Chapter 10
Climate Change Projections: Characterizing Uncertainty Using Climate Models

Ben Sanderson and Reto Knutti

Glossary

Bayes' Theorem	A law in probability theory relating the probability of a hypothesis given observed evidence to the often easier to characterize probability of that evidence given the hypothesis. The theorem states that the conditional "posterior" probability of an event A given an event B is equal to the "prior" probability of A multiplied by the likelihood of B given A is true, normalized by the prior probability of B.
Climate sensitivity	The equilibrium global mean near surface air temperature response in Kelvin to a sustained doubling of the atmospheric carbon dioxide concentration.
CMIP-3	The Coupled Model Intercomparison Project Phase 3, a set of coordinated model experiments using General Circulation Models from the world's major modeling centers.
Detection and attribution	A process whereby spatial "fingerprints" associated with individual climate forcing factors (such as aerosol or greenhouse gas concentrations) are identified and used to quantify whether an observed change exceeds the

This chapter was originally published as part of the Encyclopedia of Sustainability Science and Technology edited by Robert A. Meyers. DOI:10.1007/978-1-4419-0851-3

B. Sanderson (✉)
National Center for Atmospheric Research, NCAR, 1850 Table Mesa Dr, 80305 Boulder, CO, USA
e-mail: bsander@ucar.edu

R. Knutti
Institute for Atmospheric and Climate Science, ETH, Universitätstrasse 16, Zürich, Switzerland
e-mail: reto.knutti@env.ethz.ch

P.J. Rasch (ed.), *Climate Change Modeling Methodology: Selected Entries from the Encyclopedia of Sustainability Science and Technology*,
DOI 10.1007/978-1-4614-5767-1_10,

	range of natural internal climate variability (detection) and to attribute it to different causes, that is, different forcings (attribution).
Empirical model	A model based on fitting empirical data, and thus makes no attempt to justify its representations of the system with any physical basis.
General circulation model (GCM)	A three-dimensional mathematical model for the atmosphere and possibly the ocean, land, and sea ice.
Initial condition ensemble	A number of simulations using a single climate model, each with a small, unique perturbation to the initial state.
Last glacial maximum (LGM)	A period in the most recent ice age lasting several 1,000 years, peaking approximately 20,000 years ago at the maximum extent of the ice sheets.
Lead time	The period in between the time at which the forecast is made and the time to be forecasted.
Multi-model ensemble (MME)	A collection of structurally different models from a range of institutions used to perform a coordinated set of experiments.
Parameter space	The multidimensional domain created by considering the possible values of a number of parameters within a model.
Perturbed physics ensemble	A set of climate simulations generated by taking a single physical model and altering uncertain parameters within a range of plausibility.
Prior probability (marginal probability)	The probability of an event before any additional data is considered in a Bayesian sense.
Posterior probability	The probability of an event after considering additional relevant evidence in a Bayesian sense.
Systematic error	The difference between a model simulation and observations or a poorly represented process which is not reducible by parameter tuning.

Definition of the Subject and Its Importance

The atmosphere, ocean, land surfaces, and ice sheets of the Earth are highly complex and coupled systems, with physical laws which describe behavior from the microscopic to the planetary scale. General Circulation Models are computational analogs for these physical systems, which can be used to study how these systems might behave when boundary conditions are changed (e.g., by increasing the concentration of atmospheric greenhouse gases).

Inherent in the design of such models are a myriad of choices when deciding which components of the system are to be modeled and how to represent processes which cannot be currently modeled explicitly. In order to have any confidence in the ability of our models to have value for simulating aspects of future climate change, it is necessary for those models to reproduce observable properties of the physical system. However, model errors in the simulation of the past or present are likely to be smaller than errors in future projections because model developers can use observations and historical records in the development of their code. Additionally, some processes may not be observable or testable yet, because they might only take place in a warmer (or otherwise changed) world.

One way to characterize at least some of the uncertainty in future projections is to produce an ensemble of climate simulations, each making different but reasonable assumptions about their representation of physical processes.

In recent years, a number of groups in the international climate science community have produced General Circulation Models of the earth system, each making different choices about model complexity, resolution, and parameterization of processes which occur at scales finer than those resolved. By conducting coordinated experiments with each of these models, it has become possible to examine some of the effect that such choices have on uncertainty in future climate simulations. However, the sheer volume of data and range of models available from such an ensemble presents a new challenge for the science to address: How can a spread of non-independent "best guesses" be combined to produce meaningful statements of uncertainty which are relevant to climate-related policy decisions?

Introduction

In 1979, Jule Charney chaired a committee on anthropogenic global warming, producing a report [1] providing a brief overview of the state of the science of climate change. At the time, two General Circulation Models were available for consideration, one led by Syukuro Manabe and the other by James Hansen. The report produced an estimate for the climate sensitivity (the equilibrium global mean temperature change to a doubling of the atmospheric carbon dioxide concentrations) based on the mean result of these two models. In comparing the predicted future climate of these two models, the report stated:

> We conclude that the predictions of CO_2-induced climate changes made with the various models examined are basically consistent and mutually supporting. The differences in model results are relatively small and may be accounted for by differences in model characteristics and simplifying assumptions.

This, in many ways, represents the first effort to combine multiple results from an ensemble of climate model simulations, and the statements made using those models are still relevant to ensemble modeling. A better understanding in the uncertainties in the simulations and increased confidence can be claimed if an

ensemble of somewhat independent models produces common features in its simulations, and if the origins of the differences between simulations may be traced back to physical characteristics.

When presented with a range of simulations of future climate, one must make judgments on many levels on how that ensemble should be interpreted: How should model agreement, or lack of it, translate into a degree of confidence in the simulations? Should all models be treated equally, and if not then how should one distinguish between them? If some processes are absent from some or all of the simulations, how can the projections be updated to account for these "known and unknown unknowns"? Should each ensemble member be interpreted to be an estimate of the "truth" with some unknown error, or should the "true" earth system be considered as a potential member of the ensemble? Although some of these questions verge on the philosophical, the judgments made in answering them can have large effects on the results themselves that are obtained and the degree of uncertainty in those results.

In recent years, there have been systematic efforts to explore and characterize uncertainty using large ensembles of increasingly complex models of the earth system. These model simulations have been coordinated and analyzed to help in characterizing climate change in a series of assessment reports for the Intergovernmental Panel on Climate Change (IPCC). In 1990, 1995, 2001, and 2007, a selection of GCMs were assembled from various major modeling groups around the world to compare simulations of past, future, and other idealized scenarios of climate change. Through the successive decades, model complexity and scope have increased; the early GCMs of Manabe [2] and Hansen [3] modeled atmospheric dynamics and radiative transfer, with a simplistic representation of the hydrological cycle. By the time of the First Assessment Report of the IPCC [4] (FAR) in 1991, models included clouds, a land surface model, and prescribed ice cover. For the Second Assessment Report [5] (SAR), some models also included a representation of the ocean and interactive sea ice. In the Third Assessment Report [6] (TAR), some models considered the effects of volcanic eruptions and aerosol emissions, with a fully dynamical representation of the oceans. By the time of most recent Fourth Assessment Report [7] (AR-4), some models were beginning to include an explicit representation of the carbon cycle in the earth system. Today's models continue to model additional components of the earth system, such as interactive vegetation, dynamically resolved ice sheets, a coupled carbon-nitrogen cycle and full atmospheric chemistry. In parallel with all of these improvements, the last few decades have also seen a continued increase in model resolution. Where the models used in the FAR split the earth into cells as large as 500 km on a side, models for the AR4 can resolve at a scale of a few tens of kilometers.

This entry focuses on the advantages and additional complexities which one must consider when studying a range of different model simulations. Rather than giving a comprehensive description of the results of the models assessed in the successive generations of the IPCC, this entry will discuss the added technical and conceptual challenges encountered when considering the results of a range of non-independent models and how a range of simulations may be combined into best estimates and uncertainties for future climate evolution.

