Chapter 34
Uncertainty Quantification Using Multiple Models—Prospects and Challenges

Reto Knutti, Christoph Baumberger and Gertrude Hirsch Hadorn

Abstract Model evaluation for long-term climate predictions must be done on quantities other than the actual prediction, and a comprehensive uncertainty quantification is impossible. An ad hoc alternative is provided by coordinated model intercomparisons which typically use a "one model one vote" approach. The problem with such an approach is that it treats all models as independent and equally plausible. Reweighting all models of the ensemble for performance and dependence seems like an obvious way to improve on model democracy, yet there are open questions on what constitutes a "good" model, how to define dependency, how to interpret robustness, and how to incorporate background knowledge. Understanding those issues have the potential to increase confidence in model predictions in modeling efforts outside of climate science where similar challenges exist.

Keywords Ensemble modeling · Idealization · Model independence · Robustness · Structural model uncertainty · Uncertainty quantification

34.1 Introduction

Whether conceptual, analytical, or numerical, a model is usually an idealization, i.e., a simplified representation of a target system. A model represents certain elements or processes in order to reproduce or understand the characteristic behavior of a system, to test a hypothesis, or to predict target system quantities of interest that cannot be measured. Often, there are practical limitations that determine the complexity of a model, like the availability of data, computational cost, or even the lack of understanding of some processes that are deemed relevant. What is part of a model and what is not, and how it is represented, is driven by the purpose of the model, i.e., the research question in hand. Therefore, there is not only one possible model of one target, but there are many. The benefit of picking another model, or success of

R. Knutti (✉) · C. Baumberger · G. Hirsch Hadorn
ETH Zurich, Zürich, Switzerland
e-mail: reto.knutti@env.ethz.ch

C. Beisbart and N. J. Saam (eds.), *Computer Simulation Validation*,
Simulation Foundations, Methods and Applications,
https://doi.org/10.1007/978-3-319-70766-2_34

changing the model (or lack thereof) can usually be quantified in terms of prediction skill. Thus, while an infinite number of model structures, boundary conditions, and parameter sets is possible in principle, in practice the decisions on how to further develop a model and whether to accept or reject a proposed change can often be made on a pragmatic basis: a change is likely to be implemented if it is more firmly rooted in theory and if it improves the skill, explanatory power, or usefulness of the model without compromising other desirable properties like efficiency. Improving the model may still be very challenging. But if the model can be evaluated by repeatedly testing its predictions (as, e.g., in the case of weather prediction models), this provides a clear feedback that guides model development. We distinguish model evaluation or validation as the determination of whether a model represents reality well enough for a particular purpose from verification as the determination of whether the output of a simulation approximates the true solutions to the differential equations of the original model. In what follows, we restrict ourselves to computer simulation models. Our focus is on model evaluation rather than on verification.

Model evaluation for long-term climate predictions cannot be based on repeated confirmation of the predictions against observation-based data. Moreover, model evaluation requires uncertainty estimation, ideally in quantitative terms. However, a comprehensive uncertainty quantification, which requires testing different assumptions in a model (i.e., variations in the structure, the processes included), exploring the uncertainty in parameter choices, and quantifying the effect of boundary conditions and datasets, is effectively impossible (see Sect. 34.2; for methods of uncertainty quantification in engineering contexts where repeated confirmation is possible; see Chap. 22 by Dougherty, Dalton and Dehghannasiri in this volume). As an ad hoc alternative, the climate modeling community has therefore started to establish coordinated model intercomparisons. The resulting ensembles of different models can be used to explore uncertainties either by testing the robustness of projections or as a basis for statistical methods that estimate the uncertainty about future climate change. A model projection is usually called robust if it is simulated by most models in the ensemble (although that does not imply that it is accurate). The notion of robustness is more generally used in the sciences to characterize the invariance of a result under multiple independent determinations, be these multiple different modeling approaches or, e.g., diverse experimental devices and measurement practices (Woodward 2006; Wimsatt 2012).

Here we use climate modeling to illustrate a few major (and possibly unique) challenges of determining the robustness of simulation results and estimating their uncertainty (for a general view on validation in climate science see also Chap. 30 by Rood in this volume). These challenges include definitions of core concepts, requirements for ensembles, and metrics for robustness that would support inferences from the robustness of projections, e.g., to warranted confidence in the projections. The challenges are interesting from both a philosophical and a practical point of view. Understanding these issues and finding smarter ways to deal with the resulting plurality of models has the potential to increase the value of models for climate as well as for other environmental areas, and potentially beyond. Eventually, this may increase the confidence we can have in such models as epistemic tools and provide scientists with a clearer explanation of what they are doing, and stronger arguments when it does or does not work.

We first discuss some peculiarities of climate modeling which make a comprehensive uncertainty quantification impossible (Sect. 34.2). We then distinguish between different sources of uncertainty in predicting climate change in order to better understand the motivation of using model ensembles as a means of estimating uncertainties in climate predictions (Sect. 34.3). The usual "one model one vote" approach problematically assumes that all models are independent and equally plausible (Sect. 34.4). As a way to improve on this model democracy, we suggest reweighting all models of an ensemble for performance and dependence (Sect. 34.5), and illustrate the idea for the case of Arctic sea ice (Sect. 34.6). We discuss some open issues, such as whether better agreement with observation reduces uncertainties in predictions, how to define model dependence, and how to incorporate background knowledge in the suggested weighting scheme (Sect. 34.7), and close with a short conclusion (Sect. 34.8).

34.2 Challenges for Uncertainty Quantification in Climate Modeling

Climate and Earth system models of various complexity are used to simulate the statistics of weather and how these will change in the future as a result of the emission of greenhouse gases like carbon dioxide and other radiatively active species (Claussen et al. 2002; Knutti 2008; Flato 2011). The problem of using such models for simulations has several peculiarities.

The first peculiarity relates to the system's many dimensions: simulating the weather in principle requires resolving the atmosphere, ocean, ice, and land surface of the Earth, because of the many processes and timescales that affect weather. From the condensation of water on a tiny aerosol (on spatial scales of micrometers and timescales of fractions of a second) to the large-scale ocean circulations and melting of ice sheets (extending over thousands of kilometers and thousands of years), the processes involved occur over at least twelve orders of magnitudes in both time and space. And from soil microbes that potentially affect the growth of a tree and its effect on the local carbon and water cycle to complex chemistry affecting cloud formation, from subglacial hydrology to volcanoes affecting the radiative balance in the stratosphere, from our technological progress in developing renewable energy sources to policy instruments that affect the rate of decarbonization, the list of (potentially) relevant processes that affect future climate is extremely long. The challenge consists of nothing less than simulating the whole Earth including human behavior, which by construction is impossible; and even if it were possible, it would not be reasonable. Due to the interactions of the many aspects in the climate system, an increase in complexity typically decreases the analytic understanding of a model

(Lenhard and Winsberg 2010). However, deciding on what to include and exclude, and how to simplify, is tricky.

