, and by basing them or evaluating them on high resolution models that explicitly resolve the involved processes (see Ref 88, for an overview). Some of these strategies raise questions similar to
those already discussed, but the hope is that submodels of individual processes are easier to evaluate because these processes only work on certain scales, are linked to a relatively small number of physical, chemical or biological processes, can more directly be constrained by observations and allow for experimental testing under controlled conditions. Nonetheless, if parameterizations are empirically derived from data covering the last century, this independent support does not ensure that the parameterizations are inductively stable and can be extrapolated beyond the range of where they were evaluated. For this, they need be based on (or at least loosely inspired by) physical, chemical or biological principles, which requires better observations and an improved understanding of the (mostly subgrid scale) processes in question (e.g., for cloud formation about how aerosols affect clouds and about the physics surrounding precipitation and clouds). Improving this understanding is particularly important but also difficult for processes that in the current climate state and within the limited time we have observed it, may be barely detectable (e.g., rapid calving of a large ice sheet, or the collapse of a marine ecosystem). In other words, parameterizing the unknown is both a dangerous and hopeless exercise, yet leaving it out will result in all models being robustly wrong.

**Problem of Epistemic Opacity**

Support of a model by background knowledge in general is also limited because in case of complex climate models, the understanding of how model components (physical principles, parameterizations, initial and boundary conditions) contribute to the various metrics of performance of the model is limited. As a result, it is difficult or even impossible to say where the successes and failures of climate models to reproduce data come from, and what they imply for a particular projection. Lenhard and Winsberg⁹¹ argue that two features of complex climate models exacerbate or prevent such an analytic understanding of the models. First, the models are holistic in the sense that the effect of components and submodels can often not be tested in isolation because they are highly interactive and the complexity of their interaction makes it often impossible to independently assess the merits or shortcomings of each component, sub-model, process or parameterization. Second, the models contain ‘kludges’; that is, unprincipled and complex pieces of programming that appear to work but are ill understood. Thus, even though climate models are based on well-established physical principles and theories that allow at least some understanding of the underlying climate processes, it is often impossible to fully understand the details of how model behavior emerges from the interaction of different model components (Ref 24, p. 178). This lack of analytic understanding of models (and in fact the target system as well) makes it difficult to assess whether a models’ success in representing certain aspects of past and present climate is a good reason to assume that the model is also adequate for projecting certain aspects of future climate. As a consequence, there is often no consensus of which instances of fit with observation-based data provide support for a projection.

There are strategies to deal with this problem. The first assumes that the analytic opacity of climate models does not preclude evaluating representations of individual processes in the context of the full model as well as in isolation (Ref 2, p. 754). At least in certain cases, it seems possible, by comparing individual and combined runs, to achieve an understanding of the effect of components or sub-models within the model and to attribute success or failure to certain components or sub-models (Ref 12, p. 967). Unfortunately, a parameterization that performs better when tested individually does not necessarily improve the performance of the model as a whole. If a model is biased with respect to aerosol concentration or humidity, then an improved parameterization of cloud formation may lead to a poorer performance of the model as a whole. This raises the question whether model construction should give priority to the optimization of model fit of the whole model, or strive for a more faithful representation of climate processes even though this might in the near term lead to a deterioration in model performance.

The second strategy is to pursue an ensemble approach and experiment on models and explore model hierarchies in order to advance the understanding of the behavior of the model and its target system. Experimenting with models means investigating the contribution of different processes and parameters in producing the (simulated) phenomena of interest by eliminating or adding particular processes and varying the values of parameters and comparing the simulations produced with and without these interventions. For example, one could include a realistic topography in a climate model in order to learn whether this makes much difference to the projection one is interested in.⁹² Held⁹² suggests that climate scientists should more extensively explore hierarchies of models that range from simplified models that represent some key causal factors in highly idealized ways to more and more complex models which include representations of additional causal factors...
and/or more realistic representations of previously included factors. Studying such hierarchies and tracing certain behaviors across different types of models might provide a better understanding of the interaction of different model components and sub-models.

