Climate Change Projections: Multi-Model Ensembles

Ben Sanderson¹, Reto Knutti²

(1) National Center for Atmospheric Research, Boulder, CO, USA

(2) Eidgenössische Technische Hochschule, Zürich, Switzerland

1. Glossary

Bayes' Theorem - a law in probability theory which relates the probability of a hypothesis given observed evidence to its inverse, the probability of that evidence given the hypothesis.

Climate Sensitivity - The global mean temperature response in Kelvin to an instantaneous doubling of carbon dioxide)

CMIP-3 - The Coupled Model Inter-comparison Project, a set of coordinated experiments using GCMs from the world's major modeling centers

Detection and Attribution - a process whereby spatial ‘fingerprints’ associated with individual climate forcing factors are identified and used to relate forcing signals in a period of climate change.

Empirical Model - A model which makes no attempt to justify its representations of the system with any physical basis.
**General Circulation Model** - a three dimensional mathematical model for the atmosphere and possibly the ocean, land and cryosphere.

**Last Glacial Maximum** - a period lasting several thousand years peaking approximately 20,000 years ago at the maximum extent of the ice sheets.

**Parameter Space** - the multidimensional domain created by considering all possible values of uncertain parameters within a model.

**Perturbed Physics Ensemble** - a group of climate simulations formed by taking a single General Circulation Model and altering uncertain parameters within a range of physical plausibility.

**Systematic Error** - an inaccuracy in a model simulation which is not reducible by increased model sampling or parameter tuning.

2. **Article Outline (Table of contents)**

   Definition of the subject and its importance ........................................ 2
   Introduction............................................................................................................. 2
   Projection uncertainty and the need for ensembles ................................. 2
     *Empirical and physical models* ................................................................. 2
     *Types of uncertainty and the need for ensembles* ................................. 2
   Multi-model and perturbed physics ensembles...................................... 2
     *Range of ensemble responses* ................................................................. 2
     *The ensemble mean state* ........................................................................... 2
3. Definition of the subject and its importance

General Circulation Models of the Earth system provide comprehensive tools for simulating the future response to Earth’s changing boundary conditions. But inherent in the design of such models is a myriad of choices which must be made, in terms of deciding which components of the system are to be modeled and how to represent processes which are not modeled in a consistent fashion. In recent years, a large 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 finer scales. 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 new a 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?

4. Introduction

In 1979, Jule Charney chaired a committee on anthropogenic global warming, producing a report\(^1\) which provided an overview of the understood science of the day. 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 of global mean temperatures to a doubling of 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 CO2-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 of those models are still as relevant to the ensemble modeling efforts of today. An ensemble brings confidence if somewhat independent models can produce common features in their simulations, and if the origins of the differences between simulations may be traced back to physical characteristics - then we can claim to have a better understanding in the uncertainties in the simulations.
When presented with a range of simulations of future climate, one must implicitly make several judgments 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 should be treated as equals, and if not then how should we distinguish between them? If some processes are absent from some or all of the simulations, how can we update our projections of the future to account for these “unknown unknowns”? Should one judge the ensemble members to be estimations 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 very real effects on the results obtained and the degree of uncertainty in those results.

In recent years, there have been systematic efforts to produce large ensembles of increasingly complex models of the earth system which are used to simulate coordinated experiments and analyzed by 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 hypothetical scenarios of climate change. Through the successive decades, model complexity and scope has 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), models also included volcanic activity and a sulphur cycle, with a dynamic ocean - the first Atmosphere-
Ocean General Circulation Models (AOGCMs). In the Third Assessment Report (TAR), some models introduced a representation of the Carbon Cycle, atmospheric aerosols and an overturning circulation. Finally, the most recent Fourth Assessment Report (AR-4) included some models with interactive vegetation and full atmospheric chemistry.

This chapter focuses on discussing the advantages and additional complexities which one must consider when studying a range of simulations rather than a single model output. Rather than giving a comprehensive description of the results of the models assessed in the successive generations of the IPCC, this chapter 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.

5. Projection uncertainty and the need for ensembles

Empirical and physical models

In 150 AD, Ptomely 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 in closer to our established view of the universe,
both of these models were *empirical* in that they were not based on any physical principles. However, without a physical basis, Galileo was able to validate the Copernican model by studying the phases of the planet Venus - which were 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.