The second peculiarity, partly a consequence of the first, is that it is prohibitively expensive to build a new model for each research question. The expertise and effort required imply that a big institution typically builds only one or two (often similar) versions of a model every few years. The same model is then used to study literally hundreds of different questions. Thus, rather than a specific purpose guiding model construction, we observe that it is the model, once it is built, that determines what purposes it can be used for. The third peculiarity, also a consequence of the first, is the computational cost and volume of data involved. A climate simulation typically takes days to months running on hundreds to thousands of processors of a supercomputer, which makes it prohibitively expensive to systematically optimize the dozens of parameters it has, or try hundreds of ideas before converging on a new model. Development is therefore strongly guided by experts' understanding of what could work, based on background knowledge and experience of what ideas have worked in similar situations or in other models in the past (Held 2005).

The fourth peculiarity is that a direct confirmation of the actual prediction is often impossible. To confirm the prediction of climate in the year 2100, one would have to wait for nearly a century, and even then a single confirmation would not be sufficient given the chaotic component of atmospheric variability. The development cycle of a model is usually much shorter than the typical timescales for confirmation. Model evaluation for long-term predictions therefore must be done on quantities other than the actual long-range prediction, e.g., observations of current climate (Gleckler et al. 2008; Knutti 2008; Flato 2011; Schaller et al. 2011), its variability, past changes, or paleoclimate data (Harrison et al. 2015). The question then becomes which quantities matter most for what question (see Sect. 34.5).

The mentioned peculiarities make it practically impossible to test many different assumptions in a model (i.e., variations in the structure and the processes included), different choices for parameters, and to quantify the effects of boundary conditions and datasets in a systematic way. However, such a systematic assessment would be required for comprehensive quantification of uncertainties. Coordinated model intercomparisons offer an ad hoc work-around to this problem. Such efforts were started by the climate modeling community about two decades ago. They require whoever is willing to contribute to perform standardized simulations and provide the results to others for analysis. The resulting ensembles of different models are often referred to as "ensembles of opportunity", since they group together existing models and are not designed to span an uncertainty range (Knutti 2010a; Knutti et al. 2010; Eyring et al. 2016).

34.3 Uncertainty Quantification Using Model Ensembles

To better understand the motivation of using ensembles of different models, it is useful to characterize the sources of uncertainty in predicting climate change. Three sources of uncertainty can be distinguished: natural variability (both internal to the system and externally forced from changes in solar irradiance and volcanic eruptions), scenario uncertainty and model uncertainty. Natural internal variability is an inherent property resulting from the chaotic nature of the ocean–atmosphere system. We cannot predict the weather more than about a week in advance, because tiny uncertainties in the initial conditions grow as we run the model forward in time. The system is sensitive to its initial conditions, much like a Lorenz system with multiple attractors. That does *not* imply that the system is fundamentally unpredictable; the models indicate that some aspects like the temperature difference between winter and summer or the long-term trend resulting from increased CO_2 in the atmosphere are predictable, although bifurcations may exist in parts of the system, e.g., the Atlantic meridional overturning circulation (Lenton et al. 2008). Climate, the distribution of all weather states, therefore is very likely predictable, but the individual sequence of weather events, is not (Deser et al. 2012). This uncertainty, often referred to as ontic uncertainty because it is due to the chaotic nature of the target system, is largely accounted for by making statements about the climate averaged over 20 or more years. Hence, it is not fundamentally impossible to deal with this variability, but it is challenging because we can only evaluate the model in a probabilistic sense (i.e., by comparing distributions), and single events are of little value for judging the adequacy of a model. The second source of uncertainty, scenario uncertainty, results from uncertainty in emissions of anthropogenic forcings like CO_2, methane, -SO_2, or ozone. These are driven by technological progress, climate policy, values in society, wars, etc., all of which are difficult to predict because they are based on human behavior. This is also an ontic uncertainty, due to inherent properties of in this case socio-techno-economic systems. This uncertainty is often accounted for by considering projections (as opposed to predictions), defined as the response of climate conditional on a predefined scenario of societal development (along with emissions, land use change, etc.) (Vuuren et al. 2011).

This leaves us with model uncertainty, which is an epistemic uncertainty, i.e., a lack of knowledge about whether the model is an appropriate representation of the target system in question. A model is a representation of reality that is necessarily simplified in important ways. First, some processes in the climate system are not fully understood, e.g., changes in complex ecosystems. Second, some are rather well understood but are so complex or small-scale that their effect has to be parameterized in a simple way as a function of available large-scale properties (Gent et al. 1995; McFarlane 2011), e.g., ocean mixing processes and transports occurring on scales smaller than the resolution of the model (typically 100 km). The corresponding parameters (e.g., an equivalent diffusivity) must be calibrated to match large-scale observations and have no analog measurable equivalent quantity in reality. Third, numerical approximations and finite resolution lead to small errors when integrating the equations. In principle, this could be improved by larger computers, but, in practice, every doubling of horizontal resolution requires about ten times more computing capacity, so it will take many decades before the relevant scales (tens to hundreds of meters) can be resolved in global simulations (Schneider et al. 2017). In addition, boundary conditions (like the bathymetry of the ocean or the structure and properties of the soil) at every location are not fully known.

As a consequence of all of the above, it is often said that climate models are uncertain, but this is a misconception. Strictly, a model, once it is specified in the form of equations or code, is perfectly certain, in the sense that applying the equations twice will give the exact same results, and the effect of any change in the equations can be quantified precisely. The uncertainty comes from the model being a representation of a target in the real world, which requires specification and inference steps, in deciding what to include in the model, and how to interpret the results of the models for the real world. Of course, every climate model is false, by construction, in the sense that it is an idealized representation of a real and open system (Oreskes et al. 1994). Not only does the model ignore some climate processes but it also distorts the represented processes in different ways in order to make them mathematically and computationally tractable. The question is not whether the model is true but whether it is “true enough” (Elgin 2017), i.e., how well it represents the real system, and how useful or adequate it is for learning about a particular aspect of the real system.

This last point, the adequacy of a model, motivates the pluralism in climate modeling: because of the complexity of the system, the computational cost, and the lack of direct confirmation of prediction, there is no single agreed-on “best” model. Scientists inevitably have to make choices in what to include, how to parameterize unresolved processes, and how to manage the tradeoff between complexity, resolution, the number of simulations and number of years to simulate. Since there is disagreement on how to make these choices, to some extent even for a given purpose, there is no consensus on which one is the “best” model. So while multiple models could be seen as ontologically incompatible (strictly speaking, they make conflicting assumptions about the real world), and one could argue that scientists have to assess how well they are supported by data, the community seems happy with the model pluralism. The models are seen as complementary in the sense that they are all plausible (although not necessarily equally plausible) representations of the real system given the incomplete knowledge, data, and computational constraints; they are used pragmatically to investigate uncertainties (Parker 2006, 2010, 2013).

