PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2007) **365**, 2053–2075
doi:10.1098/rsta.2007.2076
Published online 14 June 2007

Review

The use of the multi-model ensemble in probabilistic climate projections

By Claudia Tebaldi[1,*] and Reto Knutti[2]

[1]*Institute for the Study of Society and Environment, National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80304, USA*
[2]*Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology, Universitätstrasse 16 (CHN N 12.1), 8092 Zürich, Switzerland*

Recent coordinated efforts, in which numerous climate models have been run for a common set of experiments, have produced large datasets of projections of future climate for various scenarios. Those multi-model ensembles sample initial condition, parameter as well as structural uncertainties in the model design, and they have prompted a variety of approaches to quantify uncertainty in future climate in a probabilistic way. This paper outlines the motivation for using multi-model ensembles, reviews the methodologies published so far and compares their results for regional temperature projections. The challenges in interpreting multi-model results, caused by the lack of verification of climate projections, the problem of model dependence, bias and tuning as well as the difficulty in making sense of an 'ensemble of opportunity', are discussed in detail.

Keywords: regional climate change; probabilistic projections; multi-model ensembles; global climate models; structural uncertainty; performance-based weighting

1. Introduction

(*a*) *Sources of model uncertainty*

Predictions and projections of weather and climate from time scales of days to centuries usually come from numerical models that resolve or parametrize the relevant processes. Uncertainties in constructing and applying these models are manifold, and are often grouped into initial condition, boundary condition, parameter and structural uncertainties.

Initial condition uncertainty is most relevant for the shortest time scales. Weather is chaotic, and predictions are sensitive to the value of observations used to initialize numerical models (e.g. Palmer 2005). Long-term projections of climate change are typically averaged over decades and often across several ensemble members, and are thus largely insensitive to small variations in initial conditions. Even when sequences of daily output from climate models are fed into

[*] Author for correspondence (tebaldi@ucar.edu).

One contribution of 13 to a Theme Issue 'Ensembles and probabilities: a new era in the prediction of climate change'.

impact models (e.g. ecosystem models or crop models), it is expected that the choice of a specific model run within an ensemble generated by perturbed initial conditions would not produce significantly different outcomes from an alternative choice within the same ensemble, when summary statistics of climate-driven results are computed. On the other hand, some climate models show multi-decadal climate variability related to the behaviour of the Atlantic meridional overturning circulation, and in those cases the ocean initial state can affect projections over several decades (Bryan *et al.* 2006). Other components of the climate system like soil properties or ice-sheet behaviour may contribute to these long-memory effects, but for now these remain speculative.

Boundary condition uncertainty is introduced if datasets are used to replace what in reality is an interactive part of the system, e.g. if sea surface temperature and sea ice cover are prescribed in an atmosphere-only model, or if radiative forcing (e.g. changes in solar insolation, changes in atmospheric concentrations of greenhouse gases) is prescribed over time. With regard to prescribed concentrations of greenhouse gases, the main source of uncertainty resides in the assumptions over the future world economic and social development, leading to alternative scenarios of greenhouse gas emissions whose relative likelihood cannot be easily assessed.

Parameter uncertainties are discussed in detail in several other papers in this issue. They stem from the fact that, mostly for computational constraints, small-scale processes in all components of the climate system cannot be resolved explicitly, and their effect on the large-scale climate response must be parametrized with bulk formulae that depend on large-scale variables available in the model. Parameter uncertainties can be explored and quantified by perturbed physics ensembles (PPE), sets of simulations with a single model but different choices for various parameters. Many of the attempts to quantify climate change or climate model parameters in a probabilistic sense have taken this approach (e.g. Andronova & Schlesinger 2001; Wigley & Raper 2001; Forest *et al.* 2002, 2006; Knutti *et al.* 2002, 2003, 2005, 2006; Murphy *et al.* 2004; Annan *et al.* 2005*b*; Frame *et al.* 2005; Meinshausen 2005; Piani *et al.* 2005; Stainforth *et al.* 2005; Hegerl *et al.* 2006; Schneider von Deimling *et al.* 2006). As it is also the case for multi-model ensembles in ways that we highlight in §3, the PPE approach is limited in its ability to capture the full range of uncertainties in the models' representation of the true climate system, as there are many ways to design a parametrization. While many processes behind the parametrization are well understood, or observational or theoretical arguments exist, there are cases where the evidence is rather circumstantial and values are often chosen simply because they seem to work. In fact, an alternative approach to the interpretation of parameter uncertainty is being developed at the European Center for Medium-Range Weather Forecast where the idea of deterministic, if unknown, parameters has been put aside and experiments with stochastic parametrization are currently being run and evaluated (Palmer *et al.* 2005*a*).

In general, because the true climate system is highly complex, it remains fundamentally impossible to describe all its processes in a climate model, no matter how complex the model itself is. So, choices have to be made on what processes to include, how to parametrize them and what pieces to neglect. On the one hand, what is relevant to include in a model should depend on the question of interest, and a wide spectrum of models does exist, each suitable for specific applications. But even for a given problem, the process of selecting pieces to include in a model is at least partly subjective, based on expert knowledge and experience. Any uncertainty that is introduced by choices in the model design, i.e. going beyond changing values of particular parameters, is usually referred to as structural uncertainty and would be hard to capture by changing parameters within a single model, no matter how wide the range of parameters is chosen. Similar arguments can be made for the choice of the type of grid, the resolution, truncation or the type of numerical methods used to solve the equations (e.g. whether spectral or finite volume methods are used in an atmospheric model). These numerical aspects are also part of the model structure, but sometimes considered separately from the physical aspects.

There is certainly tremendous value in exploring parametric uncertainties by the PPE approach, and its success might be partly related to the simplicity of generating those ensembles. Apart from the enormous computational capacity required, this exploration of the parameter space is rather straightforward. These ensembles offer insight into processes if a single parameter is perturbed at the time. They have also been used successfully with multiple parameter perturbations to generate probability density functions (PDFs) of transient future warming, equilibrium climate sensitivity, the present-day magnitude of the aerosol forcing and various other projections and model parameters by constraining large ensembles with observations (see references above). However, for the reasons discussed above, PPE experiments address only part of the problem, even if an ever larger one, when the expensive choice is made to vary whole parametrization *schemes*. Structural uncertainties need to be further evaluated to understand and quantify the full uncertainty in climate change projections, and to make sure that a result found by using a particular model is not an artefact of its individual structure. Exploring this component of the uncertainty, i.e. the errors in approximating the true system that are most intrinsic to each model's fundamental formulations and cannot be addressed by varying its parametrizations, is the main motivation for looking at an ensemble of different models.

(*b*) *The motivation for multi-model ensembles*

We have argued that the quantification of all aspects of model uncertainty requires multi-model ensembles, ideally as a complement to the exploration of single-model uncertainties through PPE experiments. In addition, a variety of applications, not only limited to the weather and climate prediction problems, have demonstrated that combining models generally increases the skill, reliability and consistency of model forecasts. Examples include model forecasts in the sectors of public health (e.g. malaria; Thomson *et al.* 2006) and agriculture (e.g. crop yield; Cantelaube & Terres 2005), where the combined information of several models is reported to be superior to a single-model forecast.

Similarly, for weather- and climate-related applications, predictions for the El Niño Southern Oscillation (ENSO) and seasonal forecasts from multi-model ensembles are generally found to be better than single-model forecasts (e.g. Palmer *et al.* 2005*b*). Multi-model ensembles are defined in these studies as a set of model simulations from structurally different models, where one or more initial condition ensembles are available from each model (if more initial conditions for each model are available the experiments are often said to make up a super-ensemble). Seasonal forecasts show better skill, higher reliability and consistency when several independent models are combined (e.g. Doblas-Reyes *et al.* 2003; Yun *et al.* 2003). While for a single given diagnostic or variable, the multi-model performance might not be significantly better than the single best model, studies indicate that the improvements are more dramatic if an aggregated performance measure over many diagnostics is considered. Thus, the largest benefit is seen in 'the consistently better performance of the multi-model when considering all aspects of the predictions' (Hagedorn *et al.* 2005).

There are obviously different ways to combine models. In many cases, Bayesian methods (e.g. Robertson *et al.* 2004) or weighted averages, where weights are determined by using the historical relationship between forecasts and observations (e.g. Krishnamurti *et al.* 2000), perform better than simple averages where each model is weighted equally. Intuitively, it makes perfect sense to trust, and thus weigh, the better models more. The difficulty, however, is in quantifying model skill and deriving model weights accordingly. Controversial results exist regarding the best way to combine model results, even in the case where skill or performance can be calculated by comparing model predictions to observations. The problem of constructing a weighted average for climate projection, where no verification is available, is discussed in §3.