A model of any physical system is an approximated 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 shown to be a wholly correct representation of the true system, it can only be validated by reproducing some output not used in the tuning of the model itself. This is true of Galileo’s observation of the phases of Venus - information not used in the tuning of the Ptolemy 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 we take the simplest, zero dimensional empirical model of the climate imaginable, we have:

\[ C \frac{dT'}{dt} = F' - \lambda T' \]
where $T'$ is the global mean temperature difference from an equilibrium state, $F'$ is the additional radiative forcing to the planet, $C$ is the effective heat capacity of the system and $\lambda$ is the global sensitivity parameter. This equation has two free parameters, $C$ and $\lambda$ which may be tuned such that the model can fit an observed past timeseries of $F'$ and $T'$, that of the 20th Century, for example. The model can then be validated by predicting a previously unseen time period, such as the last glacial maximum. This validation, 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 training and validation.

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, like a Newtonian model of the solar system, which we believe can reproduce observed climate by coupling underlying physical laws. However, this view is often overoptimistic. Although some components of the modeled climate, such as the instantaneous radiative forcing due to a change in atmospheric carbon dioxide concentrations are relatively well understood and consistently implemented in different GCMs, there are some other processes such as convection which cannot explicitly be globally resolved with current computing resources. These processes are approximated with some uncertain parameters, which 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 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 which are encountered with an empirical model, with the
possibility of false confidence in model performance by over-tuning the model to reproduce past climates.

**Types of uncertainty and the need for ensembles**

The uncertainties in a simulation of future climate arises from unknown initial conditions, future boundary conditions, parameter values and systematic errors. In the multi-decadal climate problem, initial condition uncertainty soon becomes unimportant as the ‘memory’ of the climate system ranges from days for weather systems over land to perhaps 10 years for North Atlantic ocean temperatures. This places a fundamental lower limit on the accuracy of any multi-decadal prediction, as this ‘first-kind’ uncertainty is impossible to reduce without a perfect knowledge of the initial state. In practice, the initial condition uncertainty is evaluated by repeating simulations with a range of different initial conditions to form ‘Initial Condition Ensembles’, which help to isolate which parts of a simulated future climate trajectory are due to internally generated noise and which parts are due to changing boundary conditions.

The second kind of uncertainty relates to the boundary conditions of the problem. In the case of the long-term climate forecast problem, this is partly a question of predicting future anthropogenic emissions of forcing agents which is addressed by considering a range of plausible future socioeconomic scenarios. There are clearly also non-anthropogenic aspects to the future boundary conditions such as forcing due to volcanic activity and changing solar activity, but these represent a relatively small fraction of the total anthropogenic climate forcing and the current state of research does not permit any real skill in forecasting these quantities.
In most cases, it is uncertainties in the accuracy of the climate model itself which are most problematic in evaluating the accuracy in any prediction. A degree of uncertainty arises when a climate model contains parameterizations for processes which cannot be explicitly resolved at the resolution of the model. Parameterizations take large scale quantities which are resolved by the model, such as temperature, wind speed and humidity, and relate them to small scale processes, such as convective mass flux and cloud profiles. Although these parameterizations are usually constructed from physical underpinnings, they introduce some unavoidable uncertainty when there are a range of parameter values which might be physically plausible. GCMs are often subject to a tuning of parameter values to reproduce features of the observed climate, but given tens or hundreds of uncertain parameters this process can be time-consuming and can yield multiple solutions.

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\textsuperscript{9,10,11} and involves taking a selection of unknown parameters within the GCM and perturbing them within the bounds of physical plausibility. This will produce a selection of models with a range of base state climates and differing responses to changes in climate forcing. Incorporating this range into an uncertainty estimate for predictions of future climate requires a framework for joint consideration of each model’s performance and its future response.

The remaining model uncertainties are due to so-called ‘systematic’ or structural errors which arise the design of the model itself, i.e. the choices of which
processes to model, the resolution of the model, numerical schemes used 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.