The diversity of models across an ensemble provides one avenue to try to estimate the consequences of model uncertainty by testing the robustness of results. For example, there are several ways one can parameterize atmospheric small-scale convection, and it is helpful to test whether the model behavior depends on the structure of that parameterization and the parameter values. Robustness in a qualitative sense is often invoked as a premise in an argument to the effect that a model result can be trusted (see Parker 2011, for a critical discussion). Robustness analysis goes a step further, using robust results to confirm certain parts of a model. Robustness analysis was developed as a modeling strategy in population ecology (Levins 1966). It has been generalized and systematized (Weisberg 2006; Wimsatt 2012), and also been applied to climate science (Lloyd, 2009, 2010). Robustness analysis uses a robust result as confirmatory evidence for more general relations of a model, which are then called "robust theorems". Robust theorems have the form: "Ceteris paribus, if [common core (causal) structure] obtains, then [robust property] will obtain" (Weisberg 2006). For instance, if all models that share a core causal structure but use a variety of simplifications show that higher CO_2 concentrations lead to substantial warming, then that result is unlikely to be just a consequence of particular choices made in a model. This robust result is then used to formulate the robust theorem: "Ceteris paribus, if [Greenhouse gases relate in law-like interaction with the energy budget of the earth] obtains, then [increased global mean temperature] will obtain" (Lloyd, 2009, 2010). But there are of course limits to such an argument: there are cases where all models are known to be robustly wrong in the same way because they all ignore a process (e.g., ice sheet dynamics) or parameterize it in a similar way. In order to avoid being misled by the robustness of results that is, in fact, pseudo-robustness (Wimsatt 2012), models must be sufficiently diverse in the relevant regards. There is considerable controversy on how to specify this requirement. A typical way is to specify "diversity" as "independence" (Wimsatt 2012) and to elaborate on a formal account for explicating this concept, for instance in a Bayesian framework (Fitelson 2001; Lloyd 2010; Stegenga and Menon 2017). However, these approaches are not uncontested (Schupbach 2016), and their appropriate specification remains a challenge.

In our discussion, we focus on determining the robustness of simulation results used to estimate the uncertainty in long-term climate predictions, which needs to be distinguished from robustness analysis used to confirm certain parts of a model. For brevity, we will focus on the most interesting and challenging case of multiple structurally different models in the Coupled Model Intercomparison Projects CMIP (Eyring et al. 2016), noting that similar ideas can, of course, be applied to what is often called perturbed physics ensembles, a model run with a variety of parameter sets (Stainforth et al. 2005). Many issues are similar, except that a single model structure can only capture so much of the range of behavior: no parameter set of one model will ever behave (in all respects) like a structurally different model that resolves other processes, although parameter calibration can compensate for some missing aspects of processes.

34.4 Problems with Model Democracy

Ensembles of opportunity like CMIP are often used for uncertainty quantification in a naïve way: the average of all models is taken as a best estimate, and the spread of the models is reported as the uncertainty of the projection. This "one model one vote" or "model democracy" (Knutti 2010), often used based on a lack of more convincing or generally agreed-upon alternatives, makes several assumptions which are rarely explicitly stated and even less frequently defended by actual evidence. First, model democracy treats all models as reasonably independent, and second, it assumes that all models are about equally plausible. Third, it assumes that the range of model projections represents what we believe is the uncertainty in the projection. In a weather forecast, the equivalent would be a probabilistic projection that is neither too broad nor overconfident, so that for many trials, observed outcomes would fall within the estimated 5–95% confidence intervals in about 90% of the trials.

Unfortunately, none of the assumptions made by model democracy is strictly fulfilled by present-day model ensembles. On the first point of dependence: many models use ideas, parts of the code, or even whole components (e.g., the sea ice model) from other models. The sheer complexity and cost lead groups to merge their efforts in jointly developing or using components of other groups (Bellouin et al. 2011). New models are almost never developed from scratch but are based on earlier models (Edwards 2011). As a consequence, some models are not providing much additional information, and multiple replications of a model may strongly bias the result toward that particular model (Annan and Hargreaves 2011; Masson and Knutti 2011a; Pennell and Reichler 2011; Knutti et al. 2013). How to actually define model dependence is not straightforward (Annan and Hargreaves 2016). The models are of course dependent in the sense that they all describe the same system, but that is not the point: they are also similarly biased with regard to how they represent reality because they share structural limitations or simplify things in the same way, and therefore their projections will likely be biased in the same way. If two models share several parts, the success of one model in simulation results has implications for the probability of the other model's success. This leaves us with the question of how to explicate an appropriate notion of dependence and specify a metric to determine model dependence (see Sects. 34.6 and 34.7).

On the second assumption, some models clearly perform better than others in some metrics (for an introduction and overview on relevant metrics, see Chap. 18 by Saam in this volume), i.e., simulation results are closer to observations of reality, with differences of up to a factor of two (Knutti et al. 2013). Reductions in the biases by 20–30% from one model intercomparison to the next imply that some models are about a decade of model development ahead of others in terms of how well they reproduce the observations. No model is clearly far superior to all others, consistent with the idea of pluralism where all models are seen as plausible representations of reality given some practical boundary conditions; but some are more plausible in certain respects than others. Some models perform well on certain metrics while others perform well on others (Gleckler et al. 2008), which reflects different modeling groups' focus in terms of development and calibration. But a model that performs well on one metric also tends to perform well on many others for at least two reasons: the climate system is coupled, so a correct representation of rainfall, for example, requires humidity (and therefore temperature), the dynamics (weather patterns), and clouds to be well represented. The other fact is a practical one: some centers simply have more resources (people and computing power) and experience than others, and their models tend to do well on many criteria.

On the third assumption, the spread of model projections does not necessarily represent what we believe is the uncertainty in the prediction. The spread of the ensemble may be too big if the ensemble contains demonstrably unrealistic members that can be rejected upfront based on physical understanding or disagreement with observations (see Chap. 6 by Beven and Lane in this volume). A model of Venus or Mars, for example, is unlikely to provide a useful projection of climate for the Earth and should thus be excluded from the respective ensemble. The model spread can also be too small if all models are missing the same relevant thing and are biased in the same way. In many cases, we do not know whether the spread tends to be too large or too small, and that likely depends on the variable, the timescale and the spatial scale (Masson and Knutti 2011b).

A further complication is the question whether the ensemble of models is centered around the truth (the so-called "truth plus error" paradigm, in which every model simulation approximates the observations of reality with a random error), or whether the observations of reality and the models are drawn from the same distribution (the "indistinguishable" paradigm, in which truth is not necessarily in the center). The former implies that predictions would get ever more certain as more models are added (in much the same way as the estimate of the average fall speed of a rock gets more and more precise as we continue to measure the time for the rock to reach the ground, if the measurement errors are random) which is certainly not the case. But the average of all models often does perform better than any individual model, suggesting some truth-centeredness at least for the observations available. This interpretation however can also change from the past to the future. For projections, the indistinguishable paradigm appears to be the more plausible interpretation in most cases (i.e., reality has about the same likelihood to approximately match any of the model realizations, and it is not necessarily in the center of the distribution) (Annan and Hargreaves 2010; Sanderson and Knutti 2012).