Improvements in the performance of a multi-model mean over single models were also found when detecting and attributing greenhouse gas warming and sulphate cooling patterns in the observed climate record (Gillett *et al.* 2002). An equally weighted average of several coupled climate models is usually found to agree better with observations than any single model (Lambert & Boer 2001).

Multi-model projections for long-term climate change were used in reports of the Intergovernmental Panel on Climate Change (IPCC), where unweighted multi-model means rather than individual model results were often presented as best guess projections (IPCC 2001). Probabilistic projections based on multi-model ensembles are rather new in the literature and are based on a variety of statistical methods. These methods are discussed in detail in §2. The field is rapidly evolving with the availability of larger ensembles of different models and with very large PPE (of the order of a 100 000 simulations; e.g. Stainforth *et al.* (2005)), and the development of new statistical approaches for their analysis.

In summary, simplifications, assumptions and choices of parametrizations have to be made when constructing a model, and they inevitably lead to errors in the model and the forecasts it produces. Improving forecasts and projections by combining models rests on the assumption that if those choices are made independently for each model, then the errors might at least partly cancel, resulting in a multi-model average that is more skilful than its constitutive terms. We will discuss the assumptions made in that line of argument, and specifically their justification in climate projections in §3.

2. Existing published methods

In 1990, the Atmospheric Model Intercomparison Project (AMIP; Gates 1992) developed a standard experimental protocol for atmospheric general circulation models (GCMs). For the first time, a systematic framework in support of model diagnosis, validation and intercomparison was put forward and since then the international community of climate modelling has participated and benefited from it widely. The natural follow-up to AMIP was CMIP, the Coupled Model Intercomparison Project (Meehl *et al.* 2000), whereby the output from coupled atmosphere–ocean general circulation models (AOGCMs) became the object of study. The first phase of CMIP was limited to control runs, while in a second phase (CMIP2) idealized scenarios of global warming with atmospheric CO_2 increasing at the rate of 1% per year were collected. More recently, in support of the activities leading to the IPCC fourth assessment report (AR4), the archive of coupled model output at the Program for Climate Model Diagnosis and Intercomparison (PCMDI, http://www-pcmdi.llnl.gov/) was extended to historical, SRES (Nakićenović *et al.* 2000) and commitment experiments, where concentrations of greenhouse gases are kept constant after reaching preset levels for the remaining length of a multi-century simulation. This ever-increasing availability of model experiments under common scenarios, whose output is standardized and to which access is facilitated, has naturally inspired the analysis of multi-model ensembles since the beginning of 2000. In the third assessment report of the IPCC (2001), many results were presented as multi-model ensemble averages, accompanied by measures of inter-model variability, most commonly inter-model standard deviations.

The article by Räisänen (1997) is probably the first one to explicitly advocate the need of a quantitative model comparison and the importance of inter-model agreement in assigning confidence to the forecasts of different models. But it was only in Räisänen & Palmer (2001) that a probabilistic view of climate change projections on the basis of multi-model experiments was first proposed. The models considered are 17 AOGCMs participating in CMIP2. Based on these models, probabilities of threshold events such as ‘the warming at the time of doubled CO_2 will be greater than 1°C’ are computed as the fraction of models that simulated such an event. These probabilities are computed at the grid point level and also averaged over the entire model grid to obtain global mean probabilities. The authors use cross-validation to test the skill of the forecast derived for many different events and forecast periods. They find better skill for temperature-related events than for precipitation-related events, and for events defined close to the time of CO_2 doubling than for events forecasted on shorter time scales, when the signal of change is weaker. They also show how probabilistic information may be used in a decision theory framework, where a cost–benefit analysis of initiating some action may have different outcomes depending on the probability distribution of the uncertain event from which protection is sought. Naturally, in their final discussion, the authors highlight the importance of adopting a probabilistic framework in climate change projections, and wish for it to become an ‘established part of the analysis of ensemble integrations of future climate’. Easier said than done. The same authors shortly thereafter applied their procedure to forecasts of extreme events (Palmer & Räisänen 2002), but the next significant step in the direction of probabilistic projections was published more than a year and a half later (Giorgi & Mearns 2002, 2003).

Räisänen & Palmer (2001) assigned one vote to each AOGCM, when counting frequencies of exceedance. The reliability ensemble average (REA) approach by Giorgi & Mearns (2002) assumes a different perspective: not all GCMs are created equal. Model performance in replicating current climate and inter-model agreement in the projections of future change should be of guidance in our synthesis of multi-model projections: models with small bias and projections that agree with the ensemble 'consensus' should be rewarded while models that perform poorly in replicating observed climate and that appear as outliers should be discounted. The REA method proposes an algorithmic estimation of model weights through which 'bias' and 'convergence' criteria are for the first time quantified. Defining the weights as

$$R_i = [(R_{B,i})^m \times (R_{D,i})^n]^{[1/(m\times n)]} = \left\{ \left[\frac{\epsilon_T}{|B_{T,i}|}\right]^m \times \left[\frac{\epsilon_T}{|D_{T,i}|}\right]^n \right\}^{[1/(m\times n)]}, \tag{2.1}$$

the weighted ensemble average is computed for separate subcontinental regions as

$$\widetilde{\Delta T} = \frac{\sum_i R_i \Delta T_i}{\sum_i R_i}, \tag{2.2}$$

where the individual model projections of change are indicated by ΔT_i. The weight for an individual model, R_i in equation (2.1), is defined as the product of two terms ($R_{B,i}$ and $R_{D,i}$), one inversely proportional to the absolute bias, $B_{T,i}$, and the other to the absolute distance between the model projected change and the final weighted ensemble average, $D_{T,i}$. At the numerator, ϵ_T, a measure of natural variability, ensures that models whose bias and deviation are not large relative to natural fluctuations would not be unjustly discounted. The exponents m and n can modulate the relative importance of the two terms in the weighted average, but are set equal to 1. In a note following this article, Nychka & Tebaldi (2003) showed that the REA estimate is in fact equivalent to a standard statistical methodology for estimating a population's central tendency in the presence of outliers. It is well known that simple averages are sensitive to 'extreme' observations, while median values provide a more robust estimate. It can be demonstrated that the final estimate $\widetilde{\Delta T}$ obtained as a weighted average through the iterative reevaluation of the REA weights in equation (2.1) is in fact the median of the sample of model projections, weighted by the part of equation (2.1) that depends only on model bias.

In the second paper (Giorgi & Mearns 2003), the same REA weights are used in the computation of frequencies of threshold exceedances as in Räisänen & Palmer (2001) and Palmer & Räisänen (2002) to derive probabilistic projections of various events (e.g. warming in a region exceeding 4°C or precipitation change exceeding 20% of current average). Thus, differential weighting of GCMs is applied for the first time in a probabilistic setting. Beside the innovative step of considering the two criteria in the formal combination of the ensemble projections, the regional nature of the analysis positioned these papers to be more directly relevant for impact studies and decision-making applications.

The REA approach motivated the work by Tebaldi *et al.* (2004, 2005) and Smith *et al.* (submitted). Their Bayesian analysis treats the unknown quantities of interest (current and future climate *signals*, model reliabilities) as random variables, for which reference prior distributions (Berger 1993) are chosen. Assumptions on the statistical distribution of the data (consisting of models' output and observations) as a function of the unknown parameters determine the likelihood, which is combined through Bayes theorem with the prior distributions to derive posterior distributions of all the uncertain quantities of interest, among which is the *climate change signal*. Simple Gaussian (normal) assumptions are stipulated for the current (X_i's) and future (Y_i's) model projections, centred around the true climate signals, μ and ν, respectively, with model-specific variances. The choice of 'reference' priors ensures that the data have maximum relevance in shaping the posterior distributions. It is in this strictly mathematical sense that the nature of the prior can be called 'uninformative', since, as has widely been discussed and acknowledged, there exists no choice of prior that can be defended as absolutely neutral. Thus, it is hypothesized that

$$\begin{aligned} X_i &\sim N(\mu, \lambda_i^{-1}), \\ Y_i &\sim N(\nu, (\theta\lambda_i)^{-1}), \end{aligned} \tag{2.3}$$

where the notation $N(\mu, \lambda^{-1})$ stands for a Gaussian distribution with mean μ and variance $1/\lambda$. Similarly, the observed current climate, X_0, is modelled as a realization from a Gaussian distribution centred around the same current climate signal μ, and whose variance is estimated through the observed record

$$X_0 \sim N(\mu, \lambda_0^{-1}). \tag{2.4}$$

Through Bayes theorem, evaluated numerically by Markov Chain Monte Carlo methods, a posterior distribution for the true climate signals is derived, and straightforwardly translated into a probability distribution for climate change, defined as ν–μ. As a consequence of the distributional assumptions, the criteria of bias and convergence, in an analytical form similar to the form of the REA weights, shape the posterior distributions. In fact, the form of the posterior means for μ and ν is approximately

$$\tilde{\mu} = \frac{(\sum_i \lambda_i X_i)}{(\sum_i \lambda_i)}, \tag{2.5}$$

and

$$\tilde{\nu} = \left(\frac{\sum_i \lambda_i Y_i}{\sum_i \lambda_i}\right), \tag{2.6}$$

where the model-specific λ_i's look very much like the REA weights, being estimated approximately as

$$\tilde{\lambda}_i = \frac{a+1}{b + \frac{1}{2}[(X_i - \tilde{\mu})^2 + \theta(Y_i - \tilde{\nu})^2]}. \tag{2.7}$$