Projection Uncertainty and the Need for Ensembles

Empirical and Physical Models

In 150 AD, Ptolemy devised a model of the motion of planets in the solar system by describing a system of concentric, geocentric circles (or "*deferents*") on which were mounted smaller circles ("*epicycles*") on which the planets themselves were mounted. This system thus had a large number of degrees of freedom (the diameters and speeds of rotation of each of the deferents and epicycles), which could be finely tuned to reproduce the motions of the bodies in the night sky. Such was the predictive power of this approach, that variations of this simple model were accepted until Copernicus' heliocentric model was published in 1543. Although Copernicus' model fits the established view of the universe more closely, both of these models were *empirical* in that they were not based on any physical principles at that time. However, even without a physical basis, Galileo was able to validate the Copernican model by studying the phases of the planet Venus – which was only consistent with the heliocentric formulation. It was not until Newton's law of universal gravitation that the model could be given a physical underpinning.

Any model of a physical system is an approximate representation of the truth. It should be able to reproduce some behavior of that system, and it might do this empirically like Ptolemy's model or by explicitly simulating physical processes within the true system like an orbital system based upon Newtonian gravitation. A model, whether empirical or physical, cannot ever be validated in the strict sense of showing it to be a wholly correct representation of the true system; it can only be evaluated by reproducing some output not used in the tuning of the model itself. This was true of Galileo's observation of the phases of Venus – information not used in the tuning of the Ptolemic model. However, any empirical model becomes very sensitive to changing boundary conditions. For example, if the mass of the Sun were to instantly double, the Copernican model of the solar system would be a very poor approximation of planetary motion, whereas a model based upon Newtonian mechanics would capture enough of the necessary physics to remain useful.

These fundamental principles are relevant to methodologies for simulating the climate today. If the simplest, zero dimensional empirical model of the climate is taken to be:

$$C\frac{dT'}{dt} = F' - \lambda T'$$

where T' is the global mean temperature difference from an equilibrium state, F' is the additional radiative forcing to the planet (i.e., the change in the top of atmosphere radiative balance caused by a forcing, e.g., increased CO_2), C is the effective heat capacity of the system, and λ is the global sensitivity parameter. This equation has two free parameters, C and λ which may be tuned such that the model can fit an observed past time series of F' and T', that of the twentieth century, for example.

The model can then be evaluated by predicting a previously unseen time period, such as the last glacial maximum. This evaluation, if successful, would give more confidence in the model but would not necessarily make it trustworthy for a prediction of the future – where the boundary conditions are outside those seen in both the training and validation period.

The added advantage of using a GCM to simulate future climate is that model simulations are in theory more trustworthy because they are based upon physical principles, which it is believed can reproduce observed climate by coupling underlying physical laws that are known to be true. However, this view is often overoptimistic; although some components of the modeled climate, such as the equations of motion in the atmosphere or the instantaneous radiative forcing due to a change in atmospheric carbon dioxide concentrations are well understood and consistently implemented in different GCMs, there are other processes such as convection which cannot explicitly be resolved with current computing resources. These processes and their effects on the large scale climate must be approximated with uncertain parameters that must be estimated by tuning the model to reproduce some observed features of the climate. What this means, in practice, is that a GCM is neither only an empirical nor an explicitly physical model; it is a hybrid of the two where model developers face many arbitrary choices in parameterizing processes which cannot be explicitly resolved. The necessity for the tuning process reintroduces some of the problems encountered with an empirical model, with the possibility of false confidence in model performance by over-tuning the model to reproduce past climate. The ambiguity in these parameterizations justifies the existence of multiple models for the same purpose [8]; each of these models is seen as a plausible approximation of the climate system given the imperfect understanding, the uncertainties in observations and the computational constraints.

Types of Uncertainty and the Need for Ensembles

Although weather and climate simulations share some properties (sometimes they are conducted with the same model), the limiting uncertainties are very different. Climate represents the distribution of all possible states in which one expects to find a system, whereas weather is the specific evolution of the system from a given initial state. A model-based weather forecast is a *prediction*, in that the initial state of the simulation is as close as possible to observations and the absolute errors grow rapidly afterwards. In weather prediction, these errors occur as specific weather systems evolve from the initial state. Because the atmosphere is a chaotic system, very small errors in the estimate of the initial state can result in very large differences in the distributions of weather systems a few days later. Initial condition uncertainty is evaluated by repeating simulations with a range of slightly different initial conditions to form “Initial Condition Ensembles.” The spread of these ensembles initially grows rapidly but eventually saturates when the “memory” of

the initial state is lost (this timescale is longer in the oceans, perhaps up to 10 years for North Atlantic ocean temperatures [9]).

On decadal to century timescales, the mean and spread of an initial condition ensemble represents a *projection* of the future climate state, although this spread is only a small fraction of the total error (sometimes known as "uncertainty of the first kind"). The second kind of uncertainty relates to the boundary conditions of the problem, some of which are naturally occurring such as the level of incoming solar radiation or volcanic activity, while others are dependent on anthropogenic factors such as the future emissions of greenhouse gases or aerosols. To address this uncertainty, one must perform a range of simulations using different plausible scenarios for changes in boundary conditions. The results of any simulation are therefore conditional on assumptions made about future human behavior. There is currently little real skill in forecasting future volcanic activity and changing solar activity so simplistic but plausible scenarios for these quantities are often used (such as repeating past values). However, in most future scenarios these represent a relatively small fraction of the total anthropogenic climate forcing.

Figure 10.1 shows the relative sources of error in a climate model projection as a function of the lead time [11]. For lead times of less than a decade, the uncertainty in the initial state combined with chaotic error growth and natural patterns of variability are the dominant sources of error but on the scale of several decades or more, it is the future emissions scenario which dominates the uncertainty. Predictions on all timescales, however, are subject to model uncertainties. These arise when a climate model contains parameterizations for unresolved or missing processes. Parameterizations take large-scale quantities resolved by the model, such as temperature, wind speed, and humidity, and relate them to unresolved processes, such as convective mass flux and cloud profiles. Although these parameterizations are usually constructed from physical underpinnings and evaluated with observed data, they introduce some unavoidable uncertainty when a range of parameter values might be physically plausible. GCMs are often subject to a tuning of parameter values to reproduce features of the observed climate, but with tens or hundreds of uncertain parameters this process is time consuming and can yield multiple solutions because of the computational cost, a systematic tuning of all parameters is unfeasible.

One method of quantifying the parameter uncertainty problem is to construct a "Perturbed Physics Ensemble" (PPE) using a single GCM. This process has been attempted using several major climate models [11–13] and involves taking a subset of important unknown parameters within the GCM and perturbing them within the bounds of physical plausibility. Such experiments might perturb, for example, a parameter which states the necessary humidity required for the formation of cloud. By varying this parameter, one can change dramatically how the model distributes clouds both in the present day and the future. These changes can affect the strength of global feedbacks which can change, for example, the amount of warming that the model predicts for a given rise in greenhouse gases. An example of a PPE is the "climateprediction.net" experiment [13], which used idle time in volunteer's computers to perform perturbed simulations of future climate. Incorporating this range into an uncertainty estimate for predictions of future

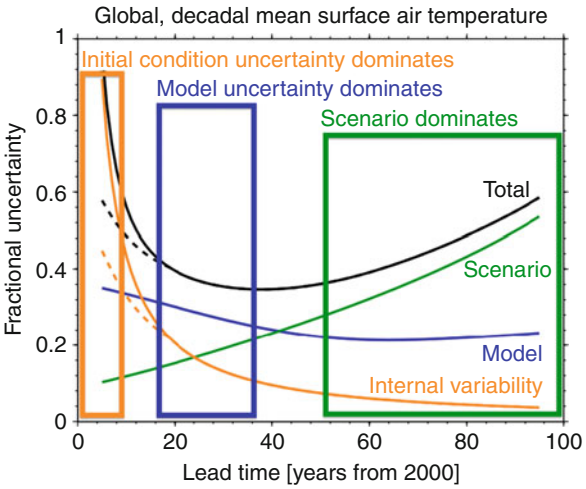

Fig. 10.1 A figure showing the fractional sources of uncertainty in a climate model projection as a function of time. The *orange* "internal variability" *line* shows the errors due to uncertain knowledge of the initial state of the system, while the *dotted line* shows the potential reduction in error if effort is made to assimilate ocean observations into the model at the start of the simulation. The *green line* shows the fractional error due to the unknown future emissions of climate altering gases, while the *blue line* shows the error due to the model imperfections. The *boxes* show where different types of uncertainty are dominant for a projection of future climate (Reproduced from [10])

climate requires a framework for joint consideration of each model's performance in simulating past and present climate as well as its future response.

