**Challenges in combining projections from multiple climate models**

Reto Knutti (1), Reinhard Furrer (2,3), Claudia Tebaldi (4), Jan Cermak (1), and Gerald A. Meehl (5)

(1) Institute for Atmospheric and Climate Science, ETH, Zurich, Switzerland
(reto.knutti@env.ethz.ch)

(2) Colorado School of Mines, CO, USA

(3) Institute of Mathematics, University of Zurich, Switzerland

(4) Climate Central, Palo Alto, CA

(5) National Center for Atmospheric Research, CO, USA

December 11, 2009, revised version submitted to Journal of Climate

**Abstract**

Recent coordinated efforts, in which numerous general circulation 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 quantifying uncertainty in future climate change. International climate change assessments also rely heavily on these models and often provide equal-weighted averages as best-guess results, assuming that individual model biases will at least partly cancel and that a model average prediction is more likely to be correct than a prediction from a single model based on the result that a multi-model average of present-day climate generally out-performs any individual model. This study outlines the motivation for using multi-model ensembles and discusses various challenges in interpreting them. Among these challenges are that the number of models in these ensembles is usually small, their distribution in the model or parameter space is unclear and the fact that extreme behavior is often not sampled. Model skill in simulating present day climate conditions is shown to relate only weakly to the magnitude of predicted change. It is thus unclear by how much our confidence in future projections should increase based on improvements in simulating present day conditions, a reduction of intermodel spread or a larger number of models. Averaging model output may further lead to a loss of signal, e.g. for precipitation change where the predicted changes are spatially heterogeneous, such that the true expected change is very likely to be larger than suggested by a model average. Finally, there is little agreement on metrics to separate 'good' and 'bad' models, and there is a concern that model development, evaluation and posterior weighting or ranking are all using the same datasets. While the multi-model average appears to still be useful in some situations, these results show that more quantitative methods to quantify model performance are critical to maximize the value of climate change projections from global models.

## 1. Introduction

With climate change over the past 50 years or so now firmly established to be mostly due to human influence via burning of fossil fuel (IPCC 2007b), concerns about future climate change of much larger magnitude than observed are increasing, and attention is increasingly directed to projections from climate models in the hope for policy-relevant information about expected changes and guidance for appropriate mitigation and adaptation measures. The degree of confidence we place on model results, however, essentially depends on whether we can quantify the uncertainty of the prediction, and demonstrate that the results do not depend strongly on modeling assumptions. Since there is no direct verification of future changes' forecasts, model performance and uncertainties need to be assessed indirectly through process understanding and model evaluation on past and present climate.

Uncertainties in future projections stem from different sources and are introduced at various stages in the modeling process. Forcing uncertainties (reflected by different economic and societal developments and political decisions) are often circumvented by focusing on (specific) projections, i.e. predictions conditional on an assumed scenario (e.g. Nakicenovic and Swart 2000). Initial and boundary conditions are mostly of minor importance for long-term climate projections. By far the largest contribution to uncertainty stems from the fact that climate models are imperfect and their projections therefore uncertain. This contribution can be further separated into model uncertainty due to limited theoretical understanding (inability to understand a process in the first place, e.g. how aerosols affect cloud formation), uncertainty in model parameters, and structural model uncertainty (inability to describe a known process accurately in the model). Parametric uncertainty is introduced by the fact that for many small-scale processes in models, their large-scale effects need to be empirically described rather than resolved, and the values in these parameterizations are not always well constrained and are not directly observable in the real world. Structural uncertainty (sometimes also termed model inadequacy) means that no set of parameters will make the model agree perfectly with observations (e.g. Sanderson et al. 2008), because certain processes are missing or are only approximated in the model (see e.g. Stainforth et al. 2007; Knutti 2008a for a more detailed discussion).

One way to study uncertainty is to consider results from multiple models. The 'multi-model' approach provides a sensitivity test to models' structural choices. Additionally, an implicit assumption exists that multiple models provide additional and more reliable information than a single model (see section 2), and higher confidence is placed on results that are common to an ensemble, although in principle all models could suffer from similar deficiencies. But for the non-expert, a collection of results is often most useful when combined and synthesized. The motivating question behind this study is how model trustworthiness can be increased by combining results of multiple models.

## 2. Model diversity: Potentials and challenges

Different scientific questions require different models in terms of resolution, components and processes, and spatial domain. However, there are also families of models of the same type, i.e. multiple models incorporating the same set of processes at similar resolutions. They partly sample the structural model uncertainty and can be seen as multiple credible approximations of the truth given some constraints in complexity and computational cost. These are often seen as coexisting rather than competing models (Parker 2006). While two models may make assumptions on smaller scales that could be seen as inconsistent, both models would agree with observations within some uncertainty (typically a sum of observational uncertainty and the structural model error) and would therefore be considered plausible. These model families usually are either variants of a single base model with perturbed parameters (so called perturbed physics ensembles, PPE) (e.g. Forest et al. 2002; Knutti et al. 2002; Murphy et al. 2004; Stainforth et al. 2005) or multi-model ensembles (MME), i.e. a somewhat arbitrary collection of different models of similar structure and complexity (e.g. Eyring et al. 2007; Plattner et al. 2008). The ensemble used here is from the recent World Climate Research Project (WCRP) Coupled Model Intercomparison Project Phase 3 (CMIP3, Meehl et al. 2007a) and consists of twenty-three state of the art atmosphere ocean general circulation models (AOGCMs) from sixteen institutions and eleven countries. One ensemble member for each model is used. The CMIP3 MME provided the basis for the projections of the latest Intergovernmental Panel on Climate Change

(IPCC) Fourth Assessment Report (AR4) (IPCC 2007b). An extensive discussion and evaluation of these models and an overview of the projections are given in the relevant IPCC chapters (Christensen et al. 2007; Hegerl et al. 2007; Meehl et al. 2007b; Randall et al. 2007). An overview of CMIP3 is given by Meehl et al. (2007a), a list of models and institutions is also provided by Gleckler et al. (2008). The data is available from the PCMDI website (http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php). The detailed structure of these models, the resolution and the exact number of models however are not relevant for the issues discussed here. Issues will be similar for future multi-model ensembles, and some of the conclusions are likely to apply to other research fields where predictive models need to be combined. Most of the discussion similarly applies to PPEs, but since structural errors are likely to be even more persistent when all models share the same core, PPEs may in fact offer even harder challenges.

**a) Prior distribution**

Synthesizing MMEs or PPEs is a problem that can be cast in a Bayesian framework, where a prior distribution determines the initial weight distribution of the sample of models or sample of predictions available for inter-comparison, and data (observations) may serve to redistribute the weight among them on the basis of their performance (the likelihood used in the Bayesian analysis). The questions most relevant to this problem then are: what is the prior distribution of these models? Is the sample randomly selected, or systematic, or neither? Is the data significantly constraining the final result or is the prior? We can fairly confidently answer the last question: for both PPEs and MMEs, data constraints are weak at best, and do not change the shape and width of the prior distribution robustly (Frame et al. 2005). It seems crucial then to consider the prior distribution, and in this respect a fundamental difference exists between PPEs and MMEs. Because of computational costs, and the large number of model parameters (typically a few dozen) a comprehensive sampling in AOGCM space is impossible. Large PPEs like climateprediction.net (Stainforth et al. 2005) use tens of thousands of members and can explore a wide range of solutions, but always structurally constrained to a single model. Most other AOGCM ensembles are small (i.e. a few tens of models). In either case a critical issue remains the definition of a uniform prior in model space, which ideally would let the data have the greater impact on the final result. There is no absolute distance metric in model space, and uniformity can only be defined with respect to a given input or output quantity. Whatever distance metric is chosen, though, in a systematic or random sampling like PPEs (e.g. Forest et al. 2002; Knutti et al. 2002; Murphy et al. 2004; Stainforth et al. 2005) it is at least clear how the models are distributed, while for MMEs like CMIP3, the models are sampled neither randomly nor systematically but the ensemble is determined by whichever modeling center had interest and resources to contribute. Most groups provide only their 'best' model, so the ensemble should be seen as a collection of carefully configured 'best estimates' rather than an attempt to sample the range of all possible models. In an ensemble of opportunity where the number of models is small, the problem of not sampling the full uncertainty range may thus be severe. The ensemble sampling also changes from one intercomparison to the next, so projections in future ensembles may change just due to the prior sampling (i.e. model selection in the intercomparison) even if the understanding of the system does not change.

A simple case where the prior distribution was found to matter even within the set of CMIP3 models was the distribution of simulations across scenarios. Figure 1a shows the mean and one standard deviation global surface temperature ranges for the SRES A2, A1B and B1 scenarios (Nakicenovic and Swart 2000) and the historical run from all CMIP3 models. The two lower emissions scenarios (B1 and A1B) appear to have a wider uncertainty range in 2100 than the high emission A2 case, although based on often used pattern scaling arguments one would expect the opposite as the uncertainty scales approximately linearly with warming (Knutti et al. 2008). Figure 1b shows the standard deviations and minimum maximum ranges in 2100 for all models and the subset of models that have run all scenarios. The apparent contradiction in uncertainty ranges in Fig. 1a occurs simply because fewer models had run the higher A2 scenario. If the subset of models with all scenarios available is considered, the uncertainty is strongly reduced for the lower emissions scenarios. Further details are given by Knutti et al. (2008). The uncertainty estimate based on the empirical distribution

of the projections of these models can only be wider than the prior distribution of the ensemble if the variance is artificially inflated. This is a challenge since it requires assessing the likelihood of outcomes that are not simulated by any model. Here we note that the consequences of underestimating the uncertainty from an ensemble of simulations have important repercussions when the simulations are used as input to impact models. That is, "best case" and "worst case" outcomes that could directly affect the severity of impacts may be missed in favor of more centrally distributed and less drastic outcomes.

**b) Model averages, independence and structural error**

There is empirical evidence from various areas of numerical modeling that a multi-model average yields better prediction or compares more favorably to observations than a single model. Examples include health (Thomson et al. 2006), agriculture (Cantelaube and Terres 2005), predictions of the El Niño Southern Oscillation (Palmer et al. 2005) and detection and attribution (Gillett et al. 2005). Weather and seasonal forecasts show improved skill, higher reliability and consistency when multiple models are combined (Krishnamurti et al. 1999; Doblas-Reyes et al. 2003; Yun et al. 2003). For a single variable, the multi-model combination might not be significantly better than the single best model, but a large benefit is seen when the aggregated performance on all aspects of the forecast is considered (Hagedorn et al. 2005). Models can simply be averaged ('one model one vote') or can be weighted e.g. using Bayesian methods, where weights are based on past relationships between forecasts and verifications. Weighted averages are found to perform better in many cases (Robertson et al. 2004; Min and Hense 2006; Peña and Van den Dool 2008; Weigel et al. 2008), provided that sufficient information is available to determine the weight (see section 2c).