The third strategy is to evaluate climate model results against use-novel data that have not been used in calibrating the model, in the cases where such data exist at all. As Frisch\textsuperscript{24} argues, in case of complex climate models that allow only a limited analytic understanding, fit with use-novel data provides stronger reasons of a model’s adequacy for projecting future climate than fit with data used in calibration. If a model is epistemically opaque and we do not know in detail how the more principled components of the model interact with parameterizations to result in the model’s output, then successful simulation of use-novel data can be an indicator that the model adequately represents the relevant climate processes and does not posit spurious relations among data introduced through tuning.

A fourth and related strategy is to exploit an ensemble of models to find strong relationships between well observed quantities and the projection of interest, so-called emergent constraints (Ref\textsuperscript{2}, p. 826; Ref\textsuperscript{11}). However, for the emergent constraints the coherence with background knowledge remains important to avoid the interpretation of spurious relationships.\textsuperscript{93} It also is critical for the use-novel data, since performance with respect to any amount of data from the past and present alone cannot, as discussed, show that the model captures all relevant processes and feedbacks that will significantly influence future climate change.

CONCLUSION

From our argumentation, a conceptual framework for discussing the adequacy of a model or family of models for long-term climate projections emerges (see Box 3). The evaluation of models needs to be specific to the purpose they are intended to serve. Models are idealized representations and it depends on the purpose whether idealizations are adequate. To be adequate for long-term projections, a climate model needs to represent sufficiently well those processes that significantly shape the long-term evolution of the climate characteristics of interest. An argument for the hypothesis that a model or family of models is adequate for a long-term climate projection can draw on the assessment of models or their results with respect to three dimensions: the empirical accuracy of model results (i.e., their fit to data), the

**BOX 3**

**FRAMEWORK FOR DISCUSSING THE EVALUATION OF THE ADEQUACY OF CLIMATE MODELS FOR LONG-TERM PROJECTIONS**

| Premises of a non-deductive argument for a model’s adequacy for projections | Strength of the resulting argument |
|---|---|
| Dimensions of evaluation | Direct objects of evaluation | Limiting factors | Ways to increase strength |
| Empirical accuracy | Model results | Model calibration | Data splitting |
| Robustness | Model results | Additional feedbacks that shape future climate | Background knowledge about relevant feedbacks to pick relevant metrics |
| Coherence with background knowledge | Model structure and model components | Lack of model independency and diversity | Increasing model independency and diversity |
| — Basic theories | — Model equations | Non-theory based parameterizations | Basing parameterizations on models that explicitly resolve the involved processes |
| — Phenomenological theories, empirical assumptions | — Parameterizations, initial + boundary conditions | Epistemic opacity of model | Improving model transparency by exploring model hierarchies |
robustness of model results (i.e., their fit to results of other models or model versions), and the coherence of model components with background knowledge (i.e., the support of model equations by basic theories and of parameterizations and initial and boundary conditions by phenomenological theories and empirical assumptions).

A certain degree of empirical accuracy, robustness and coherence with background knowledge is necessary for a model’s adequacy for long-term projections, but they neither individually nor collectively provide sufficient conditions for the adequacy of a model for such projections in a strict logical sense. The empirical accuracy of model results may to some degree be due to tuning rather than to the model’s representation of the processes that shape past and present climate (although climate scientists would deny that the degree of tuning questions our confidence in model projections, and that models are tuned to simulate a specific future projection). Even if the model represents these processes well enough, it may lack representations of feedbacks that will significantly influence future climate change at space and time scales of interest. Robustness considerations do not solve this second difficulty and mitigate the first only to the extent to which the models in question are sufficiently heterogeneous. Background knowledge that supports model components provides independent reasons to assume that the model represents sufficiently well the processes relevant for the long-term climate projection at issue. But the role of background theories is limited because of empirical parameterizations, and support of model components by background knowledge more generally faces the problem of the epistemic opacity of complex models. While such background knowledge is implicitly considered when interpreting results and assessing confidence, and provides often a strong argument for model results, its contribution is hard to express in numbers and not as easily conveyed as ‘beauty-contest like’ instances of fit. As a consequence, unfortunately, background knowledge is rarely explicitly discussed in scientific publications, and instances of model fit are often shown without arguing how they support a conclusion.