The first term of the denominator in equation (2.7) is a measure of bias, being the distance of the present climate average as simulated by model i from the optimal estimate, $\tilde{\mu}$, of current climate. The second term is a measure of convergence, computing a distance between the model's future projection from the future climate signal's posterior mean ($\tilde{\nu}$). The terms a and b are prior parameters chosen as orders of magnitude smaller with respect to the remaining terms, so that they do not have significant impact on the final estimates. As in Giorgi & Mearns (2002), outliers receive less weight here as well. Sharp criticisms have been raised against the validity of the convergence criterion when analysing a set of models that are by design 'best guesses' rather than attempting to sample a wide range of uncertainties, and whose agreement may be a consequence of inbreeding rather than reciprocal validation of individual tendencies. In particular, it has been often argued that there may exist common weaknesses in the representation of certain processes in a majority of models, and consequently outliers may not appear at random. In response to these concerns, the authors proposed a variant of the analysis in which the outliers are not heavily penalized (Tebaldi *et al.* 2004). This is achieved by *a priori* assigning a large probability to the models being less 'precise' in their future projections compared with their skill in current projections. In the statistical model, this formally translates into a prior distribution for the parameter θ in the second line of equation (2.3) that is concentrated on values less than 1. Note also that the weighting of the original REA approach (Giorgi & Mearns 2002, 2003) is undergoing revision to incorporate different measures of model performance besides bias, and to eliminate the convergence criterion following the same strain of criticisms (F. Giorgi 2006, personal communication).

Another related consequence of assuming independence among GCM projections (which is implicitly or explicitly the case for all the methods described so far) is that any statistical analysis will produce increasingly more precise estimates (e.g. narrower posterior distributions of climate change signals) as the number of models in the ensemble increases. In fact, the paper by Lopez *et al.* (2006) which compares the probabilistic estimates derived from Tebaldi *et al.* (2004, 2005) with those derived using optimal fingerprinting methodology, as in Allen *et al.* (2000), shows the results of a numerical simulation with an increasing number of GCMs. The width of the posterior distribution of temperature change (computed as a global average in this exercise) is shown to have a strong inverse relation to the number of GCMs. It is however arguable that models do not provide perfectly independent pieces of evidence, as we discuss more at length in §3. Thus, a statistical analysis that by construction relies on inter-model agreement and discounts outliers, concentrating the range of uncertainty around the larger cluster(s) of models, can be overly optimistic regarding the precision of its final estimates. Nonetheless, an implicit reliance on model agreement can be detected in many published analyses. For example, in the writing of the IPCC third assessment report (IPCC 2001), two models were discarded altogether because they produced extreme estimates of warming, i.e. showed very large climate sensitivity. Note also that despite the ever-recurrent comment about the need of accounting for model dependence, no formal approach at quantifying this dependence has been worked out yet. A distance in model space is definitely a difficult concept to formalize. A further series of developments in the statistical treatment of Tebaldi *et al.* (2004, 2005) is presented by Smith *et al.* (submitted). There, an additional level of statistical modelling is introduced in order to avoid unbalanced attribution of weight to the different AOGCMs. Also, a multivariate treatment of a set of regions over the globe is proposed instead of the separate estimation of the signals of climate change in specific regions one by one. Through this approach, the reliability of each model is evaluated across a wider set of performances (across a large set of regions rather than over a single region), thus reducing the chance of concentrating the weight in the final estimates over a restricted set of AOGCMs. A cross-validation step is added to the analysis as a way of verifying the statistical modelling assumptions and results.

The regional nature of the analyses by Giorgi & Mearns (2002, 2003), Tebaldi *et al.* (2004, 2005) and Smith *et al.* (submitted) was adopted by Greene *et al.* (2006). These authors chose to combine multi-model ensembles by a method similar to what is employed for seasonal and interannual forecasting. A Bayesian hierarchical linear model is fitted to an observational dataset of regionally aggregated seasonal and annual temperatures, where the predictors are similarly aggregated GCM projections. The true observed temperature trends (Y_{ik}s) for region i and time k are modelled as centred around a mean value, with a Gaussian error

$$Y_{ik} \sim N(\mu_{ik}, \sigma_k^2), \tag{2.8}$$

and the mean value is modelled as a linear combination of the GCMs' output

$$\mu_{ik} = \beta_{0k} + \sum_j \beta_{jk} X_{ijk}, \tag{2.9}$$

where X_{ijk} indicates the simulated temperature in region i at time k by model j. This is similar to performing model calibration, but the method adds several relevant features like error terms that are regionally differentiated and linear coefficients that are derived from a parent multivariate normal distribution (common to all regions) where the variance–covariance structure across models accounts for inter-model correlation. The calibrated ensemble is used to derive climate change projections, and given the random nature of the coefficients (in a Bayesian setting), their posterior distribution translates into a probability distribution for the climate change projection. The main assumption governing this approach is that of stationarity of the relation between observed and simulated trends, estimated in the training period of the twentieth century and applied to future simulations. In fact, this strong assumption causes obvious differences between the simple average projections from the GCMs and the projections synthesized from the calibrated ensemble, in many cases resulting in distributions over a range of values significantly shifted, more often towards lower values. The possible uncertainty in the forcings applied to the twentieth century simulations and not accounted for by the method may also be contributing to the final estimates of the calibration coefficients.

Furrer *et al.* (in press) tackled the modelling of GCM output at the grid point scale, rather than at the aggregated level of large subcontinental regions. The central idea of the approach is to model each GCM field of temperature or precipitation change as a random process on the sphere. The field is made of two additive components: a large-scale climate signal and a small-scale error field, representing both model bias and internal variability. Thus, modelling the field of change, denoted as D_i for the ith GCM, and defined simply as the difference, grid point by grid point, of the future mean projection minus the current mean projections

$$D_i = Y_i - X_i, \tag{2.10}$$

the statistical model states that

$$D_i = M\theta_i + \epsilon_i. \tag{2.11}$$

The large-scale signal, represented as the first additive term of equation (2.11), is modelled as a linear combination of a set of truncated basis functions, filling the columns of the matrix M. The basis functions are spherical harmonics, apt to represent spatial structure on a sphere, plus a set of additional vectors modelling the expected geographical patterns like, for example, a land/ocean mask. Observations are also used as one of the additional columns in the linear combination, in the hope that they will help explain some of the effect of the physical processes that create climate on Earth but are not easily represented through statistical modelling. There is no direct use of either a bias or convergence criterion in the spatial modelling of this study. The coefficients of the linear combination are the components of the vector θ_i. The small-scale residual field ϵ_i is modelled as a realization of a stationary Gaussian random field of mean zero. Both the linear coefficients θ_i and the scale parameters of the covariance function in the Gaussian process ϵ_i are model-specific, a way to account for the different GCMs' characteristics in replicating the true climate signal. The vectors θ_i of linear coefficients are samples from a distribution whose mean are the 'true' coefficients. The goal of the Bayesian analysis is to estimate the posterior distribution of the true coefficients. Once recombined with the basis functions, the posterior distribution for the true coefficients will translate into a multidimensional probability distribution of the large-scale signal of climate change, jointly quantifying the uncertainty over the global grid. Complex regional statements of climate change, then, may be easily derived. This remains the only published method to represent the uncertainty over a global map, using spatial statistics to model geographical patterns of varying degrees of smoothness (e.g. temperature change fields rather than precipitation change fields) as a function of the spatial correlation between locations.