For several generations of climate models it has been shown that the multi-model average for a variety of variables mostly agrees better with observations of present day climate than any single model, and that the average also consistently scores high in almost all diagnostics (Lambert and Boer 2001; Phillips and Gleckler 2006; Randall et al. 2007; Gleckler et al. 2008; Pincus et al. 2008; Reichler and Kim 2008; Pierce et al. 2009). While the improvement was sometimes quantified in these studies, it was rarely discussed whether the improvement was as large as expected and how it should relate to improvements in projections.

Near-surface temperature is used in the following section for illustration because models can simulate temperature reasonably well and because good observations and reanalysis datasets are available, but similar results are expected for other variables. Fig. 2a shows the mean bias of local temperature for a collection of single models for boreal winter and summer (the absolute bias at every grid point averaged across all models), whereas Fig. 2b shows the absolute value of the bias for the multi-model average. Shifts towards smaller values of the biases mean that the multi-model average performs better in simulating the climatological temperature field. There is indeed improvement in some areas, but other locations are almost unaffected by averaging, indicating that errors are similar in many models. The largest errors also tend to be in the same locations where model spread is large. These are often caused by known deficiencies in the models not resolving processes accurately due to resolution (e.g convection or coastal upwelling in the ocean, topography of mountains), not representing processes well due to inappropriate parameterizations or poor parameter choice (e.g. tropical variability related to ENSO) or not representing them at all (e.g. forcings not considered, lack of vegetation model, etc.). Note that observations (whether reanalysis or station data) are not always accurate and also exhibit biases. This point is discussed at the end of this section.

The fact that in some areas the biases are not reduced by averaging implies that the errors are not random but correlated across models. A histogram of all pair wise correlation values of two model bias patterns (Fig. 3a/b) confirms that. Correlations are largely positive and reach values up to 0.9 in cases where the two models from the same institution (e.g. the two GFDL models) or where two versions of the same model but different resolution (e.g. CCCMA) are compared. The result is that simple averaging is not very effective. It is instructive to study how the bias is reduced as the size of the ensemble used in the average is increased. Fig. 3c/d shows the root mean square (RMS) bias of the model average as a function of the number of models (i.e. averaging the models first, calculation of

bias, then averaging over space to estimate a typical local bias of the model average). The solid red curve shows the average resulting from taking many different subsets of models, the red dashed lines indicate the range covered by different random subsets. If all model errors were random, the error of the mean should decrease with the square root of the number of models (black dotted). Indeed it does, but not to zero but to a residual that is more than half the initial value. If we assume a correlation structure between the grid points of a pattern, e.g. as illustrated in Fig. 3a/b, then it is possible to calculate the theoretical RMS (Fig. 3a/b, black dashed) and to show that it converges to $\sigma\sqrt{\rho}$ with $\sigma$ the variance of the pattern and $\rho$ the average correlation (see Fig. 3 caption for full equation)

The interesting conclusions from Fig. 3c/d are that for present day temperature, half of the typical biases would remain even for an average of an infinite number of models of the same quality. The remaining bias for 22 models is two to three times larger than if the models were independent and the errors were purely random. Considerable improvement is only seen for up to about five models, after 10 models the biases are almost stable. The blue lines indicate the bias of the subset of the two best (best in this case meaning the GCMs whose DJF and JJA surface temperatures 1980-1999 agree most closely with ERA40), three best etc. models and suggest that a few good models are better than the multi-model average, and the average soon gets worse when poorer models are added. It should be noted here that the models that are best for temperature are not necessarily best for other quantities, but there is a tendency for good models to score high on many diagnostics (Gleckler et al. 2008). This reflects in part the amount of effort going into the development of a model, but also the fact that many variables in the climate are linked, such that biases in one will lead to biases on many others.

The reduction of biases by averaging depends not only on the geographical location but also on the magnitude of the initial bias. Fig. 4 shows the distribution of present day temperature biases in present day climatology for each model, the multi-model mean and the average of the five best models (with regard to simulating temperature as described above). The distribution of the model average has more pronounced long tails than the single models, i.e. the extreme errors are reduced less effectively than smaller amplitude errors, suggesting that there are indeed large errors resulting from processes that are similarly misrepresented in many models, and therefore hard to eliminate. As a caveat, observational and reanalysis datasets of course also have biases. For temperature these are quite small and conclusions would be similar for station data or other reanalysis datasets. But for evaluation on other quantities, e.g. radiation or precipitation where little or no data is directly assimilated, this is a major issue. Models used for reanalysis run at much higher resolution that those in CMIP3, but results for some variables are still based on parameterizations similar to those in CMIP3 models. In an analysis such as the above, any bias in the observations (whether station or reanalysis) would appear as a model error persistent across all climate models, and the conclusions drawn above for temperature are only justified because the model errors are generally much larger than the uncertainties in the observations (see e.g. Knutti et al. 2006, Fig. 1/2).

**c) Dependence of projection uncertainty on the number of models**

As shown in the previous section, the biases of individual models are correlated. For the present day state, sophisticated statistical methods also suggest that the equivalent number of independent models in CMIP3 is much smaller than the total number of models (Jun et al. 2008a; Jun et al. 2008b). It is unclear how this model dependency structure maps into the future, and by construction it is impossible to determine that because of the lack of 'observed truth' for the future.

Physical understanding of the assumptions underlying the models suggest that the models are not independent and distributed around the truth, yet many Bayesian methods assume model independency (Giorgi and Mearns 2002; Giorgi and Mearns 2003; Tebaldi et al. 2004; Tebaldi et al. 2005; Greene et al. 2006; Furrer et al. 2007b; Tebaldi and Knutti 2007; Smith et al. 2009; Tebaldi and Sanso 2009). The assumption of independence is equivalent to the interpretation that each model approximates the real world with some random error. Like in a case where a quantity is measured with a random error, multiple measurements will improve the accuracy of the measurement average, and the uncertainty of the mean value will shrink with the square root of the number of measurements $N$ as $N$ increases (Lopez et al. 2006). The implication of the independence assumption is that uncertainties

decrease as more models are considered, shown for illustration in Fig. 5 for the method of Furrer et al. (2007b; 2007a) and Smith et al. (2009). Because these methods determine a central tendency common to all models rather than a posterior predictive of a single ideal model projection, it is not surprising that the uncertainty (measured in terms of the width of the PDF) decreases with more models. So it comes down to a somewhat philosophical question of the quantity we are actually trying to estimate.

The signal of change underlying truth and model simulation is the abstract concept whose uncertainty is characterized by the posterior distribution, which in these Bayesian treatments decreases in width with the number of models considered. An alternative view of the simulations' relation to the uncertain future climate is to consider each model trajectory as a possible future path for Earth's climate, and accordingly represent the uncertainty in future projections by the posterior predictive distribution of a new GCM, whose width is of the same order of magnitude as the range of model projections. This proposal is made explicit in Tebaldi and Sanso (2009).

In contrast, the probabilistic method by Watterson (2008) and studies based on detection and attribution fingerprint scaling (Allen et al. 2000; Stott and Kettleborough 2002; Stott et al. 2006) assume no improvement with additional models, equivalent to the assumption that each model is a plausible representation of true world, and one of them is right, but we do not know which one (of course strictly no model is right because we know they are all incomplete, although the prediction of one could be right for a certain quantity). This issue of how much additional information is provided by more models is only now being explored and discussed. While it may be reasonable to assume some improvement in a projection when a few models are available as opposed to one (in particular when predicting many quantities), a very large number of models should not infinitely improve our confidence, as long as they are based on the same knowledge, make similar assumptions, or worse (but quite common) if they use parts of the code of existing models.

## 3. Model evaluation and weighting

Strictly, calibration and evaluation of climate model predictions is impossible as projections of climate change relate to a state never observed before. As we cannot evaluate centennial prediction, the evaluation of climate models is therefore on the observed present and past climate rather than the prediction, and if the model matches the observed data it only tells us that the data is consistent with the model. One of the difficulties is that the observations often have been used in the modeling process before, to derive parameterizations, or to tune earlier versions of models. There is therefore a risk of double counting information, overconfidence or circular logic if model evaluation and weighting is done on the same datasets that were used to develop the models. Thus it is important that models are used to simulate past climates much different from today as part of the process to establish credibility of the model responses.

If model and data do not agree, it could be for a variety of reasons, and could depend on the particular metric being evaluated, how various quantities interact in the model, and which parameter is compared to what observations. For any of these reasons the interpretation could result in a judgment that the model is in error, the observational data are inadequate, or a combination of both. Agreement between model and observed data however should be seen as a necessary but not sufficient condition (Oreskes et al. 1994). Note here that we should not expect perfect accuracy from models, but we can be satisfied with models that are adequate for a particular purpose. Weather forecast models for example do not contain a dynamic ocean component, yet prove to be useful for the purpose of predicting weather for the next few days (and adding an ocean would not improve the forecast). An energy balance climate model does not even resolve the dynamics of the atmosphere, but can easily replicate the global temperature evolution over the past century (Knutti et al. 2002; Meinshausen et al. 2009).

### a) Why do we trust models?

Confidence in climate models comes from the fact that they are at least partially based on physical principles known to be true (e.g. conservation of mass, energy and momentum), and from the fact that we understand the results in terms of physical processes (Bony et al. 2006) and can track them across hierarchies of models (Held 2005). Climate models reproduce many aspects of the current climate and

its forced and unforced variability quite well (Räisänen 2007; Randall et al. 2007; Gleckler et al. 2008). Trends over the instrumental period resulting from anthropogenic forcings are well captured in most models (Barnett et al. 2005; Hegerl et al. 2007; Knutti 2008b). An important independent line of evaluation is provided by paleoclimate evidence, e.g. from the last glacial or interglacial period. Since boundary conditions for paleoclimate are quite different from today's climate, a model's ability to simulate past climate is an illuminating test of the model assumptions. Though uncertainty in proxy data is often large for the distant past, some credibility of the model as well as the actual climate system response to different forcings can be established (e.g. Liu et al. 2009). On the other hand, in some models the response to radiative forcing in a Last Glacial Maximum state for example has been shown to be quite different from the response in current climate (Crucifix 2006; Hargreaves et al. 2007). The nature of the forcings for paleoclimate and future projections is often also quite different.