Under certain conditions, empirical accuracy, robustness and coherence with background knowledge enhance our confidence in model projections and provide thus premises for a non-deductive argument for the claim that a model is adequate for those projections. In order to assess the strength of such an argument, more information is needed, for example, about the extent to which the fit to data was dependent on tuning, about the extent to which the models that agree in their projections are independent from each other and differ thus in relevant ways, and about the extent to which parameterizations are based on models that explicitly resolve the involved processes.

The suggested framework provides not only a basis to assess evaluation arguments, but indicates also how such arguments can be strengthened by future research. First, one can work towards a better performance of models with respect to each of the three dimensions; that is, better fit of model results to data and to results of other models, and better coherence of model components with further developed background knowledge. Since resources are limited and improved model components do not necessarily lead to better model fit, this may require deciding whether priority should be given to the optimization of model fit or to a more faithful representation of climate processes. Second, one can try to increase the significance that the performance with respect to each dimension has for a model’s adequacy for long-term projections. The significance of empirical accuracy can be increased by splitting data and testing the model against data that have not been used to develop and calibrate the model, and by arguing independently that the model captures well enough the relevant climate processes. Only then does success of a model with respect to past and present provide reasons to assume that the model will also be successful with respect to future climate. The significance of robustness can be enhanced by increasing model independency and diversity. The significance of coherence of model components with background knowledge can be enhanced by improving model transparency (i.e., the understanding of how model components contribute to model fit) through the exploration of model ensembles and model hierarchies.

The suggested framework can be applied to other purposes than long-term projection (e.g., short-term prediction, detection, or attribution) and in areas outside climate. We hypothesize that the three dimensions with respect to which the adequacy of models can be evaluated are the same for different purposes and different complex empirical areas in which predictions cannot directly be evaluated, but the weight given to the different dimensions varies for different purposes. For example, whereas climate models are based on well-confirmed physical theories, there are no generally accepted basic theories in the social sciences. Even if this is harmless for short-term predictions, it hampers long-term predictions. Model fit may be decisive for short-term predictions that can repeatedly be tested against data, but support by background theories that embody an understanding of the underlying processes is important for long-term predictions since we have reason to assume that
extrapolation of trends is not reliable and calibrated parameterizations may no longer be valid on larger timescales. The challenge for scientists is on the one hand to borrow strengths from the three dimensions in a way that is optimal for their case, in order to maximize the confidence in projections made. On the other hand, it is to explicitly spell out how the fundamental theories, the assumptions, the empirical accuracy, the degree of robustness and the elements of background knowledge complement each other, in order to transparently document the line of argument supporting the projections, and to convince those who rely and make decisions based on them.

NOTES

a These notions of verification and falsification play a key role in the philosophy of logical empiricism13 and its criticism by critical rationalists,94 but they are also used in the literature on model evaluation (see e.g., Ref 15, p. 641; Ref 16, p. 577).

b In philosophy of science, this requirement is often expressed in terms of relevant similarities between the model system that the model equations describe and the target system. It then requires that the model system be, for the purpose of interest, in relevant respects and to a sufficient degree similar to the target system26,27,34,35 (for a critical voice see Ref 36).

c Katzav even argues that the difficulties are in typical cases unsurmountable (Ref 28, p. 232). Gleckler et al. (Ref 42, p. 17) share this worry: ‘it remains largely unknown what aspects of observed climate must be correctly simulated in order to make reliable predictions of climate change’.

d Here are two characteristic quotes: ‘Model tuning directly influences the evaluation of climate models, as the quantities that are tuned cannot be used in model evaluation.’ (Ref 2, Box 9.1) ‘Though perhaps an obvious point, characteristics (or metrics) that are used to explicitly tune a model or its input should not also be used to evaluate the model – this would be a form of “double-counting”.’ (Ref 4, p. 157)

e See section ‘Beyond Fit’ below. On a charitable reading, this may also be what climate scientists who argue against double-counting (cf. previous endnote) want to claim.