Based on the latest set of model experiments contributing to the AR4 of the IPCC, PDFs produced by the methods in Tebaldi *et al.* (2004), Greene *et al.* (2006) and Furrer *et al.* (in press) are compared in figure 1. The PDFs represent projections of temperature change in winter (DJF) and summer (JJA) for the end of the twenty-first century under the SRES A1B scenario. Also shown is the empirical distribution of the GCM projections, in the form of a shaded histogram. We choose four representative regions from the 22 regions first adopted by Giorgi & Francisco (2000), which have become a standard choice for regional climate change analysis on the basis of GCMs since then. Specifically, we choose Western North America (WNA), the Mediterranean basin (MED), Northern Asia (NAS) and Southeast Asia (SEA). Note that the empirical distribution is what determines the results of the method by Räisänen & Palmer (2001) and Palmer & Räisänen (2002). It is immediately noticeable how the different methods produce different curves. Tebaldi *et al.* (2004) and Furrer *et al.* (in press) produce similar curves, particularly with regard to the location of their central mass, but also for most cases similar in width. They are on average narrower than the empirical distribution of the GCM projections, as one would expect as a consequence of the scaling of the uncertainty range with the number of data points (20 in this example, as many as there are GCMs contributing to this particular experiment). The method by Greene *et al.* (2006) produces wider PDFs, probably as a consequence of the large degree of uncertainty in the estimation of the calibration coefficients. Also, some of the regions show the tendency of the latter method to produce curves whose mass is shifted with respect to the empirical distribution. This effect is due to the calibration coefficients being significantly different from those of a simple average, a consequence of the need of significantly 'reshaping' the modelled trends to match more closely with the observed ones. Tables 1 and 2 compare the probability of exceedance of several thresholds of temperature for the same regions, seasons and scenario as figure 1. These cumulative probabilities are derived from the empirical distributions of the GCMs projections and from the REA method of Giorgi & Mearns (2003). When they differ, the distributions derived by the REA method tend to concentrate the mass of the probability among a few GCM projections within the entire ensemble, thus producing steeper cumulative distributions.

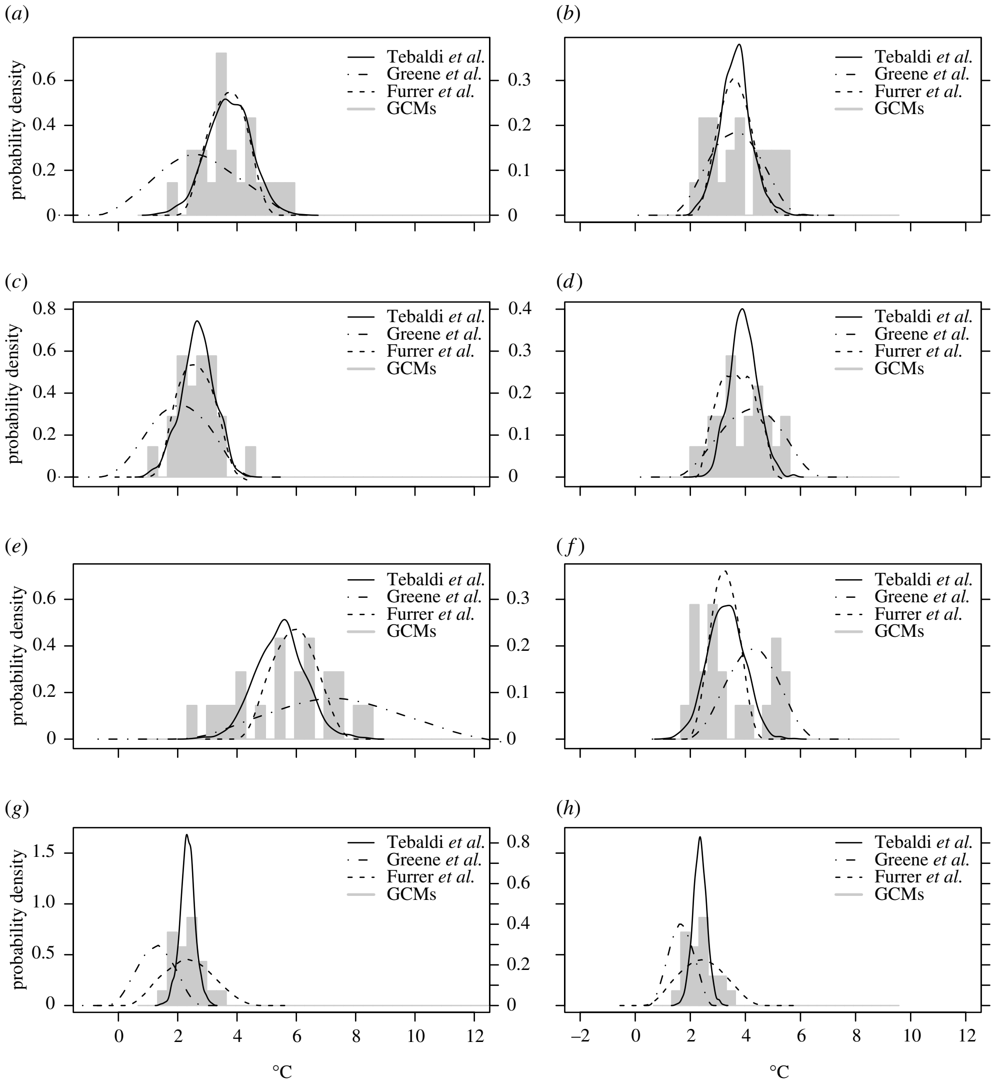

Figure 1. Comparison of PDFs derived by the three methods of Tebaldi *et al.* (2004), Greene *et al.* (2006) and Furrer *et al.* (in press) for four regions (Western North America, MEDiterranean basin, Northern ASia and Southeast Asia), two seasons (DJF and JJA) and one scenario (SRES A1B). Also represented is the histogram of the GCM projections. The temperature changes are computed as the difference between two 20-year averages, 2080–2099 versus 1980–1999. (*a*) WNA: A1B, DJF; (*b*) WNA: A1B, JJA; (*c*) MED: A1B, DJF; (*d*) MED: A1B, JJA; (*e*) NAS: A1B, DJF; (*f*) NAS: A1B, JJA; (*g*) SEA: A1B, DJF; (*h*) SEA: A1B, JJA.

A number of other papers have approached the problem of deriving regional probabilistic projections of climate change on the basis of multi-model ensembles. They are characterized by a less general approach, compared with the methods described so far, having been tailored to specific regions and impact studies. We choose three of them as good representatives of more focused studies. Benestad (2004) used statistical downscaling applied to a multi-model ensemble in order to derive probabilistic scenarios at a finer resolution for northern Europe.

Table 1. Probabilities of exceeding a series of increasing thresholds, expressed in °C, in the temperature change projections of an ensemble of GCMs, unweighted (labelled as 'empirical' in the rows of the table) and weighted by the REA method of Giorgi & Mearns (2003). (Results for temperature change in boreal winter (DJF) under SRES A1B. The temperature changes are computed as the difference between two 20-year averages, 2080–2099 versus 1980–1999.)

	$\Delta T > 0°\mathrm{C}$	$\Delta T > 1°\mathrm{C}$	$\Delta T > 2°\mathrm{C}$	$\Delta T > 3°\mathrm{C}$	$\Delta T > 4°\mathrm{C}$	$\Delta T > 5°\mathrm{C}$
WNA empirical	1.00	1.00	0.95	0.76	0.38	0.14
WNA REA	1.00	1.00	1.00	0.95	0.11	0.01
MED empirical	1.00	1.00	0.86	0.33	0.05	0.00
MED REA	1.00	1.00	1.00	0.01	0.00	0.00
NAS empirical	1.00	1.00	1.00	0.95	0.81	0.67
NAS REA	1.00	1.00	1.00	1.00	1.00	1.00
SEA empirical	1.00	1.00	0.71	0.10	0.00	0.00
SEA REA	1.00	1.00	0.81	0.06	0.00	0.00

Table 2. Probabilities of exceeding a series of increasing thresholds, expressed in °C, in the temperature change projections of an ensemble of GCMs, unweighted (labelled as 'empirical' in the rows of the table) and weighted by the REA method of Giorgi & Mearns (2003). (Results for temperature change in boreal summer (JJA) under SRES A1B. The temperature changes are computed as the difference between two 20-year averages, 2080–2099 versus 1980–1999.)