Models continuously improve in simulating the present-day climate (Reichler and Kim 2008), and general aspects of projections from newer models usually agree with older ones. Model agreement is often interpreted as increasing the confidence, however there is no obvious way to quantify whether agreement across models and their ability to simulate the present or the past implies skill for predicting the future. A more in depth discussion of these topics is given by Smith (2002), Tebaldi and Knutti (2007) and Knutti (2008a).

**b) Model evaluation and tuning**

In computationally cheap climate models the calibration of parameters can be done by minimizing some cost function using search algorithms (e.g. Andronova and Schlesinger 2001; Forest et al. 2002; Knutti et al. 2002; Knutti et al. 2003; Annan et al. 2005; Beltran et al. 2005; Frame et al. 2006; Hegerl et al. 2006; Meinshausen et al. 2009). Because of the complexity of AOGCMs and the associated computational cost, model tuning (defined as the adjustment of a model parameter within some known observational range) or calibration by automated procedures (e.g. finding optimal parameter values by minimizing some error metric) is usually unfeasible. Model calibration is mostly done in individual parts of the model and involves expert judgment. Formal metrics to quantify agreement with data are complemented with experience from other models to make choices. The number of intermediate versions of a coupled GCM that can be afforded is small, often only a few to a few tens before a final version is selected (CCSP 2008, Table 4.2). In the few cases where large perturbed parameter ensembles were calculated, the standard model was found to be surprisingly close to the best performing model (e.g. Sanderson et al. 2008, Fig. 7l) given the enormous degrees of freedom resulting from dozens of uncertain parameters. This suggests that expert judgment is very efficient in finding a good model (relative to the other models in the set) with a small number of trials. The model evaluation process is often not documented and is rarely based on clear procedures and statistical methods. Apart from the computational cost, one reason certainly is the fact that the metric to minimize is not clear. As discussed above, climate models serve multiple purposes, so it is not even clear what the best model (given some finite resources) would be. Tuning of model parameters in the sense of blindly minimizing errors without understanding the model behavior or going outside known observational uncertainty is therefore not common in GCMs, and available observations are clearly relied upon for guidance in physically plausible tuning.

Statistical methods to evaluate and weight models are not routinely used in GCM development, but they have been used to a posteriori combine models or determine parameter ranges and distributions from both PPE (Murphy et al. 2004; Piani et al. 2005; Knutti et al. 2006; Murphy et al. 2007) and in Bayesian methods using MME (Giorgi and Mearns 2002; Giorgi and Mearns 2003; Tebaldi et al. 2004; Tebaldi et al. 2005; Greene et al. 2006; Furrer et al. 2007b; Tebaldi and Knutti 2007; Smith et al. 2009; Tebaldi and Sanso 2009). However, the field is still in its infancy and no consensus exists on how models should be best evaluated. In the IPCC AR4 (Randall et al. 2007) the evaluation of models was mostly an expert assessment discussing what aspects of climate are well simulated, where models have improved and what difficulties remain. The models were assessed as a group rather than individuals and future projections did not weight individual models or select subsets. No performance metrics or rankings were proposed. Results were either presented as multi-model equal-weighted averages or as a collection of individual models to show the model spread, but without any

quantitative information of how the model spread should be interpreted or on which models may be more credible.

The main issue with model performance is that there is virtually an infinite number of metrics that can be defined, and a large number of them may be defensible for certain purposes. Whether a model is 'good' or 'bad' depends on the question at hand. Models have been evaluated on many different quantities but mostly on the present day mean climate and variability, though for the upcoming CMIP5 coordinated model experiments, paleoclimate simulations will be used for the first time as a standard part of the evaluation process (Taylor et al. 2008) . Prior to this, the present-day climate was used as a standard reference at least partly because most of the observations are about the present-day mean state and variability, and because we believe that we understand many aspects of the present climate rather well. It may also be a remnant from earlier times where multi-century transient simulations were not yet possible. The question of whether the simulation of the present day climate matters for future projections is difficult to evaluate. For example, large efforts go into improving tropical variability in models' simulations of El Niño, and significant improvements have been made, yet models do not agree on the sign of future El Niño change. This is at least partly because there is large multi-decadal and centennial timescale variability of ENSO (seen in observations and in multi-century control runs from climate models), and sampling issues related to this non-stationary base state cause difficulties in evaluating what the future behavior of El Nino may be. For many large scale changes, e.g. temperature projections which still are uncertain by about a factor of two even on the global scale, tropical variability related to ENSO is probably of minor importance, as the effect of ENSO on global temperature is only on the order of 0.1°C. On the other hand ENSO will have a strong effect on Australian water availability. Therefore, using ENSO as an example where inherent low frequency variability may make it difficult to ever provide an accurate projection of future El Nino behavior, model development and evaluation is often done based on processes of interest rather than on an analysis of what quantity would be most important to be well represented in the model to make an accurate prediction. Multi-model ensembles are of value here because they allow a determination as to why models agree or disagree, thus shedding light on where efforts are best spent to improve a prediction. Indeed the CMIP3 archive has sparked many attempts to isolate why models differ, e.g. by quantifying agreement in different feedbacks (e.g. Bony et al. 2006; Soden and Held 2006).

Good agreement with observations in one metric does not guarantee good performance in other variables, but correlation of performance across variables at least within one component of the climate system are quite high because many variables are influenced by the same processes and parameterizations. Models that represent some basic variables like temperature and precipitation well often also perform well in other variables (e.g. Gleckler et al. 2008).

**c) Model weighting**

Models can be combined by experts defining certain (sometimes ad hoc) selection criteria to pick subsets of more skillful models. In the absence of formal methods to weight models other than including or excluding them, this may be a useful approach. To give a few examples, van Oldenborgh et al. (2005) quantified the effect of climate change on ENSO using a subset of the CMIP3 models, and several studies predicted changes in Australian rainfall and runoff based on subsets and rankings of models (Perkins et al. 2007; Maxino et al. 2008; Pitman and Perkins 2008). Waugh and Eyring (2008) and Eyring et al. (2007) assessed the performance of stratospheric chemistry climate models but found only small differences between weighted and unweighted projections. Schmittner et al. (2005) produced projections of future changes in the Atlantic meridional overturning circulation and also found unweighted averages to be similar to the weighted ones, but report a decrease in the model spread after weighting. Santer et al. (2009) found that detection and attribution of water vapor changes are insensitive to model quality. Temperature and precipitation changes were calculated based on CMIP3 by considering all models (Meehl et al. 2007b), or based on weighting with current climatology (Giorgi and Mearns 2002; Giorgi and Mearns 2003; Tebaldi et al. 2004; Tebaldi et al. 2005). Greene et al. (2006) additionally used observed trends for weighting.

In the area of weather and seasonal prediction the ensemble approach is well established (e.g. Fraedrich and Leslie 1987; Doblas-Reyes et al. 2003; Hagedorn et al. 2005). Despite the success of combining models in other areas, such attempts are still rare in the climate community and many people are reluctant to deviate from the interpretation of the family of coexisting models (Parker 2006). Some scientists argue that we cannot attach weights, produce meaningful PDFs or even define the space of plausible models (Stainforth et al. 2007), because all models have essentially zero weight. This may be strictly true, but from a pragmatic point of view, model selection is routinely already done. Newer models are developed and older ones are phased out in newer intercomparisons and IPCC reports (Meehl et al. 2007b), thus giving them zero weight. Clearly there are many issues with PDFs derived through statistical analysis of MMEs, but the problem may lie more in how to communicate and interpret them than in whether or not they should be constructed in the first place. PDFs are of course always conditional on the model, statistical assumptions and observational constraints (although that generally is not informative for the decision-maker without further discussion of the model trustworthiness).

One way to test whether some observed quantity is important for a prediction is to consider the correlation between the observed quantity and the prediction across a set of models. If the correlation is weak then the observation likely has little effect on the prediction, and weighting or creating a subset based on that quantity will not impose any constraint on the prediction. It may however introduce spurious biases in the projection if the sample size is small (selecting a subset of five or so models out of twenty will have the tendency to reduce the model range and variance even if the subset is chosen randomly). If the correlation between an observation and a prediction is strong, that means that the observation may be a good predictor for the quantity of interest, and a constraint on the observation will constrain the future. The assumption of course is that the correlation across several models represents the influence of a process that affects both the observation and the prediction, and not just the simplicity of the underlying model, i.e. that all models are based on the same parameterizations. In many cases (e.g. where observed greenhouse attributable warming is related to future warming) this assumption is justified, but in particular in PPE where all models share the same structural core and many constraints are applied without understanding the processes behind it, correlation may indeed be unphysical. There is also the assumption that an observation does not manifest itself differently in different models due to nonlinear interactions, or that several observations can have different realizations in different models due to such interactions.

The correlations between predictions and observation features of the current climate mean state (which is predominantly used for model evaluation) are predominantly weak if existent at all. Fig. 6a/b shows the correlation of the CMIP3 seasonal temperature biases as compared to ERA40 near surface temperature 1980-1999 (aggregated as root mean square over space) and future global seasonal warming 2080-2099 in the A1B scenario. Correlations are vanishingly small, and even if both winter and summer are considered and related to simple quantities like the transient climate response and climate sensitivity (Fig. 6c/d) to exclude the effect of different forcings, the correlations do not improve. Thus the climate response does not seem to depend in an obvious way on the pattern of 20$^{th}$ century temperature at least in the range of models considered.

This is consistent with the fact that newer climate models reproduce the current climate significantly more accurately than older ones (Randall et al. 2007; Reichler and Kim 2008), yet the spread of projections on both global and local scale is not decreasing very much (Knutti et al. 2008). For example, the range of climate sensitivity has only decreased from 1.5° - 4.5°C to 2.0° - 4.5°C over the last two decades. In the most recent CMIP3 intercomparisons, the standard deviation of all model climate sensitivities was 0.69, reduced but not significantly so from the earlier CMIP1 and CMIP2 intercomparisons (0.78 and 0.92, respectively, one standard deviation), after several years of model development, more observational data for model evaluation and an increase in computational cost of probably at least two orders of magnitude. These results are also consistent with recent studies that found only a weak statistical relation between observations of the present day climate and climate sensitivity (Murphy et al. 2004; Piani et al. 2005; Knutti et al. 2006; Sanderson et al. 2008; Huber et al. 2009). It is also in agreement with the results of Jun et al. (2008a) who noted that there was very

little correlation between the ability of the climate models to simulate the observed patterns of the mean temperature and the observed patterns of the temperature trend.