f This is one reason why the slogan that a climate model must ‘get the right effect for the right reasons’ is potentially misleading. If ‘getting the right effect’ means reproducing past and present climate sufficiently well, then the requirement is too weak. A model can reproduce past and present climate for the right reason but still miss processes and feedbacks that are relevant for long-term projections. A second reason is that if ‘for the right reason’ implies that the model resolves all relevant processes (rather than incorporating them via parameterization), then the requirement is too demanding.

g An alternative approach regards spread of ensemble results as guide to possibility rather than probability.81 Betz32,81 and Katzav28 have argued that a focus on probabilistic projection is misguided and that models ought to be used to show that certain scenarios are real possibilities.

h The same problem besets other argument forms, for example, inference to the best explanation. Here we conclude that the causal assumptions shared by the models of an ensemble are approximately true because this is the best explanation for the fact that the results of the models agree and fit reasonable well with observational data if the amount of data is much larger than the degree of freedom in the model. Such an argument faces the additional problem that there may be other equally good explanations of model fit, for example, the fit with data may be due to tuning and the fit with result of other models may reflect a social convergence process among institutions building models.52 For a different line of criticism, see Ref 87.

i The term ‘epistemic opacity’ is due to Humphreys.89,90 While Humphreys focusses on limits in our understanding of the details of the computational process leading from the abstract model underlying a simulation to the output, we are primarily concerned with limits in our understanding of how the components of the underlying model interact and how the model relates to the target. This is also how other authors writing about climate modelling use the term (see Ref 24, p. 177; Ref 91, p. 258).

ACKNOWLEDGMENTS

We thank Claus Beisbart and two anonymous reviewers of this journal for helpful comments on an earlier version of the article.

REFERENCES

1. IPCC. Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM, 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: Cambridge University Press; 2013.

2. Flato GM, Marotzke J, Abiodun B, Braconnot P, Chou SC, Collins W, Cox P, Driouech F, Emori S, Eyring V, et al. Evaluation of climate models. In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM, 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: Cambridge University Press; 2013.

3. Lloyd EA. Confirmation and robustness of climate models. Philos Sci 2010, 77:971–984. doi:10.1086/657427.

4. Schmidt GA, Sherwood S. A practical philosophy of complex climate modelling. Eur J Philos Sci 2015, 5:149–169. doi:10.1007/s13194-014-0102-9.

5. Huber M, Mahlstein I, Wild M, Fasullo J, Knutti R. Constraints on climate sensitivity from radiation patterns in climate models. J Climate 2011, 24:1034–1052. doi:10.1175/2010JCLI3403.1.

6. Knutti R, Furrer R, Tebaldi C, Cermak J, Mehl G. Challenges in combining projections from multiple climate models. J Climate 2010, 23:2739–2758. doi:10.1175/2009JCLI3611.1.

7. Knutti R. Should we believe model predictions of future climate change? Philos Trans R Soc A 2008, 366:4647–4664. doi:10.1098/rsta.2008.0169.

8. Knutti R, Abramowitz G, Collins M, Eyring V, Gleckler PJ, Hewitson B, Mearns L. Good practice guidance paper on assessing and combing multi-model climate projections. In: Stocker T, Dahe Q, Plattner G-K, Tignor M, Midgley P, eds. Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi-Model Climate Projections. Contribution of the IPCC Working Group I, Technical Support Unit. Bern, Switzerland: University of Bern; 2010.

9. Parker WS. Confirmation and adequacy-for-purpose in climate modeling. Proc Aristotelian Soc 2009, 83:233–249. doi:10.1111/j.1467-8349.2009.00180.x.

10. Winsberg E. Computer simulations in science. In: Zalta EN, ed. The Stanford Encyclopedia of Philosophy. Stanford, Summer 2015. Available at: http://plato.stanford.edu/archives/sum2015/entries/simulations-science/.