The remaining model uncertainties are due to so-called "systematic" or structural errors arising from the model design, that is, the choices of which processes to model, the resolution of the model, the numerical schemes, and the specific form of the parameterization scheme. The structural differences between different GCMs provide a lower bound on the extent of the structural difference between any one GCM and the true climate system, but in reality the models in an ensemble such as those used for the IPCC reports share many common properties in terms of resolution, numerical methods, missing components, and parameterization schemes which might make all the models subject to similar errors. Nevertheless, considering a range of GCMs which make different modeling assumptions is an essential step when evaluating the robustness of any prediction of future climate change because it places a lower bound on the uncertainty arising from the choices made by model developers.

Multi-model and Perturbed Physics Ensembles

When making predictions of a future climate state, there is a wealth of evidence to suggest that considering a combined prediction using multiple, somewhat independent models yields more accurate results than any single model [14–16].

Additionally, the spread of simulations provides a measure of robustness in the prediction. The following section describes some reasons for the increased performance of multi-model and perturbed physics ensemble forecasts, together with some of the complexities arising from their analysis.

Range of Ensemble Responses

The spread of results from an ensemble of climate simulations is dependent upon the experimental design, or lack of it. A perturbed physics ensemble (PPE) has the luxury of allowing some control of the distribution of models in the parameter space of the model, though the structure of the underlying model places a fundamental limit on the range of observable behavior in the ensemble. For example, if a PPE is created by perturbing cloud parameters in a GCM which has no parameterization for cirrus clouds formed by gravity waves, then there is no way that such an ensemble can include uncertainty about that process. Designers of such experiments must also be aware that the decisions of how to sample the parameter space of a model will directly influence the distribution of future climate simulations [17]. In contrast, multi-model ensembles (MMEs) such as the Coupled Model Intercomparison Project (CMIP-3), explore "systematic" model differences, which sample models with different representations of the physical system, rather than simply varying parameters in a single model. These are "ensembles of opportunity" where multiple modeling groups run coordinated experiments but the ensemble itself is not sampled in any systematic fashion. Nor is the ensemble randomly sampled because each modeling group will tune their model to minimize model differences from observations, thus creating an ensemble of "best-guesses." This is quite different from the PPE case where the model is intentionally detuned to produce a wide range of behavior. Evidence for this can be seen by examining the spread of climate sensitivity in both a multi-model and a perturbed physics ensemble (Fig. 10.2). When considering a range of observational constraints on climate sensitivity, it is apparent that the multi-model values tend to cluster about the most likely value, whereas the perturbed physics ensemble contains models which span the full range of uncertainty in climate sensitivity. Although impossible to verify, it also is possible that there is also a component of social anchoring [18] which draws multi-model sensitivities toward the mean value as any group which finds their model to be an outlier may have to defend why this is the case, whereas a model with the consensus value of sensitivity is less likely to be questioned.

The Ensemble Mean State

In various fields, it has been shown that the combined performance of multiple models can exceed that of an individual ensemble member. Examples of this can be seen in models of crop yield [20], disease modeling [21], and even in the optimization routines

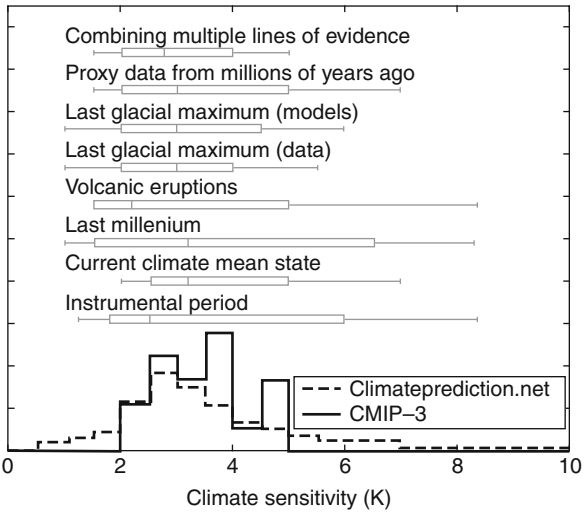

Fig. 10.2 Distribution functions of climate sensitivity (an estimate of the equilibrium response of a model to a doubling of CO_2) for models in the CMIP-3 ensemble (hinting at the range of responses from an MMF analysis), compared with a selection of models from the climateprediction.net project (hinting at the range of uncertainty from a PPE analysis). *Box and whisker plots* show estimates of the most likely values, together with 66th and 90th percentiles of likelihood for climate sensitivity taken from various lines of observational evidence (Adapted from [19]). *Histograms* represent the fraction of models in each 0.5 K bin of climate sensitivity for the atmosphere-only components of 19 models in the CMIP-3 archive and for a 2,000 member subset of the climateprediction.net ensemble [13]

used for movie recommendations based upon past viewing choices [22]. Similarly in seasonal climate predictions, it has been shown that the multi-model ensemble means yield better forecasts, in general, than using only initial condition ensembles from a single model [16]. A multi-model study incorporating a set of initial conditions for each model is often referred to as a "super-ensemble." The accuracy of the model mean often performs best in multivariate applications, that is, a single model may show increased skill in predicting one particular diagnostic, but when many variables are considered in the same metric the ensemble mean prediction tends to show greater skill than any individual model [23].

This effect can also be seen in GCM simulations of recent past climate. Figure 10.3 shows successive generations of the CMIP ensemble evaluated using a multivariate error metric comparing twentieth century observations to model simulations of that period for a variety of model diagnostics. The figure shows that model errors in simulating the current climate have decreased over time but also that for each generation of the ensemble, the multi-model mean results in a model-data discrepancy almost as good, or better than the best performing ensemble member. Various studies have found that both in detection and attribution studies [24] and in simulations of recent climate [25] that a multi-model mean provides a better multivariate simulation than any individual model.