Rather than relating global performance in simulating surface temperature to the prediction of future warming, one may argue that model performance on local to regional scales should be considered and used for regional projections. That is, if there is a high correlation between a good performing model in a certain location and a preferred value of future climate change, perhaps that agreement would provide information on what the future climate change may be. But there is still no guarantee that good present performance is a predictor of future climate change since the future climate change is unknown. Fig. 7 shows the correlation between the performance in simulating current surface temperature and predicted warming by the end of the century in the A1B scenario at each grid point. While there is correlation between the present surface temperature and future warming in some locations, the correlation is weak. On land its absolute magnitude exceeds 0.4 only in a few locations, indicating that the current temperature explains less than 20% of the model spread in most locations. But because the number of models is small, many correlations occur by chance, and the distribution of correlations for all grid points does not strongly differ from a null hypothesis where all points were purely random (Fig 7b/d). Most correlation should therefore be seen as essentially random. Based on a similar analysis but using regional patterns and multiple variables, Whetton et al. (2007) concluded that applying weights based on present day climate was useful, but correlations (in their case based on the regions defined by Giorgi and Francisco 2000) also rarely exceeded 0.4 and therefore provide a very weak constraint. Cases have been reported where correlations between observations and predictions are strong, e.g. between the seasonal cycle and climate sensitivity (Knutti et al. 2006), radiation patterns and climate sensitivity (Huber et al. 2009), the albedo feedback on seasonal and decadal timescales (Hall and Qu 2006), past and future sea ice reduction (Boe et al. 2009) or for the amplification of tropical surface temperature variability on short and long timescales (Santer et al. 2005). In some cases weighting or selecting a subset of models leads to smaller model spread for predictions, e.g. for Australian rainfall (Perkins et al. 2009; Smith and Chandler 2009), but in most cases these are carefully selected quantities based on process understanding rather than a broad aggregation of model biases across space, time and variables.

However, one may also argue that evaluating models only on a single variable is an approach that takes a too narrow focus. Reichler and Kim (2008) evaluated the CMIP3 models on a large set of variables of the present day mean climate state and combined all errors into a single number. Here we use an updated version of their results using four seasonal mean results from 37 different observed climate quantities (T. Reichler, personal communication), but only use the ranking of the models rather than the overall performance index in order to generate different subsets. The hypothesis is that the spread of projections from a subset of $N$ 'good' models should be smaller than the spread of all models. Fig. 8 shows the ratio $R$ of the standard deviation of $N$ good models to the standard deviation of all models for the precipitation trend (percent change in local precipitation per Kelvin change in global temperature calculated for each grid point and each model for the period 1900 to 2100). The ratio $R$ is determined at every grid point and the results are shown as box plots, for the two best models (leftmost), three best (second from left), etc. models. Ratios $R$ of less than unity indicate that the spread of the $N$ best models is smaller than the spread of all models. Note that the precipitation response is spatially heterogeneous, so one would not expect the box plot widths to decrease with a subset of better models, but one would expect $R$ to be lower *on average* if the subset of models was in closer agreement that the set of all models. The median (box center) and mean (red solid) of $R$ at all locations is indeed smaller than unity, but a close inspection shows that most of that effect is an artefact of the standard deviation being a biased estimator of the spread for small samples. For normally distributed numbers and sample size $N$=2 the standard deviation underestimates the real spread by roughly 20%, for N=5 by 6%, for $N$=10 by 3% (red dashed line). The maps show $R$ for a subset of 11 vs. all 22 models, and reveal a picture without any obvious structure. In fact the same figures for a random subset of models are almost undistinguishable. The somewhat surprising conclusion from this analysis is that if one would perform a ranking of all models based on a comprehensive set of present day diagnostics and select a subset of models that agree well with

observations, the tendency for a better constrained projection (on average in all locations) would be very small in most cases.

Probabilistic projections based on fingerprint scaling (Allen et al. 2000; Stott and Kettleborough 2002; Collins et al. 2006; Harris et al. 2006; Stott et al. 2006) also amount to reweighting, albeit of a single model's projections, but in this case the relation between the past and present is clearly quantified and understood in terms of the overall feedback strength; models with a stronger greenhouse warming over the past decades show a higher warming in the future. Similar arguments hold for probabilistic studies of global temperature constrained by the observed surface warming, ocean heat uptake and radiative forcing (e.g. Forest et al. 2002; Knutti et al. 2002) where the performance of a model often relates clearly to the future prediction, at least for the next few decades.

In summary, this section highlights the difficulty with weighting models based on observations. Correlations between observed quantities and predictions are small in many cases, resulting in little if any change in a weighted average and small reductions in the model spread. This does not imply that a weighting of models is impossible in principle, but it indicates that the choice of a meaningful metric is far from trivial. A few recent studies reported a reduction in model spread after evaluating the models on multiple criteria, but whether the prediction is in fact more accurate remains to be seen (and it will take a long time to find out). In most studies the weighted averages and model spread are similar to those of the unweighted ensemble, a result explained by the absence of correlation between the observations used to weight the models and the models' future projections.

**4. Model combination and loss of signal**

A last issue that deserves attention is the fact that an average of multiple models may show characteristics that do not resemble those of any single model, and some may be physically implausible. If two variables $x$ and $y$ are related in a nonlinear way, then the average of $x$ and the average of $y$ from several models will not follow the original relation between $x$ and $y$. So a model average state may not even be physically plausible. In cases where there is a bifurcation between multiple solutions, an average state may not exist. While these issues may not be serious for most large scale climate projections as long as the perturbations are not large, there is the issue of loss of signal that is serious and has not been addressed so far. One such case is the predicted change in precipitation resulting from anthropogenic warming. While models agree on the large scale drying in the subtropics and wettening of the high latitudes, the locations of the maximum changes are often a bit shifted. In some areas the models also predict opposite signs in the trends. Fig. 9a/c shows multi-model precipitation changes displayed as maps, similar to those presented in the IPCC AR4 (IPCC 2007a, Fig. SPM 7), except that trends (as before) in percent change in local precipitation per degree change in global temperature are calculated for each grid point and each model for the period 1900 to 2100 rather than showing the difference between the end of the 21st century and the present as in IPCC. The trends are used to maximize the signal using a 200 yr period and to reduce the effect of different global temperature change in the different models. If anything, the models should agree better among each other, but the maps are very similar to those in the IPCC AR4 (IPCC 2007a) and the conclusions will not depend on these choices. Changes are shown for the annual mean rainfall (Fig. 9a/b) and for the dry season (Fig. 9c/d), i.e. for the driest three consecutive months in the present at each grid point. Further details are given by Solomon et al. (2008). A histogram of the land area (restricted to 60°S-60°N, as this is essentially the area relevant for agriculture) that is undergoing a certain change in precipitation shows that almost every model (light blue lines) shows drying of more than 15%/K for the annual mean and of more than 20%/K in the dry season in some places, but the multi-model average (black) does not. The distribution of precipitation is much narrower for the model average because changes are of opposite sign or maxima are not co-located in the individual models. If the distributions of the individual models are averaged (dark blue), the distribution is about 50% wider than the multi-model distribution (black). If we interpret the CMIP3 models as a collection of predictions of which one may be the truth (of course none is exactly) but we do not know which one, then the average precipitation change expected is 50% larger than the multi-model mean suggests. Large drying may well occur in some locations even if the multi-model average has lost that signal. This is particularly disturbing because plants have thresholds beyond which they can no longer

survive, and the difference in impacts between using the individual models or the multi-model may thus be very large. The presentation of a multi-model mean map for precipitation without any further discussion of this problem may therefore be misleading, especially if used to inform adaptation decisions. The idea of robust decision making (Lempert and Schlesinger 2000; Dessai et al. 2009) requires sampling of a broad range of outcomes, and precipitation is a good example where such concepts are likely to be more useful than model averages.

## 5. Conclusions

In this study we have shown that extracting policy-relevant information and quantifying uncertainties from ensembles of opportunity of climate models is difficult. The prior distribution of the models is important but unclear, except that it is likely too narrow and not capturing the full range of plausible models. An average of models compares better to observations than a single model, but the correlation between biases amongst CMIP3 GCMs makes the averaging less effective at canceling errors than one would assume. For present day surface temperature for example, half of the biases would remain even for an infinite number of models of the same quality. Extreme biases tend to disappear less quickly than smaller biases. Thus, models are dependent and share biases and the assumption of independence made in some studies is likely to lead to overconfidence, if the uncertainty is measured by the standard error of the ensemble means (inversely proportional to the square root of the ensemble size). Quantitative methods to combine models and to estimate uncertainty are still in their infancy. Some studies have proposed ad hoc methods for weighting or selecting subsets of models but few have demonstrated any improvement in the projections' skill or that the evaluation criterion is even relevant to the forecast. International assessments by IPCC (Randall et al. 2007) or CCSP (CCSP 2008) evaluate models but provide little information of how model error/bias translates into bias in future projections. They show what models can and cannot simulate in the present, but a discussion whether this should make us confident or not for the predictions made is often missing. The issue of combining models will become more important with the availability of more computing power and more models. Future ensembles may be more heterogeneous as some models include more components (e.g. chemistry, ice sheets, dynamic vegetation, upper atmosphere, carbon cycle, land use), and some groups are starting to produce perturbed physics ensembles with their model. One would hope that a model that can reproduce many observed features is a better model than one that is unable to do so. However, defining performance metrics which demonstrably relate to prediction skill remains a largely unresolved problem. It is shown here that most straightforward metrics (e.g. root mean square errors from climatology) do not correlate with future projections on the large scale. Local biases also correlate weakly with local projections. Selecting subsets of models based on an overall evaluation of how they simulate present day climatology is shown to have a small effect on the spread of projections. While there may be benefits in selecting subsets of models in certain areas after careful process-based assessments, a general recipe or an overall model ranking for all purposes seems unlikely to exist.