34.5 Beyond Model Democracy

Reweighting the ensemble for performance and dependence seems like an obvious way to improve on model democracy: poor and duplicated models would be down weighted and models whose performances agree well with observations and are relatively independently developed would constitute stronger evidence. Yet the discussions around such methods have been controversial so far. One argument against weighting is the sensitivity of the results to the chosen metric and possible overconfidence: if we weight by something that is unrelated to the quantity of interest or dominated by variability, then there is a possibility that the result gets worse rather than better (Weigel et al. 2010), and we may not know whether it does get worse. However, this is really only an issue when the number of models is very small. For a large number of models, it would essentially converge to random weights which should not affect the results. Sometimes, there are also political sensitivities: it can be difficult to dismiss models from certain centers or countries in a coordinated modeling effort. The other main argument raised against model weighting is that there are many ways to do it and the lack of direct confirmation prevents us from testing which approach is optimal. Indeed we can define an infinite number of model performance metrics (measuring the agreement with data in some way, e.g., a root mean square difference to observations, or a spectrum, or conservation of properties), and arguing which performance metric is relevant for the quality of a model is challenging (Knutti et al. 2010b). Unlike in weather forecasting, for example, we cannot quantify skill by repeated confirmation. Many broad brush metrics (e.g., patterns of temperature or rainfall) in fact appear to be only weakly correlated to large-scale projections like global temperature across a set of models (Jun et al. 2008; Knutti et al. 2010a). The reasons for the lack of relationships can be a large structural uncertainty in the models, lack of observed trend due to large variability, or lack of observations. Another hypothesis is that most of the observed data have already been used in model development and evaluation, such that the current set of models can already be interpreted as a posterior conditional on the observations; as a consequence, using the same observations again would not add anything (Sanderson and Knutti 2012).

The argument of model weighting gets more convincing, we would argue, if we assess model quality in relation to a particular purpose (Parker 2009). The question of which model is "best" is ill-posed unless we agree on the task the model is used for. The answer depends on the task we are trying to accomplish, in much the same way as which car people would say is best depends on whether they try to go really fast, or drive off-road, or move furniture. Defining weights for predicting a certain variable X is easier both politically and scientifically. Politically because one model will get more weight for predicting X, and another one will get more weight for predicting a different variable Y, which is only natural as some groups focus their development more on X and others more on Y. Scientifically, it is easier to select processes and quantities that are relevant for predicting X: one can refer to background knowledge, i.e., knowledge of various kinds that are accepted in the scientific community about the factors that determine X. Such insight can come from process understanding, trends emerging from natural variability, detection, and attribution, or from so-called emergent constraints, which typically are strong relationships between an observable quantity and a prediction. Observing the former can provide a constraint on the latter. For example, the strength of the albedo feedback on a seasonal timescale is related to the albedo feedback on decadal timescales (Hall and Qu 2006); hence, e.g., models that lose Arctic sea ice faster in the past tend to lose it faster in the future (Boé et al. 2009; Mahlstein and Knutti 2012; Overland and Wang 2013; Notz and Stroeve 2016; Knutti et al. 2017). Not all such relationships are robust across a wide range of models and there is a danger of spurious correlation when testing a large number of predictors (Masson and Knutti 2013; Caldwell et al. 2014). But despite all difficulties, when relationships across models are well understood in terms of the underlying processes, they can provide guidance on which quantities to use for model weighting.

As an alternative to attaching weights to models, emergent constraints can also be used to define a relationship between the observable and the projection (usually through some form of regression across models). This relationship can then be used to estimate an observationally constrained projection that is relatively independent of the set of underlying models (Boé et al. 2009; Mahlstein and Knutti 2012; Cox et al. 2013). Other options are interpolations in a low-dimensional model space (Sanderson et al. 2015b) or Bayesian methods (Tebaldi et al. 2004). They all vary in their statistical methods but share the idea of deviating from model democracy by using observed evidence. Also, strategies that combine dynamic models with other types of models using data-driven methods (Mazzocchi and Pasini 2017) need to use observational data, which are unavailable for long-term predictions. Data-driven approaches are genetically independent from dynamic models and are using different modeling schemes and methodological approaches. They may fit observations better given enough degrees of freedom, but may still be biased when it comes to out-of-sample prediction.

34.6 Illustration of Model Weighting for Arctic Sea Ice

We illustrate the idea of combining projections from multiple models here for Arctic sea ice, by weighting models both for their performance relative to observations and for model dependence. The method is relatively straightforward in the sense that a single number is defined as a weight for each simulation (although the choices that need to be made are not trivial, as discussed below), and it has been used in various contexts (Sanderson et al. 2015a, b; Knutti et al. 2017; Sanderson et al. 2017). The example is taken from an earlier study by Knutti et al. (2017), and is chosen because the processes are relatively well understood, and the added value of using observations is immediately obvious: to estimate when the Arctic will likely be ice-free, the model should have about the right sea ice extent today, and about the right trend over the past decades. Sea ice loss in the past and the future is correlated across models (Boé et al. 2009; Mahlstein and Knutti 2012; Overland and Wang 2013; Notz and Stroeve 2016; Knutti et al. 2017), which is plausibly explained by some models having a stronger sea ice albedo feedback than others. Observed sea ice trends are therefore an obvious constraint. There are of course other methods to weight models (Abramowitz and Gupta 2008; Waugh and Eyring 2008; Boé et al. 2009; Massonnet et al. 2012; Abramowitz and Bishop 2015), but the method outlined here may be the most straightforward one to illustrate the concepts.

For M models in the ensemble, the weight w_i for model i is defined as

$$w_i = e^{-\frac{D_i^2}{\sigma_D^2}} \Bigg/ \left(1 + \sum_{j \neq i}^{M} e^{-\frac{S_{ij}^2}{\sigma_S^2}}\right) \tag{34.1}$$

The numerator weighs a simulation by the distance metric D_i of model i to observations (performance), while the denominator effectively takes into account how many times parts of a model are replicated based on S_{ij}, the distance metric between model i and model j, which informs about the dependence of the models in the ensemble. Both D_i and S_{ij} are evaluated here as root mean square differences of a series of variables, but different choices for the metric and the functional form of the weighting can be defended. The weights are scaled such that their sum over the whole ensemble equals one. The constants σ_D and σ_S determine how strongly the model performance and dependence ("similarity") are weighted (see below). This weighting scheme fulfills two basic requirements: a model that is infinitely far from observations and does in no way represent the real Earth (very large D_i) gets zero weight. For a model with no close neighbors, the denominator equals one and has no effect. However, duplicating an otherwise independent model ($S_{ij} = 0$) leads to a denominator being equal to two: as a consequence the two duplicates each get half of the weight, and the result is unaffected by the duplication. Because initial condition members (multiple simulations of the same model with slightly different starting conditions) are very similar, they are effectively treated as near replicates, and all available simulations can be used in a straightforward way even if the number of initial condition ensemble members varies strongly between models.

The metrics D_i and S_{ij} give equal weight to the climatological mean hemispheric mean September Arctic sea ice extent (1980–2013), and its trend over the same period, gridded climatological mean surface air temperature for each month, and climatological mean gridded interannual variability of monthly surface air temperature, but the sensitivity of the results to the choice of variable is illustrated in the results.