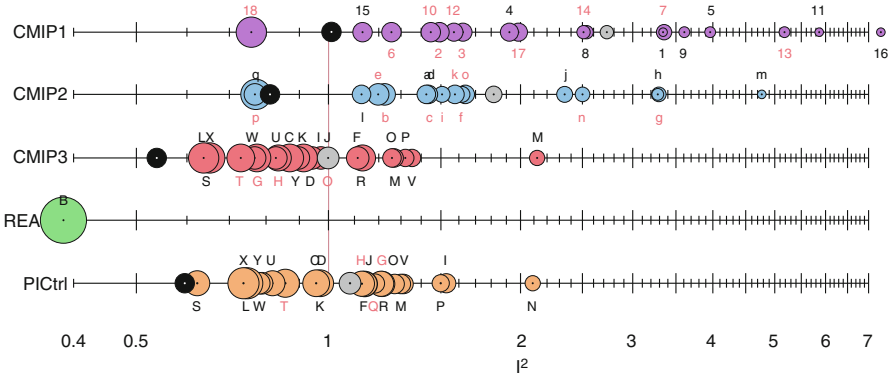

Fig. 10.3 A comparison of model errors reproduced from Reichler and Kim ([26], corrigendum) multivariate errors are evaluated for successive generations of the coupled model intercomparison project (CMIP). *Colored circles* represent individual models in the ensemble, whereas *black circles* show the performance of the multi-model mean. REA represents the NCEP reanalysis and the CMIP-3 PICNTRL is the performance of the preindustrial control simulations when evaluated against present-day observations

This approach is common in the reports of the IPCC, where an unweighted mean of future model simulations is used to show a "best-guess" simulation of future climate, while the degree of model spread is used to estimate some measure of the significance of the result. There are more sophisticated methodologies that one may use for model combination, involving Bayesian methodologies [27] or model weighting [28], but the correct implementation and interpretation of such studies is subject to some debate. It has been shown that the ranking of model performance within a multi-model ensemble such as CMIP-3 is often highly dependent on the choice of metric used to evaluate the model. A metric based on the model's ability to reproduce observed variability will produce a different ranking than a metric which evaluates the model simulation of the mean state [30], and the performance of different models on these two metrics are very weakly correlated (Fig. 10.4). In addition, violation of the model "democracy" (one model, one vote) in the IPCC process is potentially controversial, as choices of how to weight models could be interpreted as a political statement [31].

The question of why multi-model means perform better than individual models is a complex one. Certainly, the mean is not in itself a self-consistent representation of a physical system and is therefore not subject to many of the restrictions that apply when tuning one model to reproduce an observed climate. As an example, a single model may be tuned in different ways to reproduce two different observed values "A" and "B," but it might be impossible to tune the model to reproduce "A" and "B" simultaneously. However, if different models in the ensemble make different choices about the relative importance of "A" and "B," it is likely that the ensemble mean will be close to the observed values in the case of a large ensemble. Clearly, real GCMs have a large number of observable diagnostics to

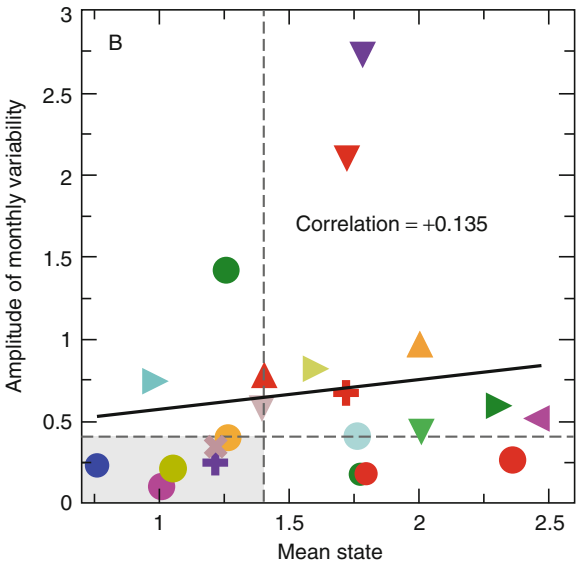

Fig. 10.4 The relationship between model skill in reproducing the mean climate state and skill in reproducing patterns of natural variability for models in the CMIP-3 ensemble. Each point represents a single model in the CMIP-3 archive, and errors are averaged over a large number of diagnostics. The *black line* shows the fitted least-squares regression (Reproduced from [29])

reproduce and a large number of tuning parameters, but it remains true that the multi-model mean is less restricted by model structure than any individual model in the ensemble. Another interpretation is that some of the model biases are random perturbations about the truth (i.e., each model reproduces the observations with some pattern of bias that is characteristic to that model and but different in each model), such that averaging many models reduces the magnitude of the biases. In the limit of completely random independent biases, the average would be perfect for an infinite number of models.

In some cases, the multi-model mean can indicate behavior unrepresentative of any of the models within the ensemble. Figure 10.5 shows the distribution of expected percentage precipitation change per unit global temperature increase in the current dry season for various models within the CMIP3 archive. Each individual model shows a wide distribution of change with some regions showing up to 30% decrease in precipitation for every degree rise in global mean temperature. If the models are averaged together in advance, however, the resulting multi-model mean has no region which displays this extreme decrease in precipitation in the dry season. The multi-model mean is thus not representative of the findings of the individual ensemble members in the respect that it fails to recover the extremes of the distribution of precipitation change. The reason for this discrepancy is, at least partially, a difference of the spatial representation of precipitation patterns in different ensemble members. Different models have different resolution, representations of orography, and parameterizations for precipitation. When combined this gives each models unique spatial modes of variability for precipitation. This allows each model to display extreme future drying in some specific regions, but critically those regions are not necessarily identical in all models in the

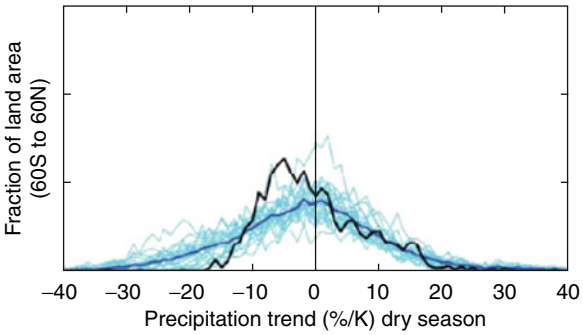

Fig. 10.5 This plot, from Knutti [31] shows the fraction of land area between 60°N and 60°S experiencing a given change in precipitation in the dry season. Precipitation change is measured in percent per unit global temperature rise in Kelvin measured over the period 1900–2100 relative to the 1900–1950 average. Each *light blue line* represents a single CMIP3 ensemble member, the *dark blue line* marks the average of all distributions. The *black line* shows the precipitation change in the multi-model mean. The expected absolute precipitation change in the multi-model mean is about 30% smaller than in any single model

ensemble, effectively smearing out the small scales and the extremes of the distribution. Thus, although the mean result of a large ensemble may provide a reduction in model bias, the averaging process itself may create an unrepresentative forecast.

Model Independence

Given a set of truly independent models distributed about the truth, one would always be able to improve simulation quality by increasing the number of models in the ensemble as truly independent errors would tend to cancel. Any study which treats CMIP ensemble members as independent realizations of a possible future is implicitly making this assumption, but one can make statistical arguments to show that the models are not distributed in a way which would be consistent with this assumption [36]. To illustrate this visually, Fig. 10.6 shows maps of temperature and precipitation from a selection of models in the CMIP-3 archive, all of which could be used with equal weight in producing a multi-model mean. However, one can see instantly that the two GFDL models have very similar biases in surface temperature, even though they are submitted as separate models to the archive. The temperature biases in the other two models shown have very different spatial patterns. The precipitation plots, however, show that there are some common biases in all four of the models. There are many reasons why these common biases might exist; all models in the CMIP-3 ensemble cannot explicitly resolve features smaller than about half a degree, which renders them incapable of simulating behavior such as atmospheric blocking or the response to local orography. Models may also share parameterization schemes and be tuned to reproduce the same observations [32],