Understanding what makes the projections of two models agree or disagree, evaluating models on key processes, developing metrics that demonstrably relate to projections and searching for emerging constraints in the system on the basis of observations (Knutti et al. 2002; Stott and Kettleborough 2002; Hall and Qu 2006; Knutti et al. 2006; Huber et al. 2009) may be ways forward. Large perturbed physics ensembles with multiple models may help find constraints valid across structurally different models. Seamless prediction, i.e. the initialization with observations and evaluation on weather and seasonal timescales using climate models could help provide constraints on feedbacks which operate on both short and long timescales. Paleoclimate provides another opportunity to evaluate models, although observational uncertainties tend to be large in the distant past. New methodologies such as stochastic-dynamic parameterization (Palmer et al. 2009), where stochastic parameterization schemes are devised to represent model uncertainty to produce the benefits of a multi-model ensemble in a single model, could eventually provide an alternative to the current multi-model ensemble methodology.

The community would benefit from new methods to intelligently combine perturbed physics and multi model ensembles, as well as statistical methods that can incorporate structural model uncertainty.

Thus, taking advantage of the characteristic that the multi-model ensemble average out-performs any individual model, methodologies could be developed to better assess uncertainty by using this information combined with characteristics of the range of model realizations. Such methods, however, require a set of models that have a reasonable spread to begin with. In fact it is likely that more could be learned from a model that is on the outer edge of the range than from another model near the center of the range. Model diversity is important for these kinds of multi-model exercises, with the range of model realizations providing information that informs the plausible spread of model realizations. It is also more useful if the data used for development is not the same as that for evaluation and weighting.

Given the demonstrated difficulties in defining model performance and the lack of consensus on selecting and weighting models, methods to combine models should be assessed carefully and compared to multi-model ensemble averages and information derived from model spread. The overconfidence achieved by improper weighting may well be more damaging than the loss of information by equal weighting or no aggregation at all. As long as there is no consensus on how to properly produce probabilistic projections, the published methods should be used to explore the consequences arising from different specifications of uncertainty.

The lack of consensus on combining models also underscores the need for decisions that are robust against alternative future climate outcomes (Lempert and Schlesinger 2000; Dessai et al. 2009). In certain cases, the simple specification of a few illustrative models as alternative plausible outcomes without probabilities (similar to the illustrative SRES scenarios) may also be a useful and transparent choice to test the sensitivity of adaptation and policy decisions to the uncertainty in future climate change. However, there is some danger of not sampling the extreme ends of the plausible range with a few cases, e.g. very high climate sensitivities which are not present in CMIP3 (Knutti and Hegerl 2008), and the danger that the illustrative models will be interpreted as equally likely even if no probabilities are specified. In any case we feel that as the amount of data from climate models grows and as the dependency structure across the ensemble get more complex when perturbed parameter versions of some models become available, metrics to evaluate models and quantitative methods to extract the relevant information and synthesize it are urgently needed.

**Acknowledgements**

Discussions with a large number of colleagues and comments from three reviewers helped improve the manuscript. We acknowledge the modelling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy (DOE). Portions of this study were supported by the Office of Science (BER), U.S. DOE, Cooperative Agreement No. DE-FC02-97ER62402, and the National Science Foundation. The National Center for Atmospheric Research is sponsored by the National Science Foundation. Members of the International Detection and Attribution Working Group (IDAG) acknowledge support from DOE Office of Science (BER) and NOAA Climate Program Office.

## References


Allen, M. R., P. A. Stott, J. F. B. Mitchell, R. Schnur, and T. L. Delworth, 2000: Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, **407,** 617-620.

Andronova, N. G. and M. E. Schlesinger, 2001: Objective estimation of the probability density function for climate sensitivity. *J. Geophys. Res.*, **106,** 22605-22612.

Annan, J. D., J. C. Hargreaves, N. R. Edwards, and R. Marsh, 2005: Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter. *Oc. Model.*, **8,** 135-154.

Barnett, T., F. Zwiers, G. Hegerl, M. Allen, T. Crowley, N. Gillett, K. Hasselmann, P. Jones, B. Santer, R. Schnur, P. Scott, K. Taylor, and S. Tett, 2005: Detecting and attributing external influences on the climate system: A review of recent advances. *J. Clim.*, **18,** 1291-1314.

Beltran, C., N. R. Edwards, A. Haurie, J.-P. Vial, and D. S. Zachary, 2005: Oracle-based optimization applied to climate model calibration. *Env. Mod. Assessm.*, **11,** 31-43.

Boe, J. L., A. Hall, and X. Qu, 2009: September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nature Geoscience*, **2,** 341-343.

Bony, S., R. Colman, V. M. Kattsov, R. P. Allan, C. S. Bretherton, J.-L. Dufresne, A. Hall, S. Hallegatte, M. M. Holland, W. Ingram, D. A. Randall, B. J. Soden, G. Tselioudis, and M. J. Webb, 2006: How well do we understand and evaluate climate change feedback processes? *J. Clim.*, **19,** 3445-3482.

Cantelaube, P. and J. M. Terres, 2005: Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A*, **57,** 476-487.

CCSP, 2008: *Climate Models: An Assessment of Strengths and Limitations. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research*. Department of Energy, Office of Biological and Environmental Research, Washington, D.C., USA, 124 pp.

Christensen, J. H., B. Hewitson, A. Busuioc, A. Chen, X. Gao, I. Held, R. Jones, R. K. Kolli, W.-T. Kwon, R. Laprise, V. Magaña Rueda, L. Mearns, C. G. Menéndez, J. Räisänen, A. Rinke, A. Sarr, and P. Whetton, 2007: Regional climate projections. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press, 847-845.

Collins, M., B. B. B. Booth, G. Harris, J. M. Murphy, D. M. H. Sexton, and M. J. Webb, 2006: Towards quantifying uncertainty in transient climate change. *Clim. Dyn.*, **27,** 127-147.

Crucifix, M., 2006: Does the Last Glacial Maximum constrain climate sensitivity? *Geophys. Res. Lett.*, **33,** L18701, doi:10.1029/2006GL027137.

Dessai, S., M. Hulme, R. Lempert, and R. A. Pielke JR, 2009: Climate prediction: a limit to adaptation? *Adapting to climate change: Thresholds, values, governance*, W. N. Adger, Lorenzoni, I., and K. L. O'Brien, Eds., Cambridge University Press, 64-78.

Doblas-Reyes, F. J., V. Pavan, and D. B. Stephenson, 2003: The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation. *Clim. Dyn.*, **21,** 501-514.

Eyring, V., D. W. Waugh, G. E. Bodeker, E. Cordero, H. Akiyoshi, J. Austin, S. R. Beagley, B. A. Boville, P. Braesicke, C. Bruhl, N. Butchart, M. P. Chipperfield, M. Dameris, R. Deckert, M. Deushi, S. M. Frith, R. R. Garcia, A. Gettelman, M. A. Giorgetta, D. E. Kinnison, E. Mancini, E. Manzini, D. R. Marsh, S. Matthes, T. Nagashima, P. A. Newman, J. E. Nielsen, S. Pawson, G. Pitari, D. A. Plummer, E. Rozanov, M. Schraner, J. F. Scinocca, K. Semeniuk, T. G. Shepherd, K. Shibata, B. Steil, R. S. Stolarski, W. Tian, and M. Yoshiki, 2007: Multimodel projections of stratospheric ozone in the 21st century. *J. Geophys. Res.-Atmos.*, **112,** D16303, DOI:10.1029/2006JD008332.

Forest, C. E., P. H. Stone, A. P. Sokolov, M. R. Allen, and M. D. Webster, 2002: Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science*, **295,** 113-117.

Fraedrich, K. and L. M. Leslie, 1987: Combining predictive schemes in short-term forecasting. *Mon. Wea. Rev.*, **115,** 1640-1644.

Frame, D. J., D. A. Stone, P. A. Stott, and M. R. Allen, 2006: Alternatives to stabilization scenarios. *Geophys. Res. Lett.*, **33,** doi:10.1029/2006GL025801.

Frame, D. J., B. B. B. Booth, J. A. Kettleborough, D. A. Stainforth, J. M. Gregory, M. Collins, and M. R. Allen, 2005: Constraining climate forecasts: The role of prior assumptions. *Geophys. Res. Lett.*, **32,** L09702.
Furrer, R., S. R. Sain, D. Nychka, and G. A. Meehl, 2007a: Multivariate Bayesian analysis of atmosphere-ocean general circulation models. *Env. Ecol. Stat.*, **14,** 249-266.
Furrer, R., R. Knutti, S. R. Sain, D. W. Nychka, and G. A. Meehl, 2007b: Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophys. Res. Lett.*, **34,** DOI: 10.1029/2006GL027754.
Gillett, N. P., R. J. Allan, and T. J. Ansell, 2005: Detection of external influence on sea level pressure with a multi-model ensemble. *Geophys. Res. Lett.*, **32,** L19714, DOI:10.1029/2005GL023640.
Giorgi, F. and R. Francisco, 2000: Uncertainties in the Prediction of Regional Climate Change. *Global Change and Protected Areas*, G. e. a. Visconti, Ed., 127-139.
Giorgi, F. and L. O. Mearns, 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.
——, 2003: Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophys. Res. Lett.*, **30,** 1629, DOI:10.1029/2003GL017130.
Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.-Atmos.*, **113,** doi:10.1029/2007JD008972.
Greene, A. M., L. Goddard, and U. Lall, 2006: Probabilistic multimodel regional temperature change projections. *J. Clim.*, **19,** 4326-4346.
Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept. *Tellus*, **57A,** 219-233.
Hall, A. and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett.*, **33,** L03502, doi:10.1029/2005GL025127.
Hargreaves, J. C., A. Abe-Ouchi, and J. D. Annan, 2007: Linking glacial and future climates through an ensemble of GCM simulations. *Clim Past*, **3,** 77-87.
Harris, G., D. M. H. Sexton, B. B. B. Booth, M. Collins, J. M. Murphy, and M. J. Webb, 2006: Frequency distributions of transient regional climate change from perturbed physics ensembles of general circulation model simulations. *Clim. Dyn.*, **27,** 357-375.
Hegerl, G. C., T. J. Crowley, W. T. Hyde, and D. J. Frame, 2006: Climate sensitivity constrained by temperature reconstructions over the past seven centuries. *Nature*, **440,** 1029-1032.
Hegerl, G. C., F. W. Zwiers, P. Braconnot, N. P. Gillett, C. Luo, J. A. Marengo Orsini, N. Nicholls, J. E. Penner, and P. A. Stott, 2007: Understanding and attributing climate change. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Quin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press, 663-745.
Held, I. M., 2005: The gap between simulation and understanding in climate modeling. *Bull. Am. Meteorol. Soc.*, **80,** 1609–1614, DOI: 10.1175/BAMS-86-11-1609
Huber, M., R. Knutti, I. Mahlstein, and M. Wild, 2009: Constraints on climate sensitivity from radiation patterns in climate models. *J. Climate,* submitted.
IPCC, 2007a: Summary for Policymakers. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press.
——, 2007b: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.
Jun, M., R. Knutti, and D. W. Nychka, 2008a: Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *Journal of the American Statistical Association Applications and Case Studies*, **103,** 934-947, DOI 10.1198/016214507000001265.
——, 2008b: Local eigenvalue analysis of CMIP3 climate model errors. *Tellus*, **60A,** 992-1000.
Knutti, R., 2008a: Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society A*, **366,** 4647-4664, doi:10.1098/rsta.2008.0169.