The choice of σ_D and σ_S determines how close a model's simulation results need to be to observations to be considered "good" (performance), and how close two models need to be in order to be considered "similar" (dependence), respectively. The choice of these parameters is not straightforward. A very small σ_D, for example, may lead to the total weight being concentrated on just one or two models, at the expense of the results' robustness. A very large value, on the other hand, will result in the weighting having almost no effect. One way to inform the choice of these parameters is to use perfect model tests, i.e., sequentially treating one of the models as reality and using the others to predict its future. Confirmation is possible, in this case, and allows optimizing the parameters for maximum skill while ensuring that the predictions are not overconfident. However, if models are similarly wrong then the perfect model tests might suggest that the method works well even if it does not in the real world. As such, perfect model tests are a necessary but not sufficient step for informing the choice of these parameters and to demonstrate the skill of the proposed method. A more in-depth discussion is provided by Knutti et al. (2017).

Weighting models can be done straightforwardly based on Eq. (34.1), but a number of choices with regards to variables, regions, time periods, and parameters are important. Hence, the results' sensitivity toward these assumptions needs to be tested, and background knowledge is required to judge which choices are plausible. If clear constraints exist from observations, then the weighting makes the models more consistent with the past and narrows the model spread of the projection. In the following, we present an example of how taking the observations into account can improve the projections relative to a model democracy case.

Figure 34.1 shows the simulated September Arctic temperature and sea ice extent for all available fully coupled climate models (i.e., structurally different models as well as multiple initial condition members). Colors from gray to yellow to red indicate increasingly higher weights. The blue line indicates observations. Weighting is not based on the time series only, but on how well the models simulated the whole Arctic climate (see figure caption for details). Figure 34.1c indicates that the observationally weighted projection range (red) is substantially narrower than the raw range, i.e., the model democracy case (gray), and agreement with the observed trends is better. Note that we would not expect perfect agreement, as the observations represent one single realization whereas the weighted model average is closer to a forced response with much weaker variability. In the case of Arctic sea ice, there is evidence that part of the strong ice loss might be due to natural variability (Kay et al. 2011; Swart et al. 2015; Screen and Francis 2016). This would be consistent with the observed decline in sea ice being steeper than the weighted model average.

34.7 Discussion and Open Issues

As we argued in recent articles (Knutti 2010; Knutti et al. 2017), model democracy is increasingly hard to justify for climate model projections. Biases in some models and variables are so large that they cannot be ignored; in the example of Arctic sea ice discussed above, a model without sea ice in the present day or one with more sea ice by 2100 than observed today would be challenging to deal with. Simple bias correction methods that consider anomalies from a reference state will not work well or at all in such cases, as the change will depend on the reference state (if no sea ice is left, the change will also be zero). So if there are observations or other sources of information that can inform, or even better narrow the range of plausible projections, it would be strange not to use them.

In our view, there are essentially three points that need to be considered: performance as measured by agreement with observed data, model dependence, and background knowledge. In the case of dynamical models (as opposed to statistical models that are fitted), good agreement with a variety of observations provides strong evidence that the models are doing the relevant things correctly, but is not a formal proof of course (Baumberger et al. 2017). While confidence in the results should be larger when they are obtained by models that reproduce relevant aspects of current climate more accurately, performance alone provides insufficient support for long-term predictions. Furthermore, if the processes likely relevant for specific projections are sufficiently well understood and captured in the models, the coherence of models with this background knowledge provides an additional reason that increases our confidence in a projection (Baumberger et al. 2017). Given the complexity of the system, a model never agrees with all the data, but that is not required. The question is whether the model provides insight that we would not have otherwise. But how do we deal with a situation where improving the model based on process understanding, either through a more physical representation of a process, through increased resolution, or by explicitly resolving a process that has been prescribed or ignored before, leads to poorer agreement with data? Such situations are not uncommon, and can result from observation biases or from compensating errors in the models. From an understanding point of view, we might trust the new model more than the old one, and further development might improve the agreement again. Yet in an operational setting where users depend on predictions, a lower skill is hard to justify. Even in a

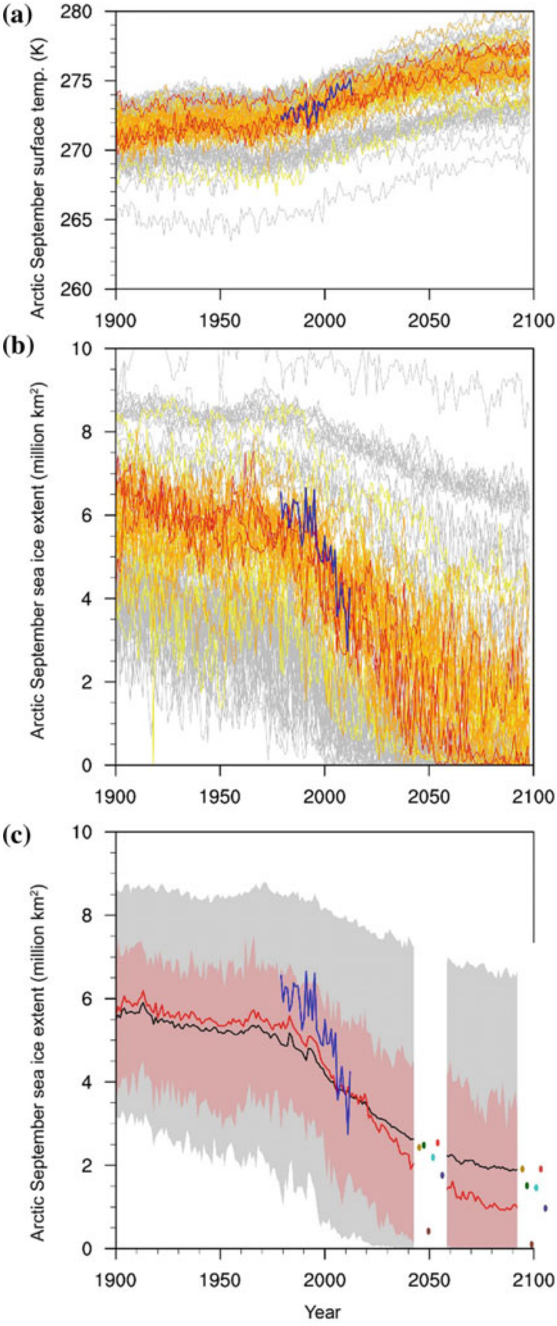

Fig. 34.1 **a** Arctic (60–90°N) September surface air temperature, **b** Arctic September sea ice extent in all CMIP3/5 simulations. Gray, yellow, orange and red indicates those that get <0.5%, >0.5%, >1%, and >5% weight, respectively, from weighting with Eq. (34.1). Observations (NCEP) are shown in blue. **c** Mean and 5–95% range for no weighting (black line, gray band) and weighting (red line and band). Colored dots near 2050 and 2100 show 2046–2055 and 2090–2099 average sea ice extent using (from left to right) the following metrics: (1) none (unweighted), (2) climatological mean (1980–2013) September sea ice extent, (3) September sea ice extent trend 1980–2013, (4) climatology of monthly surface temperature (1980–2013), (5) interannual variability of monthly surface temperature, (6) all 2–5. Figure reproduced from Knutti et al. (2017)

research context there is a tendency for a “dog and pony show”: the argument that it “looks good” is easier to sell than the fact that the underlying processes are more realistically described. This, of course, raises interesting discussion about the value of fit, and calibration (“tuning”) (Baumberger et al. 2017; Knutti 2018).