11. Knutti R. Climate model confirmation: from philosophy to predicting climate in the real world. In: Lloyd L, Winsberg E, eds. Philosophical and Conceptual Issues in Climate Modeling. Chicago: University of Chicago Press; 2016.

12. Frigg R, Thompson E, Werndl C. Philosophy of climate science part 2: modelling climate change. Philos Compass 2015, 10:967–977. doi:10.1111/phc3.12297.

13. Creath R. Logical empiricism. In: Zalta EN, ed. The Stanford Encyclopedia of Philosophy. Stanford, Spring 2014 Edition. Available at: http://plato.stanford.edu/entries/logical-empiricism/.

14. Thornton S. Karl Popper. In: Zalta EN, ed. The Stanford Encyclopedia of Philosophy. Stanford, Winter 2016 Edition. Available at: http://plato.stanford.edu/archives/win2016/entries/popper/.

15. Oreskes N, Shrader-Frechette K, Belitz K. Verification, validation, and confirmation of numerical models in the earth sciences. Science 1994, 263:641–646. doi:10.1126/science.263.5147.641.

16. Müller P. Constructing climate knowledge with computer models. WIREs Clim Change 2010, 1:565–580. doi:10.1002/wcc.60.

17. Cartwright N. Nature’s Capacities and their Measurement. Oxford: Oxford University Press; 1989.

18. McMullin E. Galilean idealization. Stud Hist Philos Sci 1985, 16:247–273. doi:10.1016/0039-3681(85)9003-2.

19. Frigg R, Hartmann S. Models in science. In: Zalta EN, ed. The Stanford Encyclopedia of Philosophy. Stanford, Fall 2012 Edition. Available at: http://plato.stanford.edu/archives/fall2012/entries/models-science/.

20. Weisberg M. Three kinds of idealization. J Philos 2007, 104:639–659. doi:phil20071041240.

21. Strevens M. Depth. An Account of Scientific Explanation. Cambridge, MA: Harvard University Press; 2008, chapter 8.

22. Held IM. The gap between simulation and understanding in climate modeling. Bull Am Meteorol Soc 2005, 86:1609–1614. doi:10.1175/BAMS-86-11-1609.

23. Held IM. Simplicity amid complexity. Science 2014, 343:1206–1207. doi:10.1126/science.1248447.

24. Frisch M. Predictivism and old evidence: a critical look at climate model tuning. Eur J Philos Sci 2015, 5:171–190. doi:10.1007/s13194-015-0110-4.

25. Giorgi F, Torma C, Coppola E, Ban N, Schär C, Somot S. Enhanced summer convective rainfall at Alpine high elevations in response to climate warming. Nat Geosci 2016. doi:10.1038/ngeo2761.

26. Giere RN. How models are used to represent reality. Philos Sci 2004, 71:742–752. doi:10.1086/425063.

27. Giere RN. An agent-based conception of models and scientific representation. Synthese 2010, 172:269–281. doi:10.1007/s11229-009-9506-z.

28. Katzav J. The epistemology of climate models and some of its implications for climate science and the philosophy of science. Stud Hist Philos Sci 2014, 46:228–238. doi:10.1016/j.shpsb.2014.03.001.

29. Meinsenhausen M, Meinsenhausen N, Hare W, Raper S, Frieler K, Knutti R, Frame D, Allen M. Greenhouse-gas emission targets for limiting global warming to 2°C. Nature 2009, 458:1158–1162. doi:10.1038/nature08017.

30. Rogelj J, Meinsenhausen M, Knutti R. Global warming under old and new scenarios using IPCC climate sensitivity range estimates. Nat Clim Change 2012, 2:248–253. doi:10.1038/nclimate1385.

31. Claussen M, Mysak LA, Weaver AJ, Crucifix M, Fichefet T, Loutre M-F, Weber SL, Alcamo J, Alexeev VA, Berger A, et al. Earth system models of intermediate complexity: closing the gap in the
spectrum of climate system models. *Clim Dyn* 2002, 18:579–586. doi:10.1007/s00382-001-0200-1.