——, 2008b: Why are climate models reproducing the observed global surface warming so well? *Geophys. Res. Lett.*, **35,** L18704, doi:10.1029/2008GL034932.
Knutti, R. and G. C. Hegerl, 2008: The equilibrium sensitivity of the Earth's temperature to radiation changes. *Nature Geoscience*, **1,** 735-743.
Knutti, R., T. F. Stocker, F. Joos, and G.-K. Plattner, 2002: Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature*, **416,** 719-723.
——, 2003: Probabilistic climate change projections using neural networks. *Clim. Dyn.*, **21,** 257-272.
Knutti, R., G. A. Meehl, M. R. Allen, and D. A. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Clim.*, **19,** 4224-4233.
Knutti, R., M. R. Allen, P. Friedlingstein, J. M. Gregory, G. C. Hegerl, G. A. Meehl, M. Meinshausen, J. M. Murphy, G. K. Plattner, S. C. B. Raper, T. F. Stocker, P. A. Stott, H. Teng, and T. M. L. Wigley, 2008: A review of uncertainties in global temperature projections over the twenty-first century. *J. Clim.*, **21,** 2651-2663.
Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285,** 1548-1550.
Lambert, S. J. and G. J. Boer, 2001: CMIP1 evaluation and intercomparison of coupled climate models. *Clim. Dyn.*, **17,** 83-106.
Lempert, R. J. and M. E. Schlesinger, 2000: Robust strategies for abating climate change - An editorial essay. *Clim. Change*, **45,** 387-401.
Liu, Z., B. L. Otto-Bliesner, F. He, E. C. Brady, R. Tomas, P. U. Clark, A. E. Carlson, J. Lynch-Stieglitz, W. Curry, E. Brook, D. Erickson, R. Jacob, J. Kutzbach, and J. Cheng, 2009: Transient Simulation of Last Deglaciation with a New Mechanism for Bolling-Allerod Warming. *Science*, **325,** 310-314.
Lopez, A., C. Tebaldi, M. New, D. A. Stainforth, M. R. Allen, and J. A. Kettleborough, 2006: Two approaches to quantifying uncertainty in global temperature changes. *J. Clim.*, **19,** 4785-4796.
Maxino, C. C., B. J. McAvaney, A. J. Pitman, and S. E. Perkins, 2008: Ranking the AR4 climate models over the Murray-Darling Basin using simulated maximum temperature, minimum temperature and precipitation. *Int. J. Clim.*, **28,** 1097-1112.
Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor, 2007a: The WCRP CMIP3 multimodel dataset - A new era in climate change research. *Bull. Am. Met. Soc.*, **88,** 1383-1394.
Meehl, G. A., T. F. Stocker, W. D. Collins, P. Friedlingstein, A. T. Gaye, J. M. Gregory, A. Kitoh, R. Knutti, J. M. Murphy, A. Noda, S. C. B. Raper, I. G. Watterson, A. J. Weaver, and Z.-C. Zhao, 2007b: Global climate projections. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press, 747-845.
Meinshausen, M., S. C. B. Raper, and T. M. L. Wigley, 2009: Emulating IPCC AR4 atmosphere-ocean and carbon cycle models for projecting global-mean, hemispheric and land/ocean temperatures: MAGICC 6.0. *Atmospheric Chemistry and Physics Discussions*, **8,** 6153-6272.
Min, S. K. and A. Hense, 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,** doi:10.1029/2006GL025779.
Murphy, J. M., B. B. B. Booth, M. Collins, G. R. Harris, D. M. H. Sexton, and M. J. Webb, 2007: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philos T R Soc A*, **365,** 1993-2028.
Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **429,** 768-772.
Nakicenovic, N. and R. Swart, 2000: *IPCC Special Report on Emissions Scenarios*. Cambridge University Press.
Oreskes, N., K. Shraderfrechette, and K. Belitz, 1994: Verification, Validation, and Confirmation of Numerical-Models in the Earth-Sciences. *Science*, **263,** 641-646.

Palmer, T. N., F. J. Doblas-Reyes, R. Hagedorn, and A. Weisheimer, 2005: Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philos T Roy Soc B*, **360,** 1991-1998.
Palmer, T. N., F. J. Doblas-Reyes, A. Weisheimer, G. J. Shutts, J. Berner, and J. M. Murphy, 2009: Towards the probabilistic Earth system model. *J. Climate***,** sumbitted.
Parker, W., 2006: Understanding model pluralism in climate science. *Foundations of Science*.
Peña, M. and H. Van den Dool, 2008: Consolidation of multimodel forecasts by ridge regression: application to Pacific sea surface temperature. *J. Climate*, **21,** 6521-6538, DOI: 10.1175/2008JCLI2226.1.
Perkins, S. E., A. J. Pitman, and S. A. Sisson, 2009: Smaller projected increases in 20-year temperature returns over Australia in skill-selected climate models. *Geophys. Res. Lett.*, **36,** -.
Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney, 2007: Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *J. Clim.*, **20,** 4356-4376.
Phillips, T. J. and P. J. Gleckler, 2006: Evaluation of continental precipitation in 20th century climate simulations: The utility of multimodel statistics. *Water Resour. Res.*, **42,** W03202, doi:10.1029/2005WR004313.
Piani, C., D. J. Frame, D. A. Stainforth, and M. R. Allen, 2005: Constraints on climate change from a multi-thousand member ensemble of simulations. *Geophys. Res. Lett.*, **32,** L23825.
Pierce, D. W., T. P. Barnett, B. D. Santer, and P. J. Gleckler, 2009: Selecting global climate models for regional climate change studies. *Proc. Natl. Acad. Sci. USA*, **106,** 8441-8446.
Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Glecker, 2008: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *J. Geophys. Res.-Atmos.*, **113,** D14209, doi:10.1029/2007JD009334.
Pitman, A. J. and S. E. Perkins, 2008: Regional Projections of Future Seasonal and Annual Changes in Rainfall and Temperature over Australia Based on Skill-Selected AR(4) Models. *Earth Interactions*, **12,** DOI: 10.1175/2008EI260.1.
Plattner, G.-K., R. Knutti, F. Joos, T. F. Stocker, W. Von Bloh, V. Brovkin, D. Cameron, E. Driesschaert, S. Dutkiewicz, M. Eby, N. R. Edwards, T. Fichefet, J. C. Hargreaves, C. D. Jones, M. F. Loutre, H. D. Matthews, A. Mouchet, S. A. Müller, S. Nawrath, A. Price, A. Sokolov, K. M. Strassmann, and A. J. Weaver, 2008: Long-term climate commitments projected with climate-carbon cycle models. *J. Clim.*, **21,** 2721- 2751, DOI: 10.1175/2007JCLI1905.1.
Räisänen, J., 2007: How reliable are climate models? *Tellus Series a-Dynamic Meteorology and Oceanography*, **59,** 2-29.
Randall, D. A., R. A. Wood, S. Bony, R. Colman, T. Fichefet, J. Fyfe, V. Kattsov, A. Pitman, J. Shukla, J. Srinivasan, R. J. Stouffer, A. Sumi, and K. Taylor, 2007: Climate models and their evaluation. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press, 589-662.
Reichler, T. and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Am. Met. Soc.*, **89,** 303-311.
Robertson, A. W., S. Kirshner, and P. Smyth, 2004: Downscaling of daily rainfall occurrence over northeast Brazil using a hidden Markov model. *J. Clim.*, **17,** 4407-4424.
Sanderson, B. M., R. Knutti, T. Aina, C. Christensen, N. Faull, D. J. Frame, W. J. Ingram, C. Piani, D. A. Stainforth, D. A. Stone, and M. R. Allen, 2008: Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. *J. Clim.*, **21,** 2384-2400.
Santer, B. D., K. E. Taylor, P. J. Gleckler, C. Bonfils, T. P. Barnett, D. W. Pierce, T. M. L. Wigley, C. Mears, F. J. Wentz, W. Bruggemann, N. P. Gillett, S. A. Klein, S. Solomon, P. A. Stott, and M. F. Wehner, 2009: Incorporating model quality information in climate change detection and attribution studies. *Proc. Natl. Acad. Sci. USA*, **106,** 14778-14783.
Santer, B. D., T. M. L. Wigley, C. Mears, F. J. Wentz, S. A. Klein, D. J. Seidel, K. E. Taylor, P. W. Thorne, M. F. Wehner, P. J. Gleckler, J. S. Boyle, W. D. Collins, K. W. Dixon, C. Doutriaux, M. Free, Q. Fu, J. E. Hansen, G. S. Jones, R. Ruedy, T. R. Karl, J. R. Lanzante, G. A. Meehl, V. Ramaswamy,