It is important to keep in mind that better agreement with observations will not necessarily reduce uncertainties in projections (Knutti and Sedláček 2012). But even in cases where it does not (Sanderson et al. 2017), we should not conclude that the effort was useless. This inability to further constrain the model range can arise either because the spread was not sufficient to begin with, or because the ensemble was already weighted due to good models being replicated a lot (Sanderson et al. 2017), or because the observations are not long enough or of sufficient quality or have too much variability to provide a constraint, or because the quantity of interest is inherently unpredictable, or because we have already used most of the information in the model development, evaluation and calibration. But in any case, we would not know until we have actually done the exercise. If the posterior after weighting is similar to the prior, then we have not reduced the spread, but we can be confident that the projection is reasonably consistent (in both magnitude and spread) with the observations we have on mean and trends. The raw model spread is just a range across models and cannot be interpreted as an uncertainty. It is an ad hoc measure of spread reflecting the ensemble design, or lack thereof, whereas the weighted results can be interpreted as an incomplete measure of uncertainty given all observations we have. The numbers may be similar, but the interpretation of the range is very different, and we should have more confidence in the latter.

Stronger constraints will come in the future (and have already in the past) from better observing systems specifically designed for climate change (early observations were mostly taken for weather prediction where long-term stability of a system was less of a concern), and from anthropogenic trends. Often past trends are more strongly related to future trends in a model than the mean state is related to future trends. But past forced trends may have been amplified or masked by natural variability, in particular over shorter periods (Deser et al. 2012; Fischer and Knutti 2016; Saffioti et al. 2016; Medhaug et al. 2017). Given the strong limits of available observations and computational capacity, model development and evaluation will, therefore, be a continuous process, and uncertainty estimates of projections will continue to change, as is the case in most other research areas. The lack of direct confirmation and the reliance on multiple potentially strongly dependent models however is somewhat unique to climate projections.

Model performance is an issue that any model developer always considers. In contrast, the issue of model dependence has gotten far less attention in the climate community. It is something that only becomes apparent after the various institutions have finalized their models. Only the most recent intercomparisons provided clear evidence that this problem can no longer be ignored, and there is less of a consensus on how to deal with it. It is likely to get more pronounced as model development gets increasingly complex and expensive. People sharing ideas or code, or developing code in a collaborative way is perfectly fine, but its impact on projections has to be considered in the interpretation of the results. In Sect. 34.6, we proposed a straightforward way how to include model dependence as a term in the weighting.

An open issue is a proper mathematical definition of model dependence that can actually be implemented in practice (Annan and Hargreaves 2016). Models' resembling each other by sharing certain parts or features is an indication for them being related, but once the simulation results of two models converge to observations, the simulation results of the various models will also get closer and closer to each other without the models necessarily being dependent. Furthermore, models that are independent from others may be irrelevant for the hypothesis in question. Because of these basic problems, it has also been questioned whether a concept of dependence is appropriate to explicate the diversity of models or other methods for reliably determining the robustness of their results (Schupbach 2016). More pragmatic concerns with applying a formal concept of probabilistic dependence in the case of climate models are for example how to find out which processes are represented in which way in different models. In most cases of using ensembles for determining the robustness of simulation results, these concerns are not an issue right now, because the distance of models' results to observations is typically far bigger than the distance between two strongly related models. Dependence and performance are treated independently in the example in Sect. 34.6, but further work may come up with different or more sophisticated alternatives.

Another open issue is an adequate selection process for ensemble members that avoids both pseudo-robustness resulting from excluding relevant plausible but diverging models (too narrow spread of results, e.g., because few centers in CMIP try to develop models with extreme behavior) and lack of robustness resulting from including irrelevant models (too broad spread of results). Which models are relevant depends on the hypothesis (purpose) for which the ensemble is used, which needs to be assessed by reference to relevant background knowledge about the problem in question and experiences with modeling practices. While this is a question that cannot be answered in general, making the considerations on the relevance of models explicit in each case would be a general requirement on using ensembles to determine the robustness of predictions. Scientists, e.g., often implicitly consider background knowledge when selecting an ensemble, but these considerations should be made explicit.

Background knowledge is important for considering whether to exclude or down-weight models which violate basic physical principles (such as conservation of water or energy), or which lack representations of processes or feedbacks that are known to play an important role for future climate. In general terms: If the models within an ensemble differ strongly in how coherent they are with background knowledge, and if it is likely that there is a correlation between how well a model is based on process understanding and the model's adequacy for long-term projections, then the coherence with background knowledge should be considered in weighting the models for estimating uncertainties in such projections. It is important to see why coherence with background knowledge cannot be built into the dimension of performance: If two models reproduce equally well observed mean climate and trends but we know from background theories that only the first represents certain feedbacks (e.g., greenhouse gas emissions from thawing permafrost) which significantly influence future climate, then the first model should be given more weight than the second. On the one hand, one could think of coherence with relevant background knowledge as a consideration additional to determining the robustness of results (Parker 2013), e.g., for determining which models to include in an ensemble in the first place. On the other hand, one might think about integrating coherence with the relevant background knowledge as a further term in the weighting. In a Bayesian framework, the first option affects the prior, which is based on the whole ensemble; the second option affects the posterior, which depends on the weighing of the models. However, there is still considerable work to do in order to find a qualitative or a quantitative way to consider coherence with relevant background knowledge. It needs, e.g., to be determined how to deal with the intransparency of what exactly is in the models, and with limitations in the state of knowledge. Moreover, a procedure for assessing this coherence, e.g., something like eliciting expert judgments, needs to be established. Accounting for coherence with relevant background knowledge is a challenging task, but it needs to be addressed in order to improve the epistemic significance of robust results.

34.8 Conclusion

We have used climate modeling to illustrate a few major (and possibly unique) challenges of determining the robustness of simulation results for long-term predictions and of estimating their uncertainty. We have proposed to weight the models of an ensemble in order to avoid biases that result when all models are treated equally. We have proposed a somewhat ad hoc scheme that considers dependence and performance of the models, yet there are challenges that need further work. These include how to quantitatively account for coherence with background knowledge as a further important requirement on ensembles, as well as definitions of core concepts and metrics in order to provide a quantitative determination of the robustness of simulation results. Such an explicit and systematic approach to robustness of results is required to support inferences from the robustness of projections and to establish confidence in the projections. These challenges are interesting from both a philosophical and a practical point of view. Improving our understanding of these issues and finding better ways to deal with the plurality of models has the potential to increase the value of models not just for climate but other environmental areas, and potentially beyond, where determining the robustness of results is a strategy to assess confidence in results. Eventually, this may provide scientists with a clearer explanation of what they are doing in modeling, and stronger arguments about when modeling as an epistemic tool does or does not work.