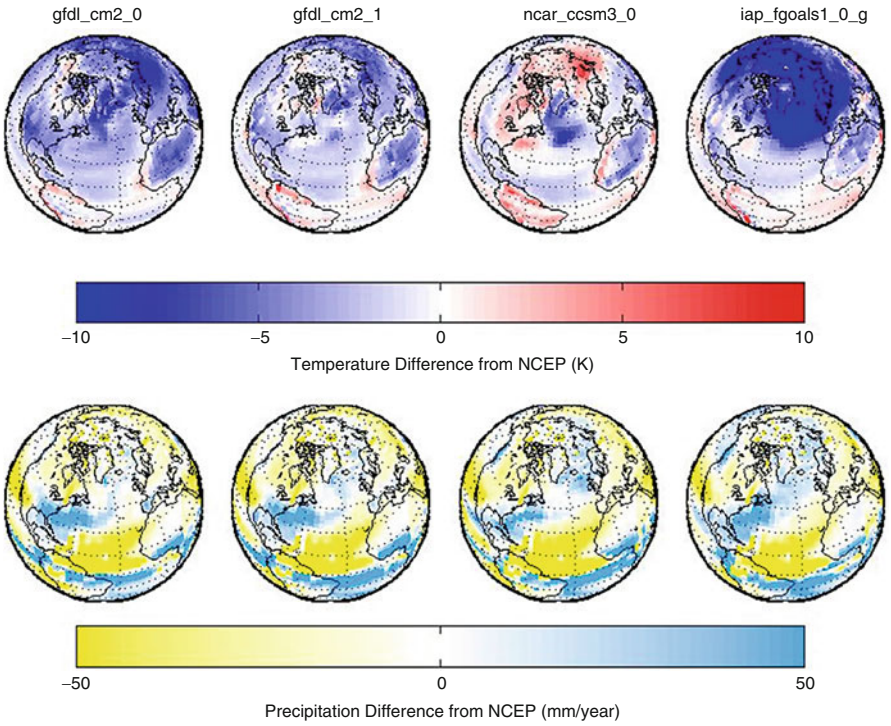

Fig. 10.6 Temperature and precipitation maps of the North Atlantic region from four models submitted to the CMIP-3 archive. Each map shows the 1980–2000 averages for June, July, and August – expressed as a difference between the model simulation and the NCEP reanalysis for the same period. The *top row* shows the anomaly for surface temperatures, while the *bottom row* shows the anomaly for annual total precipitation

and in some cases the same model can be submitted to the ensemble at multiple resolutions meaning that models can share considerable parts of code, making it very likely that model biases will be correlated. In summary, it is both expected and evident that the current generation of climate models does not provide an independent sample of estimates distributed about an underlying truth, and it is unlikely that increasing the number of similar models in the ensemble would drastically increase the accuracy of combined predictions.

Model Validation and Tuning

GCMs are frequently tuned by minimizing differences between simulations of the past century and observations. The observations can be in the form of data from satellites and in situ measurements or may be expressed as reanalyzed products

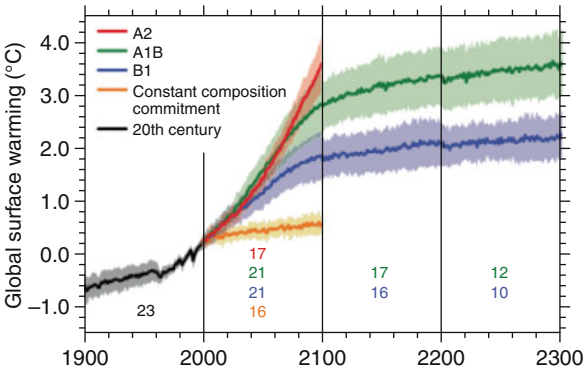

Fig. 10.7 A figure reproduced from the IPCC AR-4 report (Fig. 10.4) showing the mean and inter-model spread of simulations in the CMIP-3 model archive for simulations of the twentieth century, together with the simulations of three different scenarios for periods after the year 2000. Global mean temperatures are shown relative to the 1990–2000 mean. In each case, the *line* represents the multi-model mean and the shading shows the 1 standard deviation ensemble spread

which attempt to incorporate information from both of these. Simulations of earlier periods may also be evaluated against proxy data (estimates of temperature or rainfall etc., produced from tree rings, ice cores, etc.), although the long simulations and necessary model reconfiguration for these periods often mean they do not form part of the active model development process. Because models are tuned to agree with data over the twentieth century, they tend to agree with each other for this time period. There is little spread in the model simulations over the twentieth century. Figure 10.7 taken from the IPCC AR-4 report shows that the models behave similarly throughout the twentiethcentury when compared to any one of the scenarios for the twenty-first century. The reader should not attach any significance to the absolute values of the global mean temperature time series, which are expressed as anomalies with respect to the 1980–2000 mean for all models.

The remarkable consistency of the global mean temperature evolution in the twentiethcentury in the current generation of GCMs is made possible through the various degrees of freedom the models have in fitting this well-observed period. The response of any model is governed by a combination of transient ocean heat uptake, climate sensitivity, and the radiative forcing to the system, which effectively makes the problem poorly constrained with multiple ways to fit the twentieth century global mean temperature time series [33]. A study by Kiehl (2007) concluded that models produced this agreement by compensating between differences in climate sensitivity with differences in aerosol forcing. Figure 10.8 shows both the climate sensitivities and the later twentieth century anthropogenic forcing of climate in a selection of GCMs [34]. It is apparent that those models with a larger anthropogenic climate forcing in the twentieth century have a smaller climate sensitivity, allowing the models to successfully reproduce the twentieth century temperature record (a weak correlation between aerosol forcing and climate sensitivity is also seen in the CMIP-3 archive used for the AR4 report [35]).

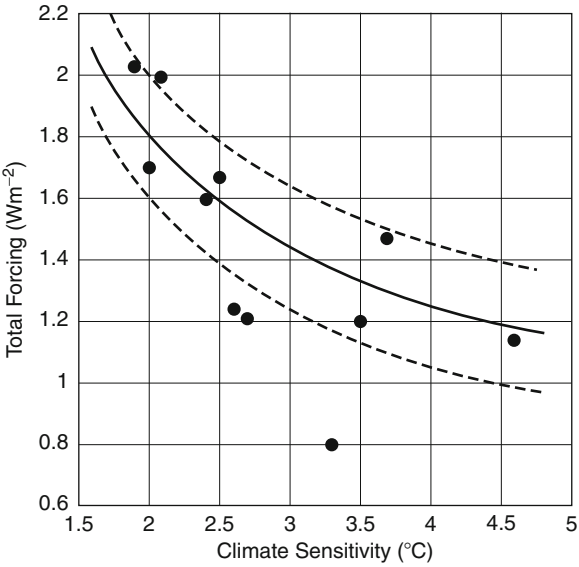

Fig. 10.8 A figure reproduced from Kiehl [35] which shows the relationship between climate sensitivity and total anthropogenic forcing of climate in the late twentieth century in a set of GCMs, represented by the *black dots*. The *solid line* represents a theoretical relationship between the two quantities necessary to produce the warming observed over the twentieth century, the *dashed lines* show the uncertainty in this relationship due to uncertainty in transient ocean heat uptake

In each of the twenty-first century scenarios illustrated in Fig. 10.7, the aerosol concentrations are predicted to decrease as increasingly stringent clean air legislation comes into effect. Meanwhile, all the scenarios show a continuing increase in greenhouse gases throughout the twenty-first century, which makes the climate sensitivity of the models the primary factor influencing their future evolution as the total anthropogenic forcing increases. The differing climate sensitivities amongst the CMIP3 models thus cause a larger spread in the twenty-first century simulations than for the twentieth century simulations. However, it should be noted that most AR4 models included the "direct" radiative effect of aerosols, but not their indirect effect on cloud properties. This means that eliminating the correlation between climate sensitivity and aerosol forcing would not necessarily reduce projection uncertainty, and the success of most models in simulating the twentieth century may be partly spurious [36].