32. Frigg R, Smith LA. The myopia of imperfect climate models: the case of UKCP09. *Philos Sci* 2013, 80:886–897. doi:10.1086/673892.

33. Frigg R, Smith LA, Stainforth DA. An assessment of the foundational assumptions in high-resolution climate projections: the case of UKCP09. *Synthese* 2015, 192:3979–4008. doi:10.1007/s11229-015-0739-8.

34. Giere RN. *Explaining Science. A Cognitive Approach*. Chicago: University of Chicago Press; 1988.

35. Weisberg M. *Simulation and Similarity. Using Models to Understand the World*. New York: Oxford University Press; 2013, chapter 8.

36. Suárez M. An inferential conception of scientific representation. *Philos Sci* 2004, 71:767–779. doi:10.1086/421415.

37. Flato GM. Earth system models: an overview. *WIREs Clim Change* 2011, 2:783–800. doi:10.1002/wcc.148.

38. Potochnik A. The diverse aims of science. *Stud Hist Philos Sci* 2015, 22:111–137. doi:10.1335/ken.2012.0008.

39. Intemann K. Distinguishing between legitimate and illegitimate values in climate modeling. *Eur J Philos Sci* 2015, 5:217–232. doi:10.1007/s13194-014-0105-6.

40. Winsberg E. Values and uncertainties in the predictions of global climate models. *Kennedy Inst Ethics J* 2012, 22:111–137. doi:10.1335/ken.2012.0008.

41. Parker WS. Values and uncertainties in climate prediction, revisited. *Stud Hist Philos Sci* 2014, 46:24–30. doi:10.1016/j.shpsa.2013.11.003.

42. Gleckler PJ, Taylor KE, Doutriaux C. Performance metrics for climate models. *J Geophys Res* 2008, 113: D06104. doi:10.1029/2007JD008972.

43. Masson D, Knutti R. Spatial-scale dependence of climate model performance in the CMIP3 ensemble. *J Climate* 2011, 24:2680–2692. doi:10.1175/2011JCLI3513.1.

44. Schaller NI, Mahlstein I, Cermak J, Knutti R. Analyzing precipitation projections: a comparison of different approaches to climate model evaluation. *J Geophys Res* 2011, 116:D10118. doi:10.1029/2010JD014963.

45. Frigg R, Thompson E, Wernsd C. Philosophy of climate science part 1: observing climate change. *Philos Compass* 2015, 10:953–964. doi:10.1111/phc3.12294.

46. Parker WS. Computer simulation, measurement, and data assimilation. *Br J Philos Sci* 2015. doi:10.1093/bjps/axv037.

47. Allen RJ, Sherwood SC. Warming maximum in the tropical upper troposphere deduced from thermal winds. *Nat Geosci* 2008, 1:399–403. doi:10.1038/ngeo208.

48. Mears CA, Wentz FJ. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* 2005, 309:1548–1551. doi:10.1126/science.1114772.

49. Santer BD, Wigley TML, Mears C, Wentz FJ, Klein SA, Seidel DJ, Taylor KE, Thorne PW, Wehner MF, Gleckler PJ, et al. Amplification of surface temperature trends and variability in the tropical atmosphere. *Science* 2005, 309:1551–1556. doi:10.1126/science.1114867.

50. Sherwood SC, Lanzante JR, Meyer CL. Radiosonde daytime biases and late-20th century warming. *Science* 2005, 309:1556–1559. doi:10.1126/science.1115640.

51. Lloyd EA. The role of ‘complex’ empiricism in the debates about satellite data and climate models. *Stud Hist Philos Sci* 2012, 43:390–401. doi:10.1016/j.shpsa.2012.02.001.

52. Edwards PN. Global climate science, uncertainty and politics. Data laden models, model filtered data. *Sci Cult* 1999, 8:437–472. doi:10.1080/09505439909526558.