G. Russell, and G. A. Schmidt, 2005: Amplification of surface temperature trends and variability in the tropical atmosphere. *Science*, **309,** 1551-1556.
Schmittner, A., M. Latif, and B. Schneider, 2005: Model projections of the North Atlantic thermohaline circulation for the 21st century assessed by observations. *Geophys. Res. Lett.*, **32,** L23710.
Smith, I. and E. Chandler, 2009: Refining rainfall projections for the Murray Darling Basin of south-east Australia - the effect of sampling model results based on performance. *Clim. Change*.
Smith, L. A., 2002: What might we learn from climate forecasts? *Proc. Natl. Acad. Sci. USA*, **99,** 2487-2492.
Smith, R. L., C. Tebaldi, D. W. Nychka, and L. O. Mearns, 2009: Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association Applications and Case Studies*, **104,** 97-116, DOI 10.1198/jasa.2009.0007.
Soden, B. J. and I. M. Held, 2006: An assessment of climate feedbacks in coupled ocean-atmosphere models. *J. Clim.*, **19,** 3354-3360.
Solomon, S., G. K. Plattner, R. Knutti, and P. Friedlingstein, 2008: Irreversible climate change due to carbon dioxide emissions. *Proc. Natl. Acad. Sci. USA*, **106,** 1704-1709.
Stainforth, D. A., M. R. Allen, E. R. Tredger, and L. A. Smith, 2007: Confidence, uncertainty and decision-support relevance in climate predictions. *Philos T R Soc A*, **365,** 2145-2161.
Stainforth, D. A., T. Aina, C. Christensen, M. Collins, N. Faull, D. J. Frame, J. A. Kettleborough, S. Knight, A. Martin, J. M. Murphy, C. Piani, D. Sexton, L. A. Smith, R. A. Spicer, A. J. Thorpe, and M. R. Allen, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433,** 403-406.
Stott, P. A. and J. A. Kettleborough, 2002: Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature*, **416,** 723-726.
Stott, P. A., J. F. B. Mitchell, M. R. Allen, T. L. Delworth, J. M. Gregory, G. A. Meehl, and B. D. Santer, 2006: Observational constraints on past attributable warming and predictions of future global warming. *J. Clim.*, **19,** 3055-3069.
Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2008: A summary of the CMIP5 experiment design, https://cmip.llnl.gov/cmip5/docs/Taylor_CMIP5_dec31.pdf
Tebaldi, C. and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Phil. Trans. Royal Soc. A*, **365,** 2053–2075.
Tebaldi, C. and B. Sanso, 2009: Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach. *Journal of the Royal Statistical Society Series a-Statistics in Society*, **172,** 83-106.
Tebaldi, C., L. O. Mearns, D. Nychka, and R. L. Smith, 2004: Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophys. Res. Lett.*, **31,** L24213.
Tebaldi, C., R. W. Smith, D. Nychka, and L. O. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multi-model ensembles. *J. Clim.*, **18,** 1524-1540.
Thomson, M. C., F. J. Doblas-Reyes, S. J. Mason, R. Hagedorn, S. J. Connor, T. Phindela, A. P. Morse, and T. N. Palmer, 2006: Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. *Nature*, **439,** 576-579.
van Oldenborgh, G. J., S. Y. Philip, and M. Collins, 2005: El Niño in a changing climate: a multi-model study. *Oc. Sci.*, **1,** 81-95.
Watterson, I. G., 2008: Calculation of probability density functions for temperature and precipitation change under global warming. *J. Geophys. Res.-Atmos.*, **113,** D12106, doi:10.1029/2007JD009254.
Waugh, D. W. and V. Eyring, 2008: Quantitative performance metrics for stratospheric-resolving chemistry-climate models. *Atm. Chem. Phys.*, **8,** 5699-5713.
Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Met. Soc.*, **134,** 241-260.
Whetton, P., I. Macadam, J. Bathols, and J. O'Grady, 2007: Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models. *Geophys. Res. Lett.*, **34,** L14701, doi:10.1029/2007GL030025.
Yun, W. T., L. Stefanova, and T. N. Krishnamurti, 2003: Improvement of the multimodel superensemble technique for seasonal forecasts. *J. Clim.*, **16,** 3834-3840.

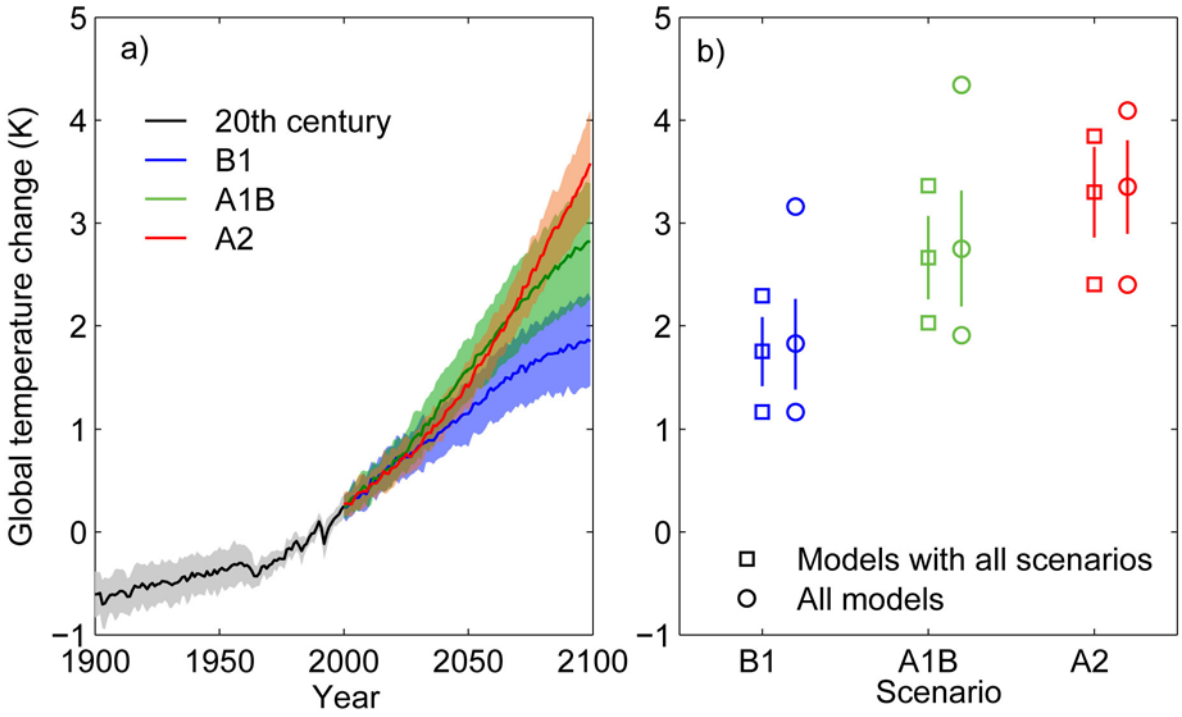


Figure 1: a) Multi-model mean and one standard deviation uncertainty ranges for global temperature (relative to the 1980-1999 average for each model) for the historic simulation and projections for three IPCC SRES scenarios. b) Mean and one standard deviation ranges (lines) plus minimum maximum ranges (symbols) for the subset of models that have run all three scenarios (squares) and for all models (circles). The model spread for the scenarios B1 and A1B depends strongly on what prior distribution of models is assumed.

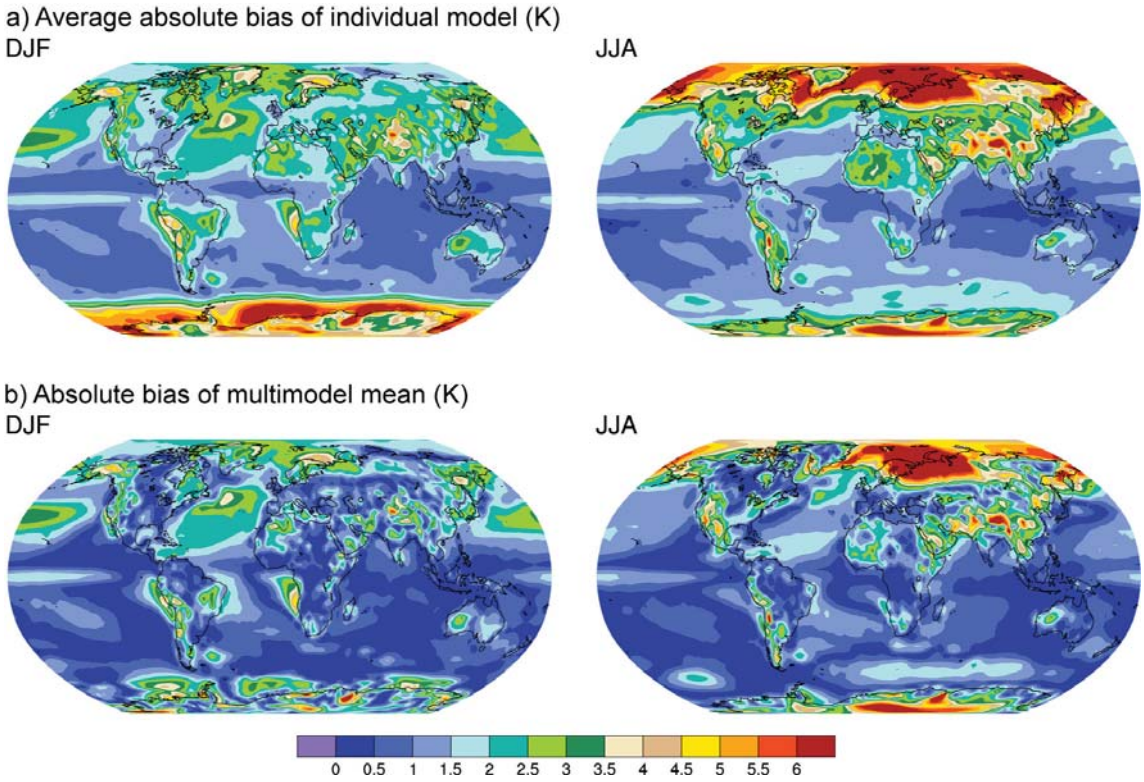


Figure 2: a) Absolute bias in 1970-1999 average surface temperature from ERA40, averaged across all CMIP3 models for December to February (DJF) and June to August (JJA). b) Same as in a) but bias shown for the multi-model average. In some locations, biases from observations are reduced but the improvement by averaging is very heterogeneous.