An additional problem lies in the lack of independent data with which to tune and verify the models. In many cases, model quality metrics are based upon mean state and variability data from the latter twentieth century data, which is very likely to be used in the development of parameterizations and tuning of the model. For example, most models use satellite data products to tune the top of atmosphere energy fluxes, and these products are often considered to be one of the more robust constraints when evaluating a model quality metric. In addition, models may often be evaluated against reanalyses, rather than the observational data itself. Reanalysis products are model simulations strongly "nudged" to reproduce an incomplete set of observations, effectively filling in the gaps with self-consistent model data output. This process introduces an additional layer of complexity, because the reanalysis climate will contain features both of the constraining observations and the underlying model. For fields where real data is sparse or where data is not assimilated directly (such as precipitation metrics), the reanalysis output might

have much more dependence on the underlying model than on any real-world data. As a result, when using reanalysis data as a constraint for multiple models, those models with a similar representation of the hydrological cycle to that used in the reanalysis will appear to perform better.

The model development process involves a considerable amount of value judgment, as a model serves many purposes and some compromise between the many different plausible performance metrics must be made. The relatively small number of degrees of freedom available to model developers makes it impossible to perfectly match a large number of observable quantities simultaneously, which means that there may be multiple possible parameter combinations which are equally valid. Each of these combinations, although they fit historical observations equally well, may have different projections of future climate change if they exhibit different climate sensitivities or aerosol responses.

In the past, model tuning has largely been a time-consuming process of expert judgment and trial and error, which leads to some uncertainty of what errors in a simulation are irreducible through parameter adjustment. Although not yet used operationally, various techniques have been proposed to automate this tuning process. One technique uses an optimal gradient descent approach to minimize some multivariate error metric [34]. This approach can yield multiple solutions, as the response surface in the parameter space of the model may show local minima. Another approach involves using a preexisting perturbed physics ensemble and fitting a nonlinear response surface [37] to interpolate between the sampled points in the parameter space. This effectively produces a "model emulator" which can predict the point in parameter space which minimizes model error, but the result is dependent on the parameter space being sufficiently densely sampled to capture the dominant features. One can also combine the predictions from a range of plausible perturbed models. The ensemble Kalman filter [38] approach has been used [39] to create a set of valid perturbed versions of a single climate model, but is subject to uncertainty that there is an unknown systematic error in the climate model which cannot be corrected by parameter modification.

A final problem lies in the incompleteness of the model representation of the climate. Many current models, for example, cannot simulate the indirect effect of aerosols on cloud amounts. Tuning an incomplete model to reproduce the observed radiative balance at the top of the atmosphere therefore involves overcompensating the cloud amounts by artificially enhancing other processes, which arguably makes the representation of the current and future state less accurate.

Statements of Probability

Multi-model Ensembles

As indicated throughout this entry, the production of a probabilistic statement for future climate from a multimember or perturbed physics ensemble has no clearly established methodology and requires a priori assumptions to be made. Arguably

the simplest assumption that can be made is one of model equality, using the democratic "one model, one vote" approach [40]. In such a method, the probability of a future event is estimated by the fraction of models in which the event happens. This hypothesis can be tested by cross-validation within an unused subset of the ensemble. However, this approach is limited by the implicit assumption that the ensemble is a random sample of plausible estimates of the true climate, where the various arguments in section "Projection Uncertainty and the Need for Ensembles" suggest this assumption may not be valid.

The next logical step is therefore to consider some measure of model skill as a weighting for each model, producing an estimate of future climate as a median of model predictions, such that models with a small bias are given a greater weight [41]. Such approaches are always highly dependent on the exact choice of metric used to evaluate the model weighting [29].

Many studies have adopted Bayesian methodologies, where prior beliefs about the range of future climate change are updated with information from models and/or observations. One example [42] takes a prior probability distribution for current and future regionally averaged climate signal (or the corresponding climate change signal) and updates this using information from models and observations. Priors can be chosen to be uninformative (flat over a large range of possible values) so that the final PDF shape is mainly influenced by the information from models and observations. The likelihood of each model simulation of the past and the future is then represented as the realization from PDFs centered around the unknown "true" present and future climate, as if the ensemble were a sample from a large idealized population of possible models. The width of the PDFs is in turn estimated jointly with the climate signals. Its magnitude depends on that model bias compared to the consensus estimate of the present day and future state. Markov Chain Monte Carlo techniques are used to approximate the result of Bayes theorem applied to priors and likelihood, allowing a joint probability distribution for the "true" climate states and the unknown parameters characterizing the model distributions to be estimated. From it, the PDF of the regional climate change signal is also straightforward to derive.

A Bayesian approach has been also applied at the grid-point scale by representing the entire field of future climate anomaly for each model in terms of a truncated set of basis functions combined with some noise estimate [43], such that each model has its own low-dimensional set of coefficients to describe the pattern of climate change. The advantage of this approach is that a similar Bayesian methodology may be applied to derive estimates for the "true" values of the coefficients, which when recombined with the basic functions results in PDFs for climate change at the grid-point level.

An issue with both traditional weighting schemes and the Bayesian approaches is the way in which outliers are treated – the so-called "convergence criterion". In the case of a large PPE, such as the climateprediction.net experiment, the logic in down-weighting outliers assumes that there is some significance in the consensus mean projection, errors are distributed randomly and that models which deviate strongly from the consensus are somewhat less trustworthy. However, in a small ensemble of best-guesses such as CMIP-3, this argument is subject to question. It is possible that

a single model in the ensemble is able to simulate processes which are not simulated in other models. This model is arguably more trustworthy than the rest of the ensemble and yet it would be down-weighted through the application of a simple convergence criterion.

Another issue with all of the methods discussed thus far is the assumption of model independence. It can be shown [44] that the width of the final PDF using a Bayesian methodology is inversely proportional to the number of the models considered in the ensemble. Whilst this would be true if all models were independent estimates of a true climate, it has been demonstrated that this not a valid assumption [32]. Although some statistical methodologies have endeavored to artificially reduce the more obvious interdependencies of the CMIP-3 ensemble [45], there is at present no generally accepted methodology for doing so. The Bayesian techniques that have been developed so far tend to produce a PDF narrower than the spread of the original ensemble, as the independence assumption causes uncertainties to decrease with added ensemble members.

A completely different approach to producing model projections is to statistically "calibrate" models, where a relationship is established between model simulations and observations over an observed period. Once this relationship has been determined, it may be applied to future climate projections to produce a "calibrated" estimate of the true future response. This approach assumes, of course, that the relationship between the projections and the true response will remain constant in the future. This approach has been applied to large scale metrics such as past and future sea-ice loss [46], as well as more complex statistical multivariate approaches which find the best fitting relationship between modes of variability in model simulated and observed past climate, again using those relationships to produce a calibrated future projection [47, 48].

Finally, some "detection and attribution" studies [49] determine spatial patterns of climate changes associated with different atmospheric forcings, using observations to determine whether models are over- or under-representing those changes in past simulations. This allows future model projections to be rescaled in light of the observations. One of the major uncertainties in such approaches is in the derivation of the calibration coefficients themselves, and whether the calibration is valid when applied to a future planet in a very different state. These uncertainties tend to result in wider PDFs than Bayesian methodologies [33].

Perturbed Physics Ensembles

While "one model, one vote" may be a questionable assumption in a multimodel ensemble, it is quite ostensibly wrong in a perturbed physics ensemble where some models have vastly inaccurate simulations of the mean climate [50]. PDFs of future climate derived from a perturbed physics ensemble have therefore often been forced to take a different approach.

Most studies thus far arising from PPEs have focused on producing PDFs for climate sensitivity, and have broadly fallen into three categories: weighting of the parameter space, using the ensemble to establish relationships between observable quantities and unknowns such as climate sensitivity, or a traditional Bayesian technique. An example of the first approach [11] takes a PPE and ascribes each model a weighting, based upon model skill in reproducing the observed climate. By interpolating between the sampled points in the parameter space, one can then produce a weighted integral of the unknown quantity (e.g., climate sensitivity). It is argued, however [51], that the PDF obtained from such an approach is fundamentally dependent upon the prior assumptions made in sampling the original parameters.