53. Edwards PN. *A Vast Machine. Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press; 2010.

54. Knutti R, Masson D, Gettelman A. Climate model genealogy: generation CMIP5 and how we got there. *Geophys Res Lett* 2013, 40:1194–1199. doi:10.1002/2012GL05256.

55. Reichler T, Kim J. How well do coupled models simulate today’s climate? *Bull Am Meteorol Soc* 2008, 89:303–311. doi:10.1175/BAMS-89-3-303.

56. Popper K. *The logic of Scientific Discovery*. London: Routledge; [2012 [1959]].

57. Hume D. In: Selby-Bigge LA, ed. *A Treatise of Human Nature*. Oxford: Clarendon Press; 1965 [1739].

58. Bellprat O, Kotlarski S, Lüthi D, De Elía R, Frigon A, Laprise R, Schär C. Objective calibration of regional climate models: application over Europe and North America. *J Climate* 2016, 29:819–838. doi:10.1175/JCLI-D-15-0302.1.

59. Horduin F, Mauritzen T, Gettelman A, Golaz JC, Balaji V, Duan Q, Folini D, Ji D, Klocke D, Qian Y, et al. The art and science of climate model tuning. *Bull Am Meteorol Soc* 2016. doi:10.1175/BAMS-D-15-00135.1.

60. Mauritzen T, Stevens B, Roeckner E, Crueger T, Esch M, Giorgetta M, Haak H, Jungclaus J, Klocke D, Matei D, et al. Tuning the climate of a global model. *J Adv Model Earth Syst* 2012, 4:M00A01. doi:10.1029/2012MS000154.

61. Katzav J, Dijkstra HA, de Laat ATJ. Assessing climate model projections: State of the art and philosophical reflections. *Stud Hist Philos M P* 2012, 43:258–276. doi:10.1016/j.shpsb.2012.07.002.