6. Multi-model and perturbed physics ensembles

When making predictions of a future, unknown climate state; there is a wealth of evidence to suggest that considering an combined prediction using multiple, somewhat independent models yields more accurate results than any single model\(^1\), together with the spread of simulations which provides a measure of robustness in the prediction. The following section describes some rationale for the performance of multi-model and perturbed physics ensemble forecasts, together with some of the complexities which arise in 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 behavior which might be observed within the
ensemble. For example, if a PPE is created by perturbing cloud parameters in a GCM which has no parameterization for methane release from arctic tundra, then there is no way that such an ensemble can include uncertainty in that potential feedback. 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\(^{15}\). In contrast, multi-model ensembles such as the Coupled Model Inter-comparison Project (CMIP-3) are not subject to any experimental design, they 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 (the global mean temperature response in Kelvin to an instantaneous doubling of carbon dioxide) in both a multi-model and a perturbed physics ensemble (Figure 1). 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 though the quality of their base climate simulation may vary massively. Although impossible to verify, it also is possible that there is also a social component which draws multi-model sensitivities towards the mean value as any group which finds that their model is 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.
Figure 1: Distributions of Climate Sensitivity for models in the CMIP-3 ensemble, compared with a selection of models from the climateprediction.net project. Box and whisker plots show estimates the most likely values, together with 66th and 90th percentiles of likelihood for climate sensitivity taken from various lines of observational evidence (adapted from Knutti and Hegerl (2008)\textsuperscript{16}). Histograms represent the fraction of models in each 0.5K bin of climate sensitivity for the atmosphere-only components of 19 models in the CMIP-3 archive and for a 2000 member subset of the climateprediction.net ensemble\textsuperscript{10}.

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\textsuperscript{17}, disease modeling\textsuperscript{18} and in the optimization routines used for movie recommendation based upon past viewing choices\textsuperscript{19}.
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\(^{13}\). A multi-model study incorporating a set of initial conditions for each model is referred to as a ‘super-ensemble’. The accuracy of the model mean often shows best in multivariate applications, i.e. 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 ensemble member\(^{20}\).

This effect can also be seen in GCM simulations of recent past climate. Figure 2 shows successive generations of the CMIP ensemble evaluated using a multivariate error metric which compares 20th century observations to model simulations of that period for a variety of model diagnostics. The figure shows that model errors have decreased over time but also that for each generation of the ensemble, the multi-model mean results in a model-data discrepancy which is almost as good, or better than the best performing ensemble member. Various studies have found that both in detection and attribution studies\(^{21}\) and in simulations of recent climate\(^{22}\) that a multi-model mean provides a better multi-variate simulation than any individual model.
Figure 2: A comparison of model errors reproduced from Reichler and Kim (2008)\textsuperscript{23}, multivariate errors are evaluated for successive generations of the Coupled Model Inter-comparison 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 pre-industrial control simulations when evaluated against present-day observations.

This approach is common in the reports of the IPCC, where an unweighted mean or future model simulations is used to show a ‘best-guess’ simulation of future climate, while the degree of model spread is used to estimate the regional significance of the result. There are more sophisticated methodologies that one may use for model combination, involving Bayesian methodologies\textsuperscript{24} or model weighting\textsuperscript{25} but the implementation of such schemes 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 ability to reproduce observed variability will produce a different result to a metric which evaluates the model simulation of the mean state\textsuperscript{26} (Figure 3). 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.

Figure 3. 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 Santer et al (2009)26.

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 means 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 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.

Figure 4: This plot, from Knutti (submitted)\textsuperscript{27} shows the fraction of land area between 60N and 60S which observe 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 blue line represents a single CMIP3 ensemble member, while the black line shows the precipitation change in the multimodel mean. The expected precipitation change in the multimodel mean is about 30% smaller than in any single model.
In some cases the multi-model mean can indicate behavior which is unrepresentative of any of the models within the ensemble. Figure 4 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 percent 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 regions which display 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 resolutions, 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 the same in each model in the 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, and one can make various arguments
to the contrary. Figure 4 shows that the assumption that model errors are
uncorrelated and distributed about a ‘true’ mean is not correct in the case of
surface temperatures for models in the CMIP3 archive. If models in the CMIP
archive had independent errors, the mean bias of the simulated surface
temperatures should fall off as 1/N where N is the number of members in the
ensemble. However, when taking a random sample of members of the CMIP
archive, errors actually appear to fall off as 1/(N+B), where B is a systematic bias
common to all of the models in the ensemble. There are many reasons why this
bias term 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, and in some cases the same model can be
submitted to the ensemble at multiple resolutions which means that models can
share considerable parts of code, making it very likely that model biases will be
correlated.
Figure 5: Reproduced from Knutti (submitted) showing the Root Mean Square Bias of a multi-model mean of surface temperature values as a function of the number of CMIP3 models averaged together (shown for December-February and June-August seasons). The dotted line shows the 1/N dependency which would be expected if model errors were independently distributed about the truth. The dashed red lines show the upper bound in model bias, taking the worst performing N models and averaging them. The dashed blue line shows the case with the best N models averaged together. The solid red line takes an average of all possible sequences of choosing N models from the 22 member ensemble. The dashed black line fits a 1/(N+B) dependency to the mean bias, where B is the systematic bias of the ensemble.