A second approach of finding relationships between observable and unknown quantities has been demonstrated using both linear [52] and nonlinear [53] transfer functions. In each case, the ensemble is used to derive some predictors which internally estimate the climate sensitivities of ensemble members. These regression coefficients can then be used together with observations of the true climate state to make a prediction of the true climate sensitivity. Clearly, these predictions are subject to uncertainty in the observational state and in the internally derived prediction error, both of which may be estimated relatively easily. The major "unknown unknown" in such an approach is the systematic or irreducible error of the underlying model, that is, how much additional uncertainty arises when the predictor is applied to the real world. A lower bound of this quantity may be obtained by examining the skill of the predictor when applied to a multimodel ensemble such as CMIP-3, but this will not account for common errors arising from lack of resolution or simulated processes.

The final approach to be considered is the use of an ensemble Kalman filter [40]. The ensemble is used together with observations to update prior beliefs about several unknown model parameters. The ensemble Kalman filter then involves an iterative process forming an idealized ensemble of plausible perturbed models. Once again the methodology is sensitive to assumptions about model error, which scale the relative importance of the model-observation discrepancies forming the overall cost function. By assuming model errors are small, the resulting idealized ensemble will be more tightly clustered about the observed state. The distribution of climate sensitivities in this idealized ensemble is then deemed to approximate a PDF for the sensitivity. One advantage of such a technique is that the predictions may be validated by producing a hindcast for the past (the Last Glacial Maximum, in this case). The LGM simulation can then be used to produce an out of sample weighting for the optimized ensemble.

Future Directions

The analysis of climate simulations from multiple models is still a problem in its relative infancy. Various techniques have been proposed in this entry, each making different assumptions about model independence, prior distributions, systematic

model errors, and about what statistical framework is appropriate. These choices remain, at present, somewhat subjective and often yield different probability distributions for unknown climate variables. The apparent contradictions between the methodologies can be understood, however, in light of the assumptions made. In contrast to a numerical weather forecast where thousands of verification cases are available to test the forecast skill, the climate projections for a century into the future are making a statement about a situation never observed before and where no model evaluation is possible. Because there is only a single realization of the future, any statement of probability expresses a degree of belief in a Bayesian sense of how different future outcomes are supported by current evidence (models, data, methods), and is therefore inherently subjective.

Clearly, any projection (and the uncertainty associated with it) must be tailored in a fashion useful to decisions on policy and planning for a changing climate. Policymakers tend to push for increases in precision, but this can lead to decreases in real accuracy if predictions are overconfident [54]. There is arguably little point in providing PDFs of future change for planning purposes if the width of those PDFs are massively sensitive to either subjective decisions or unknown errors, and the raw collection of "best-guesses" from the different models is as useful a way as any to present the ensemble of forecasts. One inherent danger with this approach, however, is the tendency to see the multi-model distribution as a discrete probability distribution for future climate. As is seen, the lack of model independence, the fact that all models are neglecting certain sources of uncertainty (e.g., the carbon cycle climate feedback uncertainty) and the fact that every modeling group will tend to submit only a best-guess climate together implies that the true uncertainty may be larger than that indicated by the spread of model simulations.

Future generations of multi-model ensembles are also likely to introduce more complex "Earth System Models," at least for some ensemble members. These models, in addition to atmosphere, ocean, land, and sea ice components are likely to introduce fully coupled carbon–nitrogen cycles, chemistry, urban, and ecosystem models into the simulation. These components of future uncertainty have not been thoroughly explored in previous generations of the CMIP experiments, and are likely to increase the spread of simulation response for the coming century. Although this could be perceived to indicate an increase in uncertainty, it is more accurately converting an "unknown unknown" into a parametric uncertainty. If different models include different components of the earth system in their models, it will also become more difficult to compare them on a like-with-like basis, as is mostly possible today. However, this underlines the importance with each generation of climate models of recognizing the uncertainties associated with what is omitted, as well as those arising from the simulations themselves.

Although there may be some use in overall metrics of model skill [55], it is likely that projections of specific phenomena will benefit from tailored metrics to rank the performance of different models (e.g., El Niño or future sea ice extent). This will also require a proper assessment of which subset of models to use for each particular application based upon both past model skill and physical

plausibility [31]. In addition, the community may benefit more from a diverse range of model predictions, where each model may be evaluated on its own performance, in place of a group of models which are artificially clustered toward a mean response leaving no way of simulating the extremes or boundaries of future climate change.

In this entry both multi-model ensembles and perturbed physics ensembles have been discussed, but there is little discussion on how the information from the two may be combined. Indeed, at present there is little to no literature on how one may combine the parametric uncertainty sampled in a PPE with the inter-model systematic differences in a multi-model ensemble. This presents a fundamental problem in that current PDFs from both of these techniques cannot incorporate the best estimates of systematic and parametric uncertainty. Future analyses must combine these various uncertainties in order to make statements about model robustness. Currently, the ability to conduct such an analysis is limited because only a small subset of the models in the CMIP-3 archive have produced a perturbed physics ensemble, and for those ensembles which do exist, the experiments have not been conducted in any coordinated fashion.

Despite all of the challenges associated with combining and interpreting results from multiple climate models, the presence of coordinated ensembles of projections provides an invaluable insight into the magnitude of some of the uncertainties which are inherent in every simulation conducted, and the ensemble provides a unique opportunity to understand why models differ. As time goes on, the length of good quality observations will increase allowing better evaluation of the transient behavior of the models (a better metric for future transient response than those based upon the model simulation of the base climate [56]). In addition, as more components of the climate system are simulated, although model convergence is not expected (at least in the short term), one can be confident that at least "unknown unknowns" in future predictions can be represented in the form of parametric uncertainty.

Finally, possibly the greatest single uncertainty in future climate remains that of human behavior. Certainly in the case of the CMIP-3 ensemble, the spread in twenty-first century simulations due to different emission scenarios generally exceeded that of the inter-model spread to any particular scenario. Simple models of the climate have already been coupled to socioeconomic models [57–59], but little progress to date has been made in coupling socioeconomic models to GCMs. As a result, potential complex feedbacks between climate change and human behavior have not been sampled in any systematic framework. Nevertheless, although an integrated treatment of uncertainty in future climate projections may seem some way off, the use of multi-model ensembles will continue to frame at least some of those uncertainties in a systematic framework, providing a robustness which would be impossible with any single model, however complex that model may become.