Acknowledgements We thank Claus Beisbart, Keith Beven, Benedikt Knüsel, Wendy Parker, and Marius Wälchli for discussions and comments which have improved the quality of this manuscript. This project was supported by the European Union’s Horizon 2020 research and innovation program under grant agreement No 641816 (CRESCENDO). We acknowledge the World Climate Research

Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led the development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

Abramowitz, G., & Gupta, H. (2008). Toward a model space and model independence metric. *Geophysical Research Letters, 35*(5), 1–4. https://doi.org/10.1029/2007GL032834.

Abramowitz, G., & Bishop, C. H. (2015). Climate model dependence and the ensemble dependence transformation of CMIP projections. *Journal of Climate, 28,* 2332–2348. https://doi.org/10.1175/JCLI-D-14-00364.1.

Annan, J. D., & Hargreaves, J. C. (2010). Reliability of the CMIP3 ensemble. *Geophysical Research Letters, 37*(2), 1–5. https://doi.org/10.1029/2009GL041994.

Annan, J. D., & Hargreaves, J. C. (2011). Understanding the CMIP3 multimodel ensemble. *Journal of Climate, 24*(16), 4529–4538. https://doi.org/10.1175/2011JCLI3873.1.

Annan, J., & Hargreaves, J. (2016). On the meaning of independence in climate science. *Earth System Dynamics Discussions*, 1–17. https://doi.org/10.5194/esd-2016-34.

Baumberger, C., Knutti, R., & Hirsch Hadorn, G. (2017). Building confidence in climate model projections: An analysis of inferences from fit. *Wiley Interdisciplinary Reviews: Climate Change, 8*(3), e454. https://doi.org/10.1002/wcc.454.

Bellouin, N., et al. (2011). The HadGEM2 family of Met Office Unified Model climate configurations. *Geoscientific Model Development*, *4*(3), 723–757. https://doi.org/10.5194/gmd-4-723-2011.

Boé, J., Hall, A., & Qu, X. (2009). September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nature Geoscience*, *2*(5), 341–343. (Nature Publishing Group). https://doi.org/10.1038/ngeo467.

Caldwell, P. M., et al. (2014). Statistical significance of climate sensitivity predictors obtained by data mining. *Geophysical Research Letters, 41*(5), 1803–1808. https://doi.org/10.1002/2014GL059205.

Claussen, M., et al. (2002). Earth system models of intermediate complexity: Closing the gap in the spectrum of climate system models. *Climate Dynamics, 18*(7), 579–586. https://doi.org/10.1007/s00382-001-0200-1.

Cox, P. M., et al. (2013). Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature*, *494*(7437), 341–344. (Nature Publishing Group). https://doi.org/10.1038/nature11882.

Deser, C., et al. (2012). Communication of the role of natural variability in future North American climate. *Nature Climate Change, 2*(11), 775–779. https://doi.org/10.1038/nclimate1562.

Edwards, P. N. (2011). History of climate modeling. *Wiley Interdisciplinary Reviews: Climate Change, 2*(1), 128–139. https://doi.org/10.1002/wcc.95.

Elgin, C. Z. (2017). *True enough*. Project MUSE: The MIT Press.

Eyring, V., et al. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016.

Fischer, E. M., & Knutti, R. (2016) Observed heavy precipitation increase confirms theory and early models. *Nature Climate Change*, *6*(11), 986–991. (Nature Publishing Group). https://doi.org/10.1038/nclimate3110.

Fitelson, B. (2001). A Bayesian account of independent evidence with applications. *Philosophy of Science, 68*(S3), S123–S140. https://doi.org/10.1086/392903.

Flato, G. M. (2011). Earth system models: An overview. *Wiley Interdisciplinary Reviews: Climate Change, 2*(6), 783–800. https://doi.org/10.1002/wcc.148.

Gent, P. R., et al. (1995). Parameterizing eddy-induced tracer transports in ocean circulation models. *Journal of Physical Oceanography*, *25*(4), 463–474. (American Meteorological Society).

Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research, 113*(D6), 1–20. https://doi.org/10.1029/2007JD008972.

Hall, A., & Qu, X. (2006). Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophysical Research Letters, 33*(3), L03502. https://doi.org/10.1029/2005GL025127.

Harrison, S. P., et al. (2015). Evaluation of CMIP5 palaeo-simulations to improve climate projections. *Nature Climate Change, 5*(8), 735–743. https://doi.org/10.1038/nclimate2649.

Held, I. M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society, 86*(11), 1609. https://doi.org/10.1175/BAMS-86-11-1609.

Jun, M., Knutti, R., & Nychka, D. W. (2008). Spatial analysis to quantify numerical model bias and dependence. *Journal of the American Statistical Association, 103*(483), 934–947. https://doi.org/10.1198/016214507000001265.

Kay, J. E., Holland, M. M., & Jahn, A. (2011). Inter-annual to multi-decadal Arctic sea ice extent trends in a warming world. *Geophysical Research Letters, 38*(15), 2–7. https://doi.org/10.1029/2011GL048008.

Knutti, R. (2008). Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 366*(1885), 4647–4664. https://doi.org/10.1098/rsta.2008.0169.

Knutti, R. (2010). The end of model democracy? *Climatic Change*, *102*(3–4), 395–404. https://doi.org/10.1007/s10584-010-9800-2.

Knutti, R. (2018). Climate model confirmation: From philosophy to predicting climate in the real world. In *Climate modelling* (pp. 325–359). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65058-6_11.

Knutti, R., & Sedláček, J. (2012). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change, 3*(4), 369–373. (Nature Publishing Group). https://doi.org/10.1038/nclimate1716.

Knutti, R., Furrer, R., et al. (2010a). Challenges in combining projections from multiple climate models. *Journal of Climate, 23*(10), 2739–2758. https://doi.org/10.1175/2009JCLI3361.1.

Knutti, R., Abramowitz, G., et al. (2010b). Good practice guidance paper on assessing and combining multi model climate projections, meeting report of the intergovernmental panel on climate change expert meeting on assessing and combining multi model climate projections. In T. F. Stocker, et al. (Eds.), *IPCC working group I technical support unit*. Switzerland: University of Bern, Bern.

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters, 40*(6), 1194–1199. https://doi.org/10.1002/grl.50256.

Knutti, R., et al. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters, 44*(4), 1–10. https://doi.org/10.1002/2016GL072012.

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *41*(3), 253–262. (Elsevier). https://doi.org/10.1016/j.shpsb.2010.07.001.

Lenton, T. M., et al. (2008). Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0705414105.

Levins, R. (1966). The strategy of model building in population biology. *American Naturalist*, 421–431. https://doi.org/10.2307/27836590.

Lloyd, E. A. (2009). I—Elisabeth A. Lloyd: Varieties of support and confirmation of climate models. *Aristotelian Society Supplementary Volume*, *83*(1), 213–232. https://doi.org/10.1111/j.1467-8349.2009.00179.x.