62. Knutti R. Why are climate models reproducing the observed global surface warming so well? *Geophys
Res Lett 2008, 35:L18704. doi:10.1029/2008GL034932.
63. Knutti R, Allen MR, Friedlingstein P, Gregory JM, Hegerl GC, Meehl GA, Meinshausen M, Murphy JM, Plattner G-K, Raper SCB, et al. A review of uncertainties in global temperature projections over the twenty-first century. J Climate 2008, 21:2651–2663. doi:10.1175/2007JCLI2119.1.
64. Masson D, Knutti R. Predictor screening, calibration, and observational constraints in climate model ensembles: an illustration using climate sensitivity. J Climate 2013, 26:887–898. doi:10.1175/JCLI-D-11-00540.1.
65. Douglas H, Magnus PD. State of the field: why novel prediction matters. Stud Hist Philos Sci 2013, 44:580–589. doi:10.1016/j.shpsa.2013.04.001.
66. Steele K, Werndl C. Climate models, calibration, and confirmation. Br J Philos Sci 2013, 64:609–635. doi:10.1093/bjps/axs036.
67. Parker WS. When climate models agree: the significance of robust model predictions. Philos Sci 2011, 78:579–600. doi:10.1086/661566.
68. Stainforth DA, Allen MR, Tredger ER, Smith IA. Confidence, uncertainty and decision-support relevance in climate predictions. Philos Trans R Soc A 2007, 365:2145–2161. doi:10.1098/rsta.2007.2074.
69. Hargreaves JC. Skill and uncertainty in climate models. WIREs Clim Change 2010, 1:556–564. doi:10.1002/wcc.58.
70. Braconnot P, Harrison SP, Kageyama M, Bartlein PJ, Masson-Delmotte V, Abe-Ouchi A, Otto-Bliesner B, Zhao Y. Evaluation of climate models using palaeoclimatic data. Nat Clim Change 2012, 2:417–424. doi:10.1038/nclimate1456.
71. Hargreaves JC, Annan JD. On the importance of palaeoclimate modelling for improving predictions of future climate change. Clim Past 2009, 5:803–814. doi:10.5194/cp-5-803-2009.
72. Hargreaves JC, Annan JD. Can we trust climate models? WIREs Clim Change 2014, 5:435–440. doi:10.1002/wcc.288.
73. Taylor KE, Stouffer RJ, Meehl GA. A summary of the CMIP5 experiment design. Bull Am Meteorol Soc 2012, 93:485–498. doi:10.1175/BAMS-D-11-00094.1.
74. Parker WS. Ensemble modeling, uncertainty and robust predictions. WIREs Clim Change 2013, 4:213–223. doi:10.1002/wcc.220.
75. Tebaldi C, Knutti R. The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc A 2007, 365:2053–2075. doi:10.1098/rsta.2007.2076.
76. Sanderson BM, Knutti R. On the interpretation of constrained climate model ensembles. Geophys Res Lett 2012, 39:L16708. doi:10.1029/2012GL052665.
77. Bishop CH, Abramowitz G. Climate model dependence and the replicate earth paradigm. Clim Dyn 2013, 41:885–900. doi:10.1007/s00382-012-1610-y.
78. Jun M, Knutti R, Nychka DW. Local eigenvalue analysis of CMIP3 climate model errors. Tellus A 2008, 60:992–1000. doi:10.1111/j.1600-0870.2008.00356.x.
79. Masson D, Knutti R. Climate model genealogy. Geophys Res Lett 2011, 38:L08703. doi:10.1029/2011gl046864.
80. Hwang YT, Frierson DMW. Link between the double-intertropical convergence zone problem and cloud biases over the Southern ocean. Proc Natl Acad Sci USA 2013, 110:4935–4940. doi:10.1073/pnas.1213302110.
81. Stainforth DA, Downing TE, Washington R, Lopez A, New M. Issues in the interpretation of climate model ensembles to inform decisions. Philos Trans R Soc A 2007, 365:2163–2177. doi:10.1098/rsta.2007.2073.
82. Betz G. Underdetermination, model-ensembles and surprises: on the epistemology of scenario-analysis in climatology. J Gen Philos Sci 2009, 40:3–21. doi:10.1007/s10838-009-9083-3.
83. Betz G. Are climate models credible worlds? Prospects and limitations of possibilistic climate prediction. Eur J Philos Sci 2015, 5:191–215. doi:10.1007/s13194-015-0108-y.
84. Sanderson BM, Knutti R, Caldwell P. Addressing interdependency in a multimodel ensemble by interpolation of model properties. J Climate 2015, 28:5150–5170. doi:10.1175/JCLI-D-14-00361.1.
85. Pirtle Z, Meyer R, Hamilton A. What does it mean when climate models agree? A case for assessing independence among general circulation models. Environ Sci Policy 2010, 13:351–361. doi:10.1016/j.envsci.2010.04.004.
86. Weisberg M. Robustness analysis. Philos Sci 2006, 73:730–742. doi:10.1086/518628.
87. Karatz J. Hybrid models, climate models, and inference to the best explanation. Br J Philos Sci 2013, 64:107–129. doi:10.1093/bjps/axs002.
88. McFarlane N. Parameterizations: representing key processes in climate models without resolving them. WIREs Clim Change 2011, 2:482–497. doi:10.1002/wcc.122.
89. Humphreys P. Extending Ourselves: Computational Science, Empiricism, and Scientific Method. New York: Oxford University Press; 2004.
90. Humphreys P. The philosophical novelty of computer simulation methods. Synthese 2009, 169:615–626.
91. Lenhard J, Winsberg E. Holism, entrenchment, and the future of climate model pluralism. Stud Hist Philos Sci 2010, 41:253–262. doi:10.1016/j.shpsb.2010.07.001.
92. Parker WS. Simulation and understanding in the study of weather and climate. *Perspect Sci* 2014, 22:336–356. doi:10.1162/POSC_a_00137.

93. Caldwell PM, Bretherton CS, Zelinka MD, Klein SA, Santer BD, Sanderson BM. Statistical significance of climate sensitivity predictors obtained by data mining. *Geophys Res Lett* 2014, 41:1803–1808. doi:10.1002/2014gl059205.