	$\Delta T > 0°\mathrm{C}$	$\Delta T > 1°\mathrm{C}$	$\Delta T > 2°\mathrm{C}$	$\Delta T > 3°\mathrm{C}$	$\Delta T > 4°\mathrm{C}$	$\Delta T > 5°\mathrm{C}$
WNA empirical	1.00	1.00	1.00	0.67	0.38	0.19
WNA REA	1.00	1.00	1.00	1.00	1.00	0.01
MED empirical	1.00	1.00	1.00	0.81	0.48	0.14
MED REA	1.00	1.00	1.00	1.00	1.00	0.00
NAS empirical	1.00	1.00	0.95	0.43	0.33	0.24
NAS REA	1.00	1.00	0.98	0.33	0.19	0.15
SEA empirical	1.00	1.00	0.71	0.14	0.00	0.00
SEA REA	1.00	1.00	1.00	0.00	0.00	0.00

The approach is the same as in Räisänen & Palmer (2001) and Palmer & Räisänen (2002) but is applied to trends downscaled from GCM output by a simple delta method, where regional anomalies simulated by GCMs are applied to observed local trends at a fine network of stations. Luo *et al.* (2005) were interested in the impacts of climate change over the wheat production of a small region of South Australia. Using output from GCMs and regional climate models (RCMs) run under different scenarios, they derive a regression relation between local change in temperature and precipitation and global average warming on a monthly basis. In turn, a relation between global warming and CO_2 concentrations and climate sensitivity is also derived. By sampling in a Monte Carlo framework along three different dimensions of uncertainty (climate scenarios, climate sensitivity and local change projections, the latter exemplified by different GCM-specific patterns), histograms of temperature and precipitation change in the regions are constructed and the results fed to a weather generator and used as input in crop models. Dettinger (2005) proposes resampling a given ensemble of trajectories of temperature or precipitation change, in the specific case for a small region corresponding to a watershed in California. In this way, the sample size is augmented and more robust estimates of the PDF of change are derived directly from the new larger ensemble. The resampling method decomposes the original ensemble into principal components (after a standardization procedure to preserve mean and standard deviation of the ensemble as a whole) and resamples the PCA results to form new members that, when added to the ensemble, preserve its spectral characteristics.

3. Challenges

There are a number of issues that need to be considered when constructing and interpreting multi-model climate projections, whether in the form of probability distributions of climate change or simple averages and measures of variability across models. Most of these difficulties, discussed in the following sections, are recognized by the climate modelling community, but are still poorly understood and quantified. Therefore, references and suggestions on how to tackle, let alone resolve, these issues are inevitably sparse. We choose to group the problems into four categories, namely, metrics of model validation, model dependence, experimental design of multi-model experiments (or lack of) and model tuning, although the different issues are partly related.

(*a*) *Metrics, skill and the lack of verification*

The predictive skill of a model is usually measured by comparing the predicted outcome with the observed one. Note that any forecast produced in the form of a confidence interval, or as a probability distribution, cannot be verified or disproved by a single observation or realization since there is always a non-zero probability for a single realization to be within or outside the forecast range just by chance. Skill and reliability are assessed by repeatedly comparing many independent realizations of the true system with the model predictions through some metric that quantifies agreement between model forecasts and observations (e.g. rank histograms). For projections of future climate change over decades and longer, there is no verification period, and in a strict sense there will never be any, even if we wait for a century. The reason is that the emission scenario assumed as a boundary condition is very likely not followed in detail, so the observations from the single climate realizations will never be fully compatible with the boundary conditions and scenario assumptions made by the models. And even if the scenario were to be followed, waiting decades for a single verification dataset is clearly not an effective verification strategy. This might sound obvious, but it is important to note that climate projections, decades or longer in the future by definition, cannot be validated directly through observed changes. Our confidence in climate models must therefore come from other sources.

The judgement of whether a climate model is skilful or not does not come from its prediction of the future, but from its ability to replicate the mean climatic conditions, climate variability and transient changes for which we have observations, and from its ability to simulate well-understood climate processes. For example, climate models are evaluated on how well they simulate the present-day mean climate (e.g. atmospheric temperature, precipitation, pressure, vertical profiles, ocean temperature and salinity, ocean circulation, sea ice distributions, vegetation, etc.), the seasonal cycle and climate variability on various time scales (e.g. the North Atlantic oscillation, ENSO, etc.). Their response to specified forcing is compared to the observed warming over the industrial period. They are evaluated against proxy data from past climate states, e.g. the Last Glacial Maximum, the Mid-Holocene, the last Interglacial period, or even further back in time (e.g. Kiehl & Shields 2005; Otto-Bliesner *et al.* 2006*a*,*b*). Further, individual processes in the model are studied in detail and evaluated against both theory and observations. While all those activities have helped in improving the models, and have greatly increased our confidence that the models capture the most relevant processes, simulating the past and present correctly does not guarantee that the models will be correct in the future. In other words, while there is a lot of circumstantial evidence for the models to be trusted for the future, there is no definitive proof for model skill in projecting future climate. In fact, most models agree reasonably well with observations of the present-day mean climate and simulate a realistic warming over the twentieth century (of course, the specific performance depends on each model/metric combination), yet their predictions diverge substantially for the twenty-first century, even when forced with the same boundary conditions.

The difficulty in quantifying model performance for future climate can be circumvented using two approaches. The first approach may be to ignore model performance altogether. As criticizable as this should be, it is in fact the case that multi-model simple averages are widely used, for example in the IPCC (2001) report. While this is likely to improve the best guess projections because model errors tend to cancel, it hardly makes optimal use of the information available. Intuitively, and using an extreme example, a model that performs well in every case where it has been compared to observations is more likely to be correct in the future than the one that is inconsistent with even the basic observed features of the climate. In an alternative approach, therefore, models can be combined in a weighted average (or subsets of models can be used, which means giving zero weight to some), where the model weight is determined by some measure of model performance. The crux lies in defining a metric for model performance which might be relevant for predicting future climate, but must be based on observations from the past or present. There is no unique way of doing that.

Tebaldi *et al.* (2005), for example, following the approach set forward by Giorgi & Mearns (2002), based the model weights on how well the model simulates the climatological mean temperature in the region of interest. Alternatively, one could argue that since the model prediction is a trend (warming over a certain time period), one should evaluate the models by how well they simulate the observed warming trend over the last decades, or on both the mean and trend, as done by Greene *et al.* (2006). Spatial patterns may be considered, thus favouring an approach similar to Furrer *et al.* (in press) which considers the full global fields instead of isolated regions, although that particular approach does not explicitly weight models. When precipitation is the focus, one should probably also look at temperature and dynamics of the atmosphere, since they both affect precipitation. Murphy *et al.* (2004) defined a climate prediction index that includes a large number of fields against which the model versions are compared. However, it is unclear which fields are important for a model to give a credible climate change response, it is unclear how the different diagnostics should be weighted, and it is probable that some of the diagnostics are dependent on each other, and thus redundant to some degree. No single climate model is best with respect to all variables (IPCC 2001; Lambert & Boer 2001), thus the weight given to each model in a probabilistic projection will always depend on the metric used to define model performance. For a given metric and for present-day climate, weighted averages of models were shown to compare better to observations than to raw averages with equal weights (Min & Hense 2006). It is unlikely, however, that the weights for future projections should be the same as those found to be optimal for present-day climate. The choice of a metric to weight models for future projections is therefore pragmatic, subjective and often also influenced by what can be observed with sufficient accuracy. It may depend on the prediction of interest, but even for a given prediction (e.g. future precipitation change in a certain region), there is no consensus of what the best metric is to quantify model performance. Arguably, then, the best approaches should attempt to use multiple diagnostics and metrics of performance, combining them in ways that account for possible correlations within the set, for uncertainty in observations and for inherent limitations in the model's ability of representing the quantities of interest. A novel approach that addresses many of these issues has recently been proposed by Goldstein & Rougier (2004) and Rougier (in press).

Whatever the approach taken in building multi-model probabilistic projections, the same issues of verifiability apply to their results as to the individual models. Some of the statistical analyses presented in §2 have adopted a 'perfect model' approach to verification, i.e. a cross-validation approach in statistical speech, to verify—at a minimum—that each model projection is itself compatible with the final synthesis product, when the latter is built on the basis of the remaining models only.

(*b*) *Model dependence and mean bias*

The idea that the performance of a forecast can be improved by averaging or combining results from multiple models is based on the fundamental assumption that errors tend to cancel if the models are independent, and thus uncertainty should decrease as the number of models increases. Most studies explicitly or implicitly assume this model independence in their statistical analyses, while many others intuitively make use of it in that they are more confident about a result that is common to many models than about results only seen in a single model.

Indeed, models are developed relatively independently by different groups around the world. However, models are also similar in many respects. All models have similar resolution, and thus can or cannot resolve the same processes. It is probable that the errors introduced into the models by the parametrization or the unresolved processes are similar across models. Models often use the same theoretical arguments for their parametrizations. For example, isopycnal diffusion (diffusion along surfaces of constant densities; Redi 1982) and the Gent/McWilliams eddy mixing parametrization (an advective term mimicking the mixing effect of ocean eddies; Gent & McWilliams 1990; Gent *et al.* 1995) are used in many ocean models. Also, the observations to tune the parametrizations or to evaluate models are often the same for most models. Any deficiency in the structure of the parametrization or biases in the observations used to constrain the free parameters will be persistent across many models. Furthermore, models share grids and numerical methods to solve the equations, and each method is known to have some deficiencies. In some cases, in particular when successful models are open to the community, entire model components are borrowed by other models to reduce the effort of developing an alternative module.

For the most recent coordinated modelling effort archived at PCMDI, several groups submitted more than one model or model version, e.g. one model was run at two different resolutions but the same physics; one ocean was coupled to two different atmospheres. In those cases, the models are clearly not independent, and their biases against observations are probably highly correlated. Sharing components and knowledge is not bad *a priori*, but it will result in persistent biases in a multi-model mean, whether weighted or not.