In summary, it is both expected and evident that the current generation of climate models is not 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.
7. Model validation and Tuning

GCMs are frequently tuned by minimizing the bias associated with simulations of the past century compared to observable data from satellites and ground stations and reanalyses which attempt to incorporate information from both of these. Simulations of earlier periods may also be evaluated against proxy data, although the long simulations and necessary model reconfiguration for these periods often means they do not form part of the active model development process. In practice, this often means that there is surprisingly little spread in the model simulations over the well observed ‘satellite era’ in the late 20th Century. Figure 6, taken from the IPCC AR-4 report shows that the inter-model spread in global mean simulations is very small in the latter part of the 20th Century, especially when compared to the inter-model spread in any one of the scenarios for the 21st Century.
Figure 6: A figure reproduced from the IPCC AR-4 report (Figure 10.4) showing the mean and inter-model spread of simulations in the CMIP-3 model archive for simulations of the 20th 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.

The remarkable consistency of the global mean warming trend in the latter 20th Century 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 20th Century global mean temperature time-series. One of the most common ways
this cancellation of errors can occur is in the compensation of differences in climate sensitivity with differences in aerosol forcing. Figure 7 shows both the climate sensitivities and the latter 20th Century anthropogenic forcing of climate in a selection of GCMs in the CMIP3 archive\textsuperscript{30}. It is apparent that those models with a larger anthropogenic climate forcing in the 20th Century also have a smaller climate sensitivity, allowing the models to successfully reproduce the 20th Century temperature record. It can be shown\textsuperscript{30} that the majority of the spread in the GCM anthropogenic forcing arises from differences in the simulation of the aerosol direct and indirect effects, which are subject to a large degree of uncertainty. However, in each of the 21st Century scenarios illustrated in Figure 6, 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 21st 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 21st Century simulations than for the 20th Century simulations.
Figure 7: A figure reproduced from Kiehl (2007) which shows the relationship between climate sensitivity and total anthropogenic forcing of climate in the late 20th Century in a selection of GCMs used in the CMIP3 ensemble, represented by the black dots. The solid line represents a theoretical relationship between the two quantities necessary to produce the warming observed over the 20th Century, the dashed lines show the uncertainty in this relationship due to uncertainty in transient ocean heat uptake.

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 20th Century, data which is very likely to be used in the 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 which are strongly ‘nudged’ to reproduce a spatially 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 (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.

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. 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 pre-existing perturbed physics ensemble and fitting a nonlinear response surface to interpolate between the sampled points in the parameter space. This effectively produces a ‘model emulator’ where 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. Another approach is to
combine the predictions from a range of plausible perturbed models. The ensemble Kalman filter\textsuperscript{33} approach has been used\textsuperscript{34} 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. The current generation of GCMs, for example, typically ignores cirrus clouds generated by orographic waves induced by flows over mountains\textsuperscript{35}, and yet in the real world these clouds definitely exist. Tuning an incomplete model to reproduce the observed radiative balance at the top of the atmosphere therefore involves overcompensating the cirrus cloud formation by artificially enhancing other processes, which arguably makes the representation of the current and future state less accurate.

8. **Statements of Probability**

**Multi-model ensembles**

As indicated throughout this chapter, the production of a probabilistic statement for future climate from a multi-member or perturbed physics ensemble has no clearly established methodology and requires \textit{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\textsuperscript{36}. 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 4 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 or model predictions, such that models with a small bias are given a greater weight\textsuperscript{37}. Such approaches are always highly dependent on the exact choice of metric used to evaluate the model weighting \textsuperscript{26}.