Lloyd, E. (2010). Confirmation and robustness of climate models. *Philosophy of Science*, *77*(5), 971–984. Retrieved July 7, 2014, from http://www.jstor.org/stable/10.1086/657427.

Mahlstein, I., & Knutti, R. (2012). September Arctic sea ice predicted to disappear near 2 °C global warming above present. *Journal of Geophysical Research, 117*(D6), 1–11. https://doi.org/10.1029/2011JD016709.

Masson, D., & Knutti, R. (2011a). Climate model genealogy. *Geophysical Research Letters, 38*(8), L08703. https://doi.org/10.1029/2011GL046864.

Masson, D., & Knutti, R. (2011b). Spatial-scale dependence of climate model performance in the CMIP3 ensemble. *Journal of Climate, 24*(11), 2680–2692. https://doi.org/10.1175/2011JCLI3513.1.

Masson, D., & Knutti, R. (2013). Predictor screening, calibration, and observational constraints in climate model ensembles: An illustration using climate sensitivity. *Journal of Climate, 26*(3), 887–898. https://doi.org/10.1175/JCLI-D-11-00540.1.

Massonnet, F., et al. (2012). Constraining projections of summer Arctic sea ice. *The Cryosphere*, *6*(6), 1383–1394. https://doi.org/10.5194/tc-6-1383-2012.

Mazzocchi, F., & Pasini, A. (2017). Climate model pluralism beyond dynamical ensembles. *Wiley Interdisciplinary Reviews: Climate Change, 8*(6), e477. https://doi.org/10.1002/wcc.477.

McFarlane, N. (2011). Parameterizations: Representing key processes in climate models without resolving them. *Wiley Interdisciplinary Reviews: Climate Change, 2*(4), 482–497. https://doi.org/10.1002/wcc.122.

Medhaug, I., et al. (2017). Reconciling controversies about the “global warming hiatus”. *Nature*, *545*(7652), 41–47. (Nature Publishing Group). https://doi.org/10.1038/nature22315.

Notz, D., & Stroeve, J. (2016). Observed Arctic sea-ice loss directly follows anthropogenic CO_2 emission. *Science, 354*(6313), 747–750. https://doi.org/10.1126/science.aag2345.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*(5147), 641. AAAS. Retrieved June 4, 2014, from http://www.sciencemag.org/cgi/content/abstract/sci;263/5147/641.

Overland, J. E., & Wang, M. (2013). When will the summer Arctic be nearly sea ice free? *Geophysical Research Letters, 40*(10), 2097–2101. https://doi.org/10.1002/grl.50316.

Parker, W. S. (2006). Understanding pluralism in climate modeling. *Foundations of Science*, *11*(4), 349–368. (Springer). http://www.springerlink.com/index/138424X1082M7277.pdf.

Parker, W. S. (2009). Confirmation and adequacy-for-purpose in climate modelling. *Aristotelian Society Supplementary, 83*(1), 233–249. https://doi.org/10.1111/j.1467-8349.2009.00180.x.

Parker, W. S. (2010). Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *41*(3), 263–272. (Elsevier). https://doi.org/10.1016/j.shpsb.2010.07.006.

Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science, 78*(4), 579–600. https://doi.org/10.1086/661566.

Parker, W. S. (2013). Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change, 4*(3), 213–223. https://doi.org/10.1002/wcc.220.

Pennell, C., & Reichler, T. (2011). On the effective number of climate models. *Journal of Climate, 24*(9), 2358–2367. https://doi.org/10.1175/2010JCLI3814.1.

Saffioti, C., et al. (2016). Reconciling observed and modeled temperature and precipitation trends over Europe by adjusting for circulation variability. *Geophysical Research Letters, 43*(15), 8189–8198. https://doi.org/10.1002/2016GL069802.

Sanderson, B. M., & Knutti, R. (2012). On the interpretation of constrained climate model ensembles. *Geophysical Research Letters, 39*(16), L16708. https://doi.org/10.1029/2012GL052665.

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015a). A representative democracy to reduce interdependency in a multimodel ensemble. *Journal of Climate, 28*(13), 5171–5194. https://doi.org/10.1175/JCLI-D-14-00362.1.

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015b). Addressing interdependency in a multimodel ensemble by interpolation of model properties. *Journal of Climate, 28*(13), 5150–5170. https://doi.org/10.1175/JCLI-D-14-00361.1.

Sanderson, B. M., Wehner, M., & Knutti, R. (2017). Skill and independence weighting for multi-model assessments. *Geoscientific Model Development*, *10*(6), 2379–2395. https://doi.org/10.5194/gmd-10-2379-2017.

Schaller, N., et al. (2011). Analyzing precipitation projections: A comparison of different approaches to climate model evaluation. *Journal of Geophysical Research, 116*(D10), 1–14. https://doi.org/10.1029/2010JD014963.

Schneider, T., et al. (2017). Climate goals and computing the future of clouds. *Nature Climate Change, 7*(1), 3–5. (Nature Publishing Group). https://doi.org/10.1038/nclimate3190.

Schupbach, J. N. (2016). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science*, *69*(February), axw008. https://doi.org/10.1093/bjps/axw008.

Screen, J. A., & Francis, J. A. (2016). Contribution of sea-ice loss to Arctic amplification is regulated by Pacific Ocean decadal variability. *Nature Climate Change, 6*(9), 856–860. https://doi.org/10.1038/nclimate3011.

Stainforth, D. A., et al. (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature, 433*(7024), 403–406. https://doi.org/10.1038/nature03301.

Stegenga, J., & Menon, T. (2017). Robustness and independent evidence. *84*(July), 414–435. http://www.journals.uchicago.edu/doi/10.1086/692141.

Swart, N. C., et al. (2015). Influence of internal variability on Arctic sea-ice trends. *Nature Climate Change, 5*(2), 86–89. (Nature Publishing Group). https://doi.org/10.1038/nclimate2483.

Tebaldi, C., et al. (2004). Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophysical Research Letters, 31*(24), 1–5. https://doi.org/10.1029/2004GL021276.

Vuuren, D. P., et al. (2011). The representative concentration pathways: An overview. *Climatic Change, 109*(1–2), 5–31. https://doi.org/10.1007/s10584-011-0148-z.

Waugh, D. W., & Eyring, V. (2008). Quantitative performance metrics for stratospheric-resolving chemistry-climate models. *Atmospheric Chemistry and Physics*, *8*(18), 5699–5713. https://doi.org/10.5194/acp-8-5699-2008.

Weigel, A. P., et al. (2010). Risks of model weighting in multimodel climate projections. *Journal of Climate, 23*(15), 4175–4191. https://doi.org/10.1175/2010JCLI3594.1.

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

Wimsatt, W. C. (2012). Robustness, reliability, and overdetermination (1981). In L. Soler et al. (Eds.), *Characterizing the robustness of science: After the practice turn in philosophy of science* (pp. 61–87). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-2759-5_2.

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology, 13*(2), 219–240. https://doi.org/10.1080/13501780600733376.