If the models were independent, like independent realizations of a random process, the uncertainty around the estimate of the process mean should approach zero as increasingly more models are averaged. Although model performance does improve when combining models, this behaviour does not seem to be entirely defensible, at least not for very large number of models. Some errors in a model might be random, but others are the result of our limited understanding of the processes and our ability to parametrize them efficiently in coarse resolution models. Neither of those is improved by adding more and more models of the same quality. There are known problems common to many models. For example, most models tend to overestimate short-wave and underestimate long-wave surface fluxes (Wild 2005; Wild *et al.* 2006), and it is unclear whether those biases affect the projections of future climate in a persistent way. The random errors will tend to cancel, while the ones common to many models will not. There might also be ‘unknown unknowns’, i.e. misrepresentations of processes, missing processes or uncertainties in processes that we are not aware of.

In sum, the current generation of models cannot be considered to be fully independent, nor are the models distributed around the true representation of the climate system. Therefore, in the absence of new knowledge about the processes and a substantial increase in computational resources to correctly resolve or parametrize them, our confidence in models should not increase unboundedly, and thus our uncertainty should not continue to decrease when the number of models increases.

(*c*) *The ensemble of opportunity*

Multi-model datasets are often described as ‘ensembles of opportunity’. This refers to how they are assembled, namely by asking for model results from anyone who is willing to contribute. Various non-scientific aspects determine the size and composition of such ensembles, e.g. whether modelling groups are interested at all and whether they have funding and computational resources to do the requested simulations.

The implication of how these multi-model ensembles are created is that the sampling is neither systematic nor random. The distribution of the models is thus completely arbitrary, and might be different in a subsequent ensemble, therefore changing the result even if the knowledge about the climate system has not changed. Even if we were to solve the problem of weighting the individual members, the posterior would still depend on the prior distribution which is determined by human decisions and cannot be interpreted in a scientific sense.

Another aspect of the same problem is that the models are not designed to span the full range of behaviour or uncertainty that is known to exist. This is not surprising, since all models are tuned and improved to match the observations as closely as possible. Once scientists are satisfied with a model, they rarely go back and see whether there might be another set of model parameters that gives a similar fit to observations but shows a different projection for the future. In other words, the process of building and improving a model is usually a process of convergence, where subsequent versions of a model build on previous versions, and parameters are only changed if there is an obvious need to do so. With a few exceptions of large PPE (Murphy *et al.* 2004; Stainforth *et al.* 2005; Collins *et al.* 2006) based on AOGCMs and simpler models (e.g. Wigley & Raper 2001; Forest *et al.* 2002; Knutti *et al.* 2002), the range of parameters and possible responses is not normally explored in GCMs by varying parameters simultaneously, mostly due to computational limitations.

Maybe the most prominent example is climate sensitivity, the equilibrium global mean surface temperature increase for a doubling of the atmospheric carbon dioxide concentration. Climate sensitivity in GCMs is usually not tuned, but the result of a sum of mostly atmospheric feedback processes. The range of climate sensitivities covered by the GCMs is approximately 2.0–4.5°C. A number of recent studies have attempted to quantify the range of climate sensitivity based on simpler models and perturbed physics GCM ensembles, and using observations of the radiative balance, observed warming of the surface and ocean, radiative forcing and of proxy evidence or the last millennium and Last Glacial Maximum. Most of the results indicate a substantial probability that climate sensitivity might be higher than 4.5°C, maybe up to 6°C or more (Andronova & Schlesinger 2001; Forest *et al.* 2002, 2006; Knutti *et al.* 2002; Murphy *et al.* 2004; Frame *et al.* 2005; Piani *et al.* 2005; Stainforth *et al.* 2005; Hegerl *et al.* 2006). Yet no GCM in the multi-model ensembles comes even close to sampling such high values of climate sensitivity (although a small fraction of the models in PPE experiments do sample those high values). Therefore, it is probable that the range covered by the multi-model ensemble covers a minimum rather than the full range of uncertainty.

(*d*) *Model tuning and evaluation*

Models can be evaluated using a variety of observations, from station data, satellite data, proxy data to reanalysis data, and more. Does a model that agrees well with observations more likely capture the important processes of the system it attempts to describe? This is certainly true to a large degree. However, in some cases, model agreement with observations can be improved by changing parameters that are unrelated to the problem. For example, if a model overestimates the temperature in a certain region, this could be improved by slightly changing the albedo of the dominant plant type in that region, although the problem might actually be related to an incorrect atmospheric circulation pattern. There is a danger of getting the right result for the wrong reason by tuning the wrong end of the model. Another example is aerosol forcing, where the spread of the total aerosol forcing across models is relatively small compared with the spread of the individual components of the forcing, indicating that different models get a similar response but for different reasons. Similarly, warming over the twentieth century is consistent with observations in many models. But it depends on the transient ocean heat uptake, climate sensitivity and the radiative forcing combined, and any bias in one of those quantities can be compensated by changes in the others (Knutti *et al.* 2002). Therefore, agreement with observations can be spurious, and can arise from a cancelling of errors, not necessarily guaranteeing that processes are correctly simulated.

The problem gets worse when the datasets which are used to tune the model are identical to those used later to evaluate the performance of the model and derive model weights in multi-model averages. Model evaluation should ideally be performed on independent datasets. Otherwise, the possibility of circular reasoning arises: a model can agree well with observations simply because the very same observations have been used to derive or tune the model.

Model tuning is often a subjective process. Annan *et al.* (2005*a*,*b*) have shown that the ensemble Kalman filter (Evensen 1993, 1994) can be a very efficient alternative to create a perturbed set of model versions consistent with observations, and that such an approach could be more objective. However, they note the problem of imperfectly known model error, i.e. the fact that the model is to some degree inadequate such that there is no perfect agreement with observations no matter how carefully the model is tuned. Also, it remains to be shown that an automated tuning approach can produce model solutions substantially better than those produced by experts making choice on the parameters based on their experience and understanding of the processes. Despite those difficulties, objective tuning methods, data assimilation and model evaluation on seasonal forecasts or in the context of weather prediction provide promising areas of research that have barely been explored with climate models so far.

Observations are also uncertain. Besides the fact that sparsity or poor quality of observations may be of obstacle in model tuning, or obfuscate model shortcomings, biased observations would cause all models to be biased in the same way, and any attempt of combing models will suffer from the same problem. While some datasets (e.g. station data) are independent of the models, the problem gets worse when reanalysis data are used. Certain types of reanalysis data are not constrained by observations and are purely model-derived, and the reanalysis models used for that are based on similar numerical methods, assumptions and parametrizations as the ones of the climate models, in which the datasets are used for evaluation afterwards (this problem is especially relevant for variables related to the hydrological cycle). Fortuitous agreement of climate models with reanalysis products might therefore result to some degree from common biases in both climate models and reanalysis models.

Finally, some models are known to ignore certain processes. For example, in the most recent collection of simulations of the twentieth century, several models do not include variations in the solar and volcanic forcing, and therefore should not reproduce the observed warming trends and patterns correctly. Nevertheless, they are usually evaluated against the observed trends.

In sum, although model agreement with observations is very valuable in improving the model, and is a necessary condition for a model to be trusted, it does not definitely prove that the model is right for the right reason. There are well-known examples where errors in different components of a single model tend to cancel. The use of the same datasets for tuning and model evaluation raises the question of circular reasoning.

4. Future directions and conclusions

Probabilistic climate projections from multi-model ensembles is a relatively recent topic in climate research, and it has gained a lot of momentum over the last few years. PDFs of projected changes attempt to represent the uncertainties that are embodied by a spectrum of modelling choices, and by the inherent imperfection of each and every one of them. These quantitative representations of the uncertainty are apt at being propagated into impact models (e.g. economic, crop or water resource management models) and can be used to study strategies for decision making under uncertainty. However, the list of open questions and issues associated with the interpretation of multi-model ensembles for climate projections as well as with statistical methodological aspects is still long, and compels us to look at the results of the analyses described in §2 as experimental and still lacking robustness, as evident by the significant disagreement among the PDFs produced by the different approaches.

In §3 we highlighted what we consider the main challenges to combining multi-model ensembles: the choice of metrics and diagnostics of performance, especially as they suggest model reliability for future projections; inter-model dependencies and common biases that should be quantified to avoid 'double counting' and over-optimistic reliance on consensus estimates; and representativeness of the sample of models with regard to common fundamental uncertainties. We think future analyses and better understanding of these ensembles will have to achieve fundamental progress in these areas.