Bibliography

Primary Literature

1. Ad Hoc Study Group on Carbon Dioxide and Climate (1979) Carbon dioxide and climate: a scientific assessment. National Academy of Sciences, Washington, DC
2. Manabe S et al (1979) A global ocean-atmosphere climate model with seasonal variation for future studies of climate sensitivity. Dyn Atmos Oceans 3:393–426
3. Hansen JE et al (1983) Efficient three-dimensional global models for climate studies: models I and II. Mon Weather Rev 111:609–662
4. Houghton JT, Jenkins GJ, Ephraums JJ (eds) (1991) Scientific assessment of climate change – report of working group I. Cambridge University Press, Cambridge, p 365
5. Houghton JT, Meira Filho LG, Callender BA, Harris N, Kattenberg A, Maskell K (eds) (1995) Contribution of working group I to the second assessment of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, p 572
6. Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Xiaosu D (eds) (2001) Contribution of working group I to the third assessment report of the Intergovernmental Panel on Climate Change (IPCC). Cambridge University Press, Cambridge, p 944
7. Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) (2007) Contribution of working group I to the fourth assessment report of the Intergovernmental Panel on Climate Change, 2007. Cambridge University Press, Cambridge/New York
8. Parker WS (2006) Understanding pluralism in climate modeling. Found Sci 11:349–368
9. Collins M, Allen MR (2002) Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. J Climate 15:3104–3109
10. Hawkins E, Sutton R (2009) The potential to narrow uncertainty in regional climate predictions. BAMS 90:1095–1107
11. Murphy JM et al (2004) Quantifying uncertainties in climate change from a large ensemble of general circulation model predictions. Nature 430:768–772
12. Stainforth DA et al (2005) Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature 433:403–406
13. Annan J, Hargreaves J (2006) Using multiple observationally-based constraints to estimate climate sensitivity. Geophys Res Lett 33(4):L06704
14. Palmer TN, Doblas-Reyes FJ, Hagedorn R, Weisheimer A (2005) Probabilistic prediction of climate using multi-model ensembles: from basics to applications. Philos Trans R Soc B 360:1991–1998
15. Weigel AP, Liniger MA, Appenzeller C (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q J R Meteorol Soc 134:241–260
16. Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: basic concept. Tellus 57A:219–233
17. Frame DJ, Booth BBB, Kettleborough JA, Stainforth DA, Gregory JM, Collins M, Allen MR (2005) Constraining climate forecasts: the role of prior assumptions. Geophys Res Lett 32: L09702. doi:10.1029/2004GL022241
18. van der Sluijs J et al (1998) Anchoring devices in science for policy: the case of consensus around climate sensitivity. Soc Stud Sci 28(2):291–323
19. Knutti R, Hegerl GC (2008) The equilibrium sensitivity of the Earth's temperature to radiation changes. Nat Geosci 1:735–743
20. Cantelaube P, Terres J-M (2005) Seasonal weather forecasts for crop yield modelling in Europe. Tellus Ser A 57:476–487. doi:10.1111/j.1600-0870.2005.00125.x
21. Thomson MC, Doblas-Reyes FJ, Mason SJ, Hagedorn R, Connor SJ, Phindela T, Morse AP, Palmer TN (2006) Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature 439:576–579. doi:10.1038/nature04503

22. Schclar A et al (2009) Ensemble methods for improving the performance of neighborhood-based collaborative filtering. In: Proceedings of the third ACM conference on recommender systems, ACM, New York, 23–25 Oct 2009, pp 261–264
23. Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. Tellus A57:219–233. doi:10.1111/j.1600-0870.2005.00103.x
24. Gillett NP, Zwiers FW, Weaver AJ, Hegerl GC, Allen MR, Stott PA (2002) Detecting anthropogenic influence with a multi-model ensemble. Geophys Res Lett 29:1970. doi:10.1029/2002GL015836
25. Lambert SJ, Boer GJ (2001) CMIP1 evaluation and intercomparison of coupled climate models. Clim Dyn 17:83–106. doi:10.1007/PL00013736
26. Reichler T, Kim J (2008) How well do coupled models simulate today's climate? Bull Am Meteorol Soc 89:303–311
27. Robertson AW, Lall U, Zebiak SE, Goddard L (2004) Improved combination of multiple atmospheric GCM ensembles for seasonal predition. Mon Weather Rev 132:2732–2744. doi:10.1175/MWR2818.1
28. Krishnamurti TN, Kishtawal CM, Zhang Z, Larow T, Bachiochi D, Williford E, Gadgil S, Surendran S (2000) Multimodel ensemble forecasts for weather and seasonal climate. J Climate 13:4196–4216. doi:10.1175/1520-0442(2000) 013!4196:MEFFWAO2.0.CO;2
29. Santer BD, Taylor KE, Gleckler PJ, Bonfils C, Barnett TP, Pierce DW, Wigley TML, Mears C, Wentz FJ, Brüggemann W, Gillett NP, Klein SA, Solomon S, Stott PA, Wehner MF (2009) Incorporating model quality information in climate change detection and attribution studies. PNAS 106:14778–14783
30. Knutti R (2010) The end of model democracy? Clim Change 102(3–4):395–404. doi:10.1007/s10584-010-9800-2
31. Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. J Climate 23:2739–2758
32. Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc A 365(1857):2053–2075
33. Knutti R, Stocker TF, Joos F, Plattner G-K (2002) Constraints on radiative forcing and future climate change from observations and climate model ensembles. Nature 416:719–723. doi:10.1038/416719a
34. Jackson C et al (2004) An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions. J Climate 17(14):2828–2841
35. Kiehl JT (2007) Twentieth century climate model response and climate sensitivity. Geophys Res Lett 34:22710
36. Knutti R (2008) Why are climate models reproducing the observed global surface warming so well? Geophys Res Lett 35(18):5
37. Sanderson BM et al (2008) Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. J Climate 21(11):2384–2400
38. Evensen G (2003) The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dyn 53:343
39. Annan JD et al (2005) Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter. Ocean Model 8:135
40. Raisanen J, Palmer TN (2001) A probability and decision-model analysis of a multimodel ensemble of climate change simulations. J Climate 14(15):3212–3226
41. Giorgi F, Mearns LO (2002) Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the 'reliability ensemble averaging' (REA) method. J Climate 15:1141
42. Tebaldi C et al (2004) Regional probabilities of precipitation change: a Bayesian analysis of multimodel simulations. Geophys Res Lett 31:24213
43. Furrer R, Sain S, Nychka D, Meehl G (2007) Multivariate Bayesian analysis of atmosphere–ocean general circulation models. Environ Ecol Stat 14:249–266

44. Lopez A et al (2006) Two approaches to quantifying uncertainty in global temperature changes. J Climate 19:4785
45. Smith R, Tebaldi C, Nychka D, Mearns L (2009) Bayesian modeling of uncertainty in ensembles of climate models. J Am Stat Assoc 104:97–116
46. Boé J et al (2009) September sea-ice cover in the Arctic ocean projected to vanish by 2100. Nat Geosci 2(4):1–3
47. Greene AM et al (2006) Probabilistic multimodel regional temperature change projections. J Climate 19:4326
48. Buser CM et al (2009) Bayesian multi-model projection of climate: bias assumptions and interannual variability. Clim Dyn 33(6):849–868
49. Stott PA, Kettleborough JA (2002) Origins and estimates of uncertainty in predictions of twenty first century temperature rise. Nature 416:723–726
50. Sanderson BM et al (2008) Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics GCM simulations. Clim Dyn 30(2–3):175–190
51. Frame DJ et al (2005) Constraining climate forecasts: the role of prior assumptions. Geophys Res Lett 32(9):L09702
52. Piani C et al (2005) Constraints on climate change from a multi-thousand member ensemble of simulations. Geophys Res Lett 32(23):L23825
53. Knutti R et al (2006) Constraining climate sensitivity from the seasonal cycle in surface temperature. J Climate 19(17):4224–4233
54. Dessai S et al (2008) In: Adger N, Lorenzoni I, O'Brien K (eds) Climate prediction: a limit to adaptation. Living with climate change: are there limits to adaptation. Cambridge University Press, Cambridge, pp 49–57
55. Gleckler PJ et al (2008) Performance metrics for climate models. J Geophys Res 113:D06104
56. Allen MR, Frame DJ (2007) ATMOSPHERE: call off the quest. Science 318:582–583
57. Edmonds J, Wise M, Pitcher H, Richels R, Wigley T, MacCracken C (1997) An integrated assessment of climate change and the accelerated introduction of advanced energy technologies. Mitig Adapt Strateg Glob Change 1:311–339
58. Messner S, Strubegger M (1995) User's guide for MESSAGE III, WP-95-69. International Institute for Applied Systems Analysis, Laxenburg
59. Bouwman AF, Kram T (2006) Integrated modelling of global environmental change. An overview of IMAGE 2.4. Netherlands Environmental Assessment Agency (MNP), MNP publication number 500110002/2006, Bilthoven

Books and Reviews

Kharin VV, Zwiers FW (2002) Climate predictions with multimodel ensembles. J Climate 15(7):793–799

Knutti R et al (2008) A review of uncertainties in global temperature projections over the twenty-first century. J Climate 21:2651–2663