Another approach to the problem is to take the Bayesian standpoint\textsuperscript{38}, where a prior probability distribution for a regional climate change signal is updated by information from models and observations. If the prior is taken to be uninformative, it is uniform between zero and infinity. Each model simulation of the past and the future is then represented by a normally distributed PDF centered around an unknown ‘true’ present and future climate. The width of each of the model PDF is increased if that model shows a large bias compared to an optimal estimate of the present day state, or if the model future prediction deviates from the optimal estimate of the future state. Bayes theorem allows a probability distribution for the ‘true’ climate states to be expressed as a function of these component PDFs, producing the optimal estimates for the present and future state, so the problem is self referential and must be solved iteratively to produce an estimate for the PDF of future climate change.

The Bayesian approach can be applied at the grid-point scale by representing the future climate anomaly for each model in terms of a truncated set of basis functions combined with some noise estimate\textsuperscript{39}, such that each model has its own small 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 basis functions results in PDFs for climate change at the gridpoint 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 vary largely from the consensus are somewhat less trustworthy. However, in a small ensemble of best-guess such as CMIP-3, this argument is subject to question. It becomes 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 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. Although some statistical methodologies have endeavored to artificially reduce the more obvious interdependencies of the CMIP-3 ensemble, there is at present no generally accepted methodology for doing so.

A completely different approach to model calibration is to statistically 'calibrate' models, where a linear relationship is established between regional model
predictions and observations over the 20th century. 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.

It has been noted that the various approaches detailed above often result in differing estimates for PDFs of future climate change. The Bayesian techniques tend to indicate a smaller spread than that of the original ensemble, as the independence assumption causes uncertainties to decrease with added ensemble members. The calibration approach tends to produce wider PDFs owing to uncertainty in the calculation of the calibration coefficients.

**Perturbed Physics Ensembles**

While ‘one model, one vote’ may be a questionable assumption in a multi-model ensemble, it is quite ostensibly wrong in a perturbed physics ensemble where some models are have vastly inaccurate simulations of the mean climate. 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 former approach 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, that the PDF obtained from such an approach is fundamentally dependent upon the prior assumptions made in sampling the original parameters.

The alternative approach of finding relationships between observable and unknown quantities has been demonstrated using both linear and nonlinear transfer functions. In each case, the ensemble is used to derive some predictors which can be used to 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, i.e. 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 multi-model 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. 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 which forms 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 which are combined to produce 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.

9. 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 chapter, each making different assumptions about model independence, prior distributions, systematic model errors and whether a Bayesian or frequentist treatment 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, if the significance of the various assumptions are well understood.

Clearly, any projection (and the uncertainty associated with it) must be tailored in a fashion which is useful to decisions on policy and planning for a changing climate. 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 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 we have seen, the lack of model independence and the fact that every modeling group will tend to submit only a best-guess climate together implies that the true uncertainty is likely much 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, 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.

In this chapter we have discussed both multi-model ensembles and perturbed physics ensembles, but little 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. As time goes on, the length of good quality, consistent observations will increase allowing increased verification of the transient behavior of the models, which it could be argued is a more relevant constraint on future transient response than the model simulation of the base climate. In addition, as more components of the climate system are simulated, although we should not expect model convergence (at least in the short term), we 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 21st century simulations due to different emission scenarios generally greatly exceeded that of the inter-model spread to any particular scenario. Simple models of the climate have already been coupled to socio-economic models.
51, but little progress to date has been made in coupling socio-economic 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.

10. Further Reading


11. Bibliography

5 Contribution of Working Group I to the Second Assessment of the Intergovernmental Panel on Climate Change, JT Houghton, LG Meira Filho, BA Callender, N Harris, A Kattenberg and K Maskell (Eds), Cambridge University Press, UK. pp 572
Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), J. T. Houghton, Y. Ding, D.J. Griggs, M. Noguer, P. J. van der Linden and D. Xiaosu (Eds.), Cambridge University Press, UK. pp 944


Using multiple observationally-based constraints to estimate climate sensitivity

James Annan, Julia Hargreaves, 2006, Geophysical Research Letters, 33, (4), L06704


Constraining climate forecasts: The role of prior assumptions


27 Challenges in combining projections from multiple climate models, Reto Knutti, Reinhard Furrer, Claudia Tebaldi, Jan Cermak, and Gerald A. Meehl (submitted)
51 Integrated modelling of global environmental change. An overview of IMAGE 2.4 © Netherlands Environmental Assessment Agency (MNP), Bilthoven, October 2006 MNP publication number 500110002/2006