Looking at the future availability of model output from concerted international efforts, we see new challenges appearing, when some climate models start to include new components of the Earth system that are not standard across the larger model population. Extensions beyond the traditional AOGCM framework of ocean, atmosphere, sea ice and land include, for example, biogeochemical cycles (e.g. carbon and nitrogen cycles), ecosystem models, atmospheric chemistry, extensions of the atmospheric model beyond the troposphere and stratosphere, urban models, embedding RCMs into global models, and more. While the latest multi-model ensemble created for the IPCC AR4 was relatively homogeneous in the sense that all models included just the four main components, it is unclear whether this will be the case in future coordinated model efforts. If future sets of models are less uniform and coherent in their structure and in the processes they include or neglect, their interpretation and combination will be more difficult. On the other hand, they will probably sample a wider range of structural uncertainties in that case, and will be reducing the concern about common biases.

For the decision-making process, it is important to know whether uncertainty in the evolution of future climate will remain at a similar level or whether it will be reduced substantially in the next decades. This uncertainty depends on the uncertainties in the emission scenarios, caused by uncertainties in social, economical and technical development, as well as uncertainties in climate model projections for a given scenario, caused by our incomplete understanding of the climate system and the ability to describe it in a reasonably efficient computational model. While a probabilistic picture of climate model uncertainty is evolving (as demonstrated by many papers in this issue), emission scenarios so far do not have likelihoods attached to them (Schneider 2001). For these reasons, it seems unlikely at this stage that projection uncertainty will decrease significantly in the very near future even on a global scale (Stott & Kettleborough 2002), and this is even more true for local projections. In principle, increased computational capacities combined with accurate long-term observations have the potential to substantially reduce the climate model uncertainties on the long run. However, even if the perfect climate model did exist, any projection is always conditional on the scenario considered. Given the often non-rational and unpredictable behaviour of humans, their decisions and the difficulty in describing human behaviour and economics in models, the perfect climate forecast (as opposed to a projection that is conditional on the scenario) is a goal that will probably be impossible due to the uncertainties in emission scenarios and the feedback loops involving the agents that the forecast is directed towards. Nonetheless, a comprehensive picture of the uncertainty in climate projections remains a key goal to aim for, and we should welcome the opportunity of taking advantage of independent resources and minds at work on it, by intelligently combining their—always different to some degree—results.

We thank A. M. Greene and R. Furrer for providing results from their published methods. We thank R. Furrer, D. Nychka, L. Mearns, T. Stocker, J. Meehl and R. Smith for discussions and feedback. We thank an anonymous reviewer and D. Sexton for their thoughtful comments and suggestions, which greatly improved this paper's final version. C.T. was supported by NCAR through the Weather and Climate Impacts Assessment Science Program. R.K. was supported by the Swiss National Science Foundation and NCAR. NCAR is supported by NSF.

References

Allen, M. R., Stott, P. A., Mitchell, J. F. B., Schnur, R. & Delworth, T. L. 2000 Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature* **407**, 617–620. (doi:10.1038/35036559)

Andronova, N. & Schlesinger, M. E. 2001 Objective estimation of the probability distribution for climate sensitivity. *J. Geophys. Res.* **106**, 22 605–22 612. (doi:10.1029/2000JD000259)

Annan, J. D., Hargreaves, J. C., Edwards, N. R. & Marsh, R. 2005*a* Parameter estimation in an intermediate complexity Earth system model using an ensemble Kalman filter. *Ocean Model.* **8**, 135–154. (doi:10.1016/j.ocemod.2003.12.004)

Annan, J. D., Hargreaves, J. C., Ohgaito, R., Abe-Ouchi, A. & Emori, S. 2005*b* Efficiently constraining climate sensitivity with paleoclimate simulations. *Sci. Online Lett. Atmos.* **1**, 181–184.

Benestad, R. 2004 Tentative probabilistic temperature scenarios for northern Europe. *Tellus A* **56**, 89–101. (doi:10.1111/j.1600-0870.2004.00039.x)

Berger, J. O. 1993 *Statistical decision theory and Bayesian analysis*, p. 617, 2nd edn. New York, NY: Springer.

Bryan, F. O., Danabasoglu, G., Nakashiki, N., Yoshida, Y., Kim, D.-H. & Tsutsui, J. 2006 Response of the North Atlantic thermohaline circulation and ventilation to increasing carbon dioxide in CCSM3. *J. Clim.* **19**, 2382–2397. (doi:10.1175/JCLI3757.1)

Cantelaube, P. & Terres, J.-M. 2005 Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A* **57**, 476–487. (doi:10.1111/j.1600-0870.2005.00125.x)

Collins, M., Booth, B. B. B., Harris, G., Murphy, J. M., Sexton, D. M. H. & Webb, M. J. 2006 Towards quantifying uncertainty in transient climate change. *Clim. Dynam.* **27**, 127–147. (doi:10.1007/s00382-006-0121-0)

Dettinger, M. 2005 From climate-change spaghetti to climate-change distributions for 21st century California. *San Francisco Estuary Watershed Sci.* **3**, article 4.

Doblas-Reyes, F. J., Pavan, V. & Stephenson, D. B. 2003 The skill of multimodel seasonal forecasts of the wintertime North Atlantic Oscillation. *Clim. Dynam.* **21**, 501–514. (doi:10.1007/s00382-003-0350-4)

Evensen, G. 1993 The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynam.* **53**, 343–367. (doi:10.1007/s10236-003-0036-9)

Evensen, G. 1994 Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**, 10 143–10 162. (doi:10.1029/94JC00572)

Forest, C. E., Stone, P. H., Sokolov, A. P., Allen, M. R. & Webster, M. D. 2002 Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science* **295**, 113–117. (doi:10.1126/science.1064419)

Forest, C. E., Stone, P. H. & Sokolov, A. P. 2006 Estimated PDFs of climate system properties including natural and anthropogenic forcings. *Geophys. Res. Lett.* **33**, L01705. (doi:10.1029/2005GL023977)

Frame, D. J., Booth, B. B. B., Kettleborough, J. A., Stainforth, D. A., Gregory, J. M., Collins, M. & Allen, M. R. 2005 Constraining climate forecasts: the role of prior assumptions. *Geophys. Res. Lett.* **32**, L09702. (doi:10.1029/2004GL022241)

Furrer, R., Sain, S., Nychka, D. & Meehl, G. In press. Multivariate Bayesian analysis of atmosphere–ocean general circulation models. *Environ. Ecol. Stat.*

Gates, W. L. 1992 AMIP: the atmospheric model intercomparison project. *Bull. Am. Meteorol. Soc.* **73**, 1962–1970. (doi:10.1175/1520-0477(1992)073<1962:ATAMIP>2.0.CO;2)

Gent, P. R. & McWilliams, J. C. 1990 Isopycnal mixing in ocean circulation models. *J. Phys. Oceanogr.* **20**, 150–155. (doi:10.1175/1520-0485(1990)020<0150:IMIOCM>2.0.CO;2)

Gent, P. R., Willebrand, J., McDougall, T. J. & McWilliams, J. C. 1995 Parameterizing eddy-induced tracer transports in ocean circulation models. *J. Phys. Oceanogr.* **25**, 463–474. (doi:10.1175/1520-0485(1995)025<0463:PEITTI>2.0.CO;2)

Gillett, N. P., Zwiers, F. W., Weaver, A. J., Hegerl, G. C., Allen, M. R. & Stott, P. A. 2002 Detecting anthropogenic influence with a multi-model ensemble. *Geophys. Res. Lett.* **29**, 1970. (doi:10.1029/2002GL015836)

Giorgi, F. & Francisco, R. 2000 Uncertainties in regional climate change predictions. A regional analysis of ensemble simulations with the HADCM2 GCM. *Clim. Dynam.* **16**, 169–182. (doi:10.1007/PL00013733)

Giorgi, F. & Mearns, L. 2002 Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the 'reliability ensemble averaging' (REA) method. *J. Clim.* **15**, 1141–1158. (doi:10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2)

Giorgi, F. & Mearns, L. 2003 Probability of regional climate change calculated using the reliability ensemble average (REA) method. *Geophys. Res. Lett.* **30**, 1629–1632. (doi:10.1029/2003GL017130)

Goldstein, M. & Rougier, J. 2004 Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM J. Sci. Comput.* **26**, 467–487. (doi:10.1137/S106482750342670X)

Greene, A., Goddard, L. & Lall, U. 2006 Probabilistic multimodel regional temperature change projections. *J. Clim.* **19**, 4326–4343. (doi:10.1175/JCLI3864.1)

Hagedorn, R., Doblas-Reyes, F. J. & Palmer, T. N. 2005 The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus A* **57**, 219–233. (doi:10.1111/j.1600-0870.2005.00103.x)

Hegerl, G. C., Crowley, T. J., Hyde, W. T. & Frame, D. J. 2006 Climate sensitivity constrained by temperature reconstructions over the past seven centuries. *Nature* **440**, 1029–1032. (doi:10.1038/nature04679)

IPCC 2001 In *Climate change 2001: the scientific basis. Contribution of working group I to the third assessment report of the Intergovernmental Panel on Climate Change* (eds J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, D. Xiaosu, X. Dai, K. Maskell & C. A. Johnson), p. 881. Cambridge, UK: Cambridge University Press.

Kiehl, J. T. & Shields, C. A. 2005 Climate simulation of the latest Permian: implications for mass extinction. *Geology* **33**, 757–760. (doi:10.1130/G21654.1)

Knutti, R., Stocker, T. F., 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)

Knutti, R., Stocker, T. F., Joos, F. & Plattner, G.-K. 2003 Probabilistic climate change projections using neural networks. *Clim. Dynam.* **21**, 257–272. (doi:10.1007/s00382-003-0345-1)

Knutti, R., Joos, F., Müller, S. A., Plattner, G.-K. & Stocker, T. F. 2005 Probabilistic climate change projections for CO_2 stabilization profiles. *Geophys. Res. Lett.* **32**, L20 707. (doi:10.1029/2005GL023294)

Knutti, R., Meehl, G. A., Allen, M. R. & Stainforth, D. A. 2006 Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Clim.* **19**, 4224–4233. (doi:10.1175/JCLI3865.1)

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., Larow, T., Bachiochi, D., Williford, E., Gadgil, S. & Surendran, S. 2000 Multimodel ensemble forecasts for weather and seasonal climate. *J. Clim.* **13**, 4196–4216. (doi:10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2)

Lambert, S. J. & Boer, G. J. 2001 CMIP1 evaluation and intercomparison of coupled climate models. *Clim. Dynam.* **17**, 83–106. (doi:10.1007/PL00013736)

Lopez, A., Tebaldi, C., New, M., Stainforth, D. A., Allen, M. R. & Kettleborough, J. 2006 Two approaches to quantifying uncertainty in global temperature changes. *J. Clim.* **19**, 4785–4796. (doi:10.1175/JCLI3895.1)

Luo, Q., Jones, R., Williams, M., Bryan, B. & Bellotti, W. 2005 Probabilistic distributions of regional climate change and their application in risk analysis of wheat production. *Clim. Res.* **29**, 41–52.

Meehl, G., Boer, G. J., Covey, C., Latif, M. & Stouffer, R. J. 2000 The Coupled Model Intercomparison Project (CMIP). *Bull. Am. Meteorol. Soc.* **81**, 313–318. (doi:10.1175/1520-0477(2000)081<0313:TCMIPC>2.3.CO;2)

Meinshausen, M. 2005 What does a 2°C target mean for greenhouse gas concentrations? A brief analysis based on multi-gas emission pathways and several climate sensitivity uncertainty estimates. In *Avoiding dangerous climate change* (eds H. J. Schellnhuber, W. Cramer, N. Nakicenovic, T. Wigley & G. Yohe), pp. 265–279. Cambridge, UK: Cambridge University Press.

Min, S.-K. & Hense, A. 2006 A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models. *Geophys. Res. Lett.* **33**, L08708. (doi:10.1029/2006GL025779)

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M. & Stainforth, D. A. 2004 Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* **429**, 768–772. (doi:10.1038/nature02771)

Nakićenović, N. *et al.* 2000 *Special report on emission scenarios: Intergovernmental Panel on Climate Change*, p. 599. Cambridge, UK: Cambridge University Press.

Nychka, D. & Tebaldi, C. 2003 Comments on calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the reliability ensemble averaging (REA) method. *J. Clim.* **16**, 883–884. (doi:10.1175/1520-0442(2003)016<0883:COCOAU>2.0.CO;2)

Otto-Bliesner, B. L., Brady, E., Clauzet, G., Tomas, R., Levis, S. & Kothavala, Z. 2006*a* Last Glacial Maximum and Holocene climate in CCSM3. *J. Clim.* **19**, 2526–2544. (doi:10.1175/JCLI3748.1)

Otto-Bliesner, B. L., Marshall, S. J., Overpeck, J. T., Miller, G. H., Hu, A. & CAPE Last Interglacial Project members. 2006*b* Simulating Arctic climate warmth and icefield retreat in the last interglaciation. *Science* **311**, 1751–1753. (doi:10.1126/science.1120808)

Palmer, T. 2005 Global warming in a nonlinear climate—can we be sure? *Europhys. News* **2**, 42–46.

Palmer, T. N. & Räisänen, J. 2002 Quantifying the risk of extreme seasonal precipitation events in a changing climate. *Nature* **415**, 512–514. (doi:10.1038/415512a)

Palmer, T., Shutts, G., Hagedorn, R., Doblas-Reyes, F., Jung, T. & Leutbecher, M. 2005*a* Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet. Sci.* **33**, 163–193. (doi:10.1146/annurev.earth.33.092203.122552)

Palmer, T. N., Doblas-Reyes, F. J., Hagedorn, R. & Weisheimer, A. 2005*b* Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Phil. Trans. R. Soc. B* **360**, 1991–1998. (doi:10.1098/rstb.2005.1750)

Piani, C., Frame, D. J., Stainforth, D. A. & Allen, M. R. 2005 Constraints on climate change from a multi-thousand member ensemble of simulations. *Geophys. Res. Lett.* **32**, L23 825. (doi:10.1029/2005GL024452)

Räisänen, J. 1997 Objective comparison of patterns of CO_2 induced climate change in coupled GCM experiments. *Clim. Dynam.* **13**, 197–211. (doi:10.1007/s003820050160)

Räisänen, J. & Palmer, T. N. 2001 A probability and decision-model analysis of a multimodel ensemble of climate change simulations. *J. Clim.* **14**, 3212–3226. (doi:10.1175/1520-0442(2001)014<3212:APADMA>2.0.CO;2)

Redi, M. H. 1982 Oceanic isopycnal mixing by coordinate rotation. *J. Phys. Oceanogr.* **12**, 1154–1158. (doi:10.1175/1520-0485(1982)012<1154:OIMBCR>2.0.CO;2)

Robertson, A. W., Lall, U., Zebiak, S. E. & Goddard, L. 2004 Improved combination of multiple atmospheric GCM ensembles for seasonal predition. *Mon. Weather Rev.* **132**, 2732–2744. (doi:10.1175/MWR2818.1)

Rougier, J. In press. Probabilistic inference for future climate using an ensemble of climate model evaluations. *Clim. Change.*

Schneider, S. H. 2001 What is ‘dangerous’ in climate change? *Nature* **411**, 17–19. (doi:10.1038/35075167)

Smith, R., Tebaldi, C., Nychka, D. & Mearns, L. Submitted. Bayesian modeling of uncertainty in ensembles of climate models.

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

Stott, P. A. & Kettleborough, J. A. 2002 Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature* **416**, 723–726. (doi:10.1038/416723a)

Tebaldi, C., Mearns, L., Nychka, D. & Smith, R. 2004 Regional probabilities of precipitation change: a Bayesian analysis of multimodel simulations. *Geophys. Res. Lett.* **31**, L24 213. (doi:10.1029/2004GL021276)

Tebaldi, C., Smith, R., Nychka, D. & Mearns, L. 2005 Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multi-model ensembles. *J. Clim.* **18**, 1524–1540. (doi:10.1175/JCLI3363.1)

Thomson, M. C., Doblas-Reyes, F. J., Mason, S. J., Hagedorn, R., Connor, S. J., Phindela, T., Morse, A. P. & Palmer, T. N. 2006 Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. *Nature* **439**, 576–579. (doi:10.1038/nature04503)

von Deimling, T. S., Held, H., Ganopolski, A. & Rahmstorf, S. 2006 Climate sensitivity estimated from ensemble simulations of glacial climate. *Clim. Dynam.* **27**, 149–163. (doi:10.1007/s00382-006-0126-8)

Wigley, T. M. L. & Raper, S. C. B. 2001 Interpretation of high projections for global-mean warming. *Science* **293**, 451–454. (doi:10.1126/science.1061604)

Wild, M. 2005 Solar radiation budgets in atmospheric model intercomparisons from a surface perspective. *Geophys. Res. Lett.* **32**, L07 704. (doi:10.1029/2005GL022421)

Wild, M., Long, C. N. & Ohmura, A. 2006 Evaluation of clear-sky solar fluxes in GCMs participating in AMIP and IPCC-AR4 from a surface perspective. *J. Geophys. Res.* **111**, D01104. (doi:10.1029/2005JD006118)

Yun, W. T., Stefanova, L. & Krishnamurti, T. N. 2003 Improvement of the multimodel supersensemble technique for seasonal forecasts. *J. Clim.* **16**, 3834–3840. (doi:10.1175/1520-0442(2003)016<3834:IOTMST>2.0.CO;2)