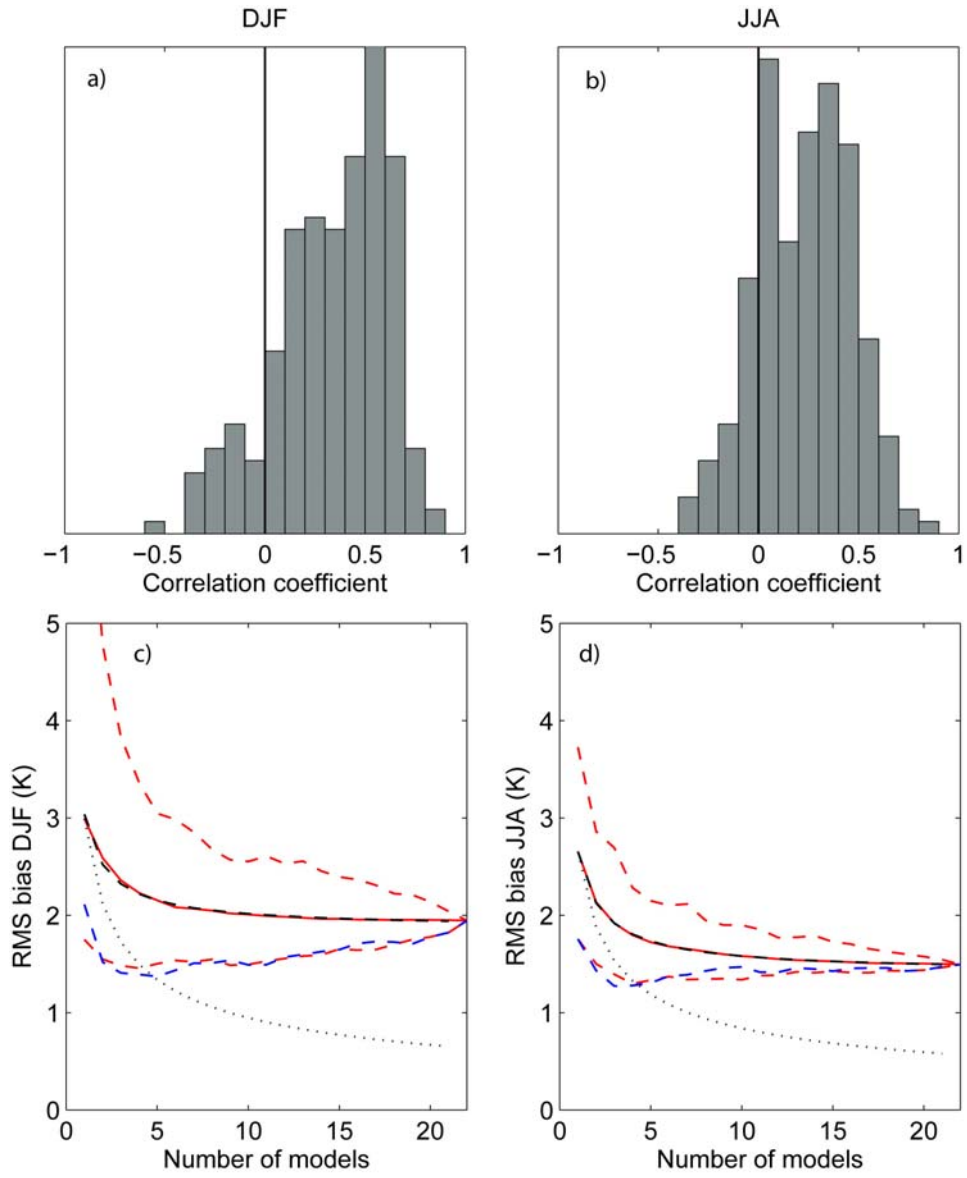


Figure 3: a/b) Histogram of correlation coefficients for all possible pairs of 1980-1999 surface temperature bias maps of the CMIP3 models for December to February (DJF) and June to August (JJA). For independent bias patterns, correlations should be distributed around zero on average. Positive correlations indicate that biases have similar patterns. c/d) Root mean square error of 1980-1999 surface temperature (averaged over space, relative to ERA40) shown as a function of the number of models included in the model average. Red dashed indicates the range covered by randomly sampling the models for the subset, red solid indicates the average. The RMS error converges to a constant value that is more than half of the initial value for one model. The black dashed line is the theoretical RMS based on the correlation structure similar to a/b) and is given by $\sigma\sqrt{[(1+(N-1)\,\rho)/N]}$ with $\sigma$ the variance and $\rho$ the average correlation. In this case $\sigma$ and $\rho$ were chosen to fit the red solid line. If the model biases were independent, the RMS error should decrease with the square root of the number of models (dotted). The blue line results if the models are sorted by how well they agree with DJF and JJA observations combined, and indicate that the average of a few good models outperforms an average of more models with poorer performance.

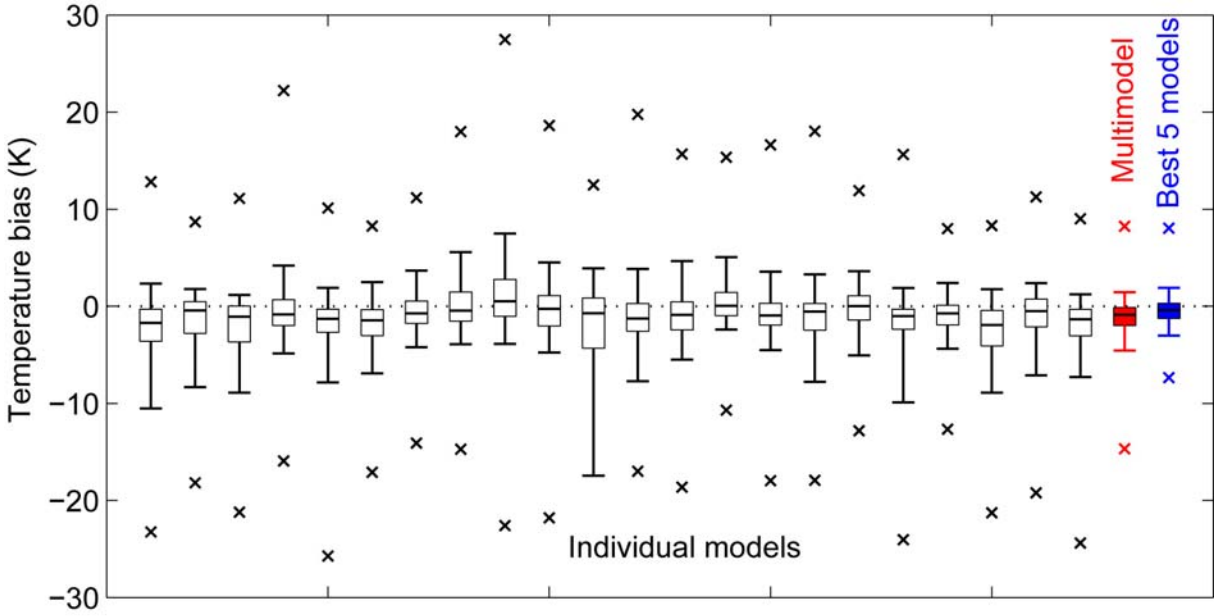


Figure 4: Box plots of surface temperature biases from ERA40 for all models and grid boxes, along with the average of all models (red) and the average of the five models that show the smallest RMS temperature biases (blue). The box marks the median and interquartile range, the line marks the 5 to 95% range, symbols mark the minimum to maximum range. One minimum value is out of range and not shown. While averaging reduces the biases, the tails of the distributions shrink proportionally less than the central part of the distributions.

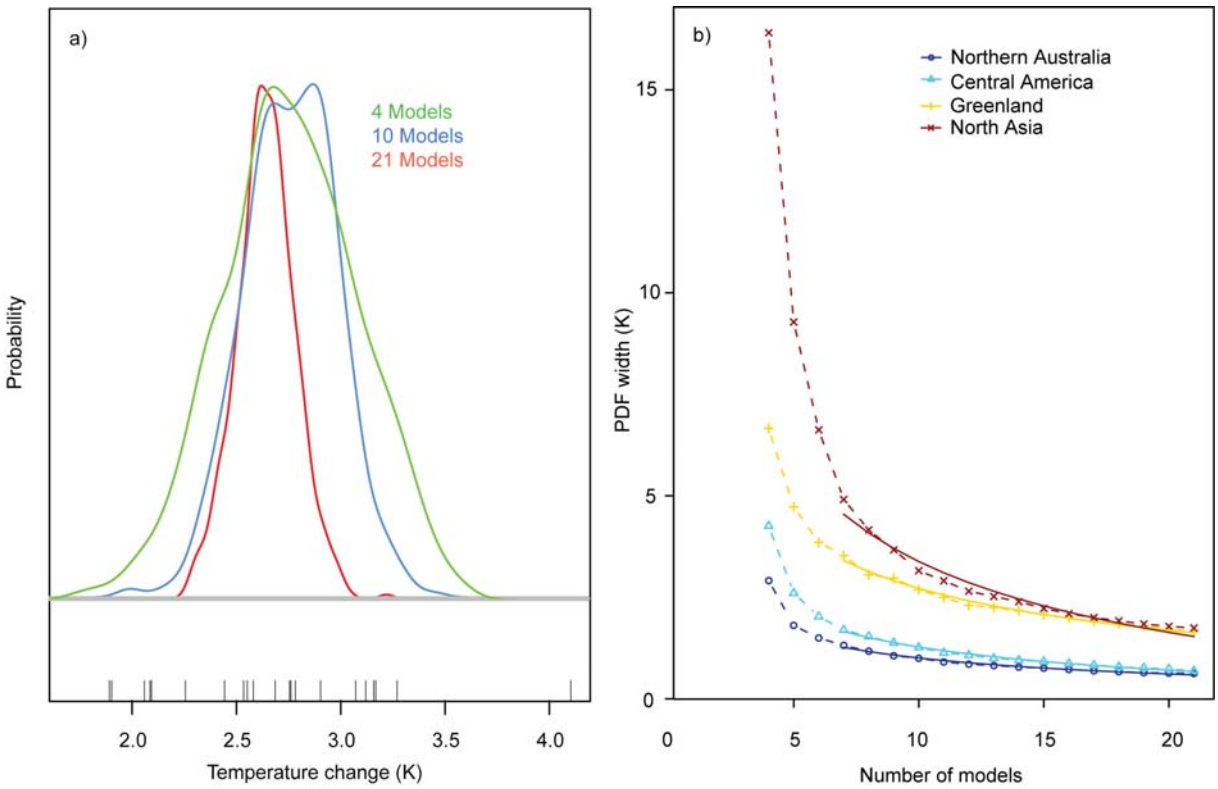


Figure 5: a) Probability density functions for annual global temperature change 2080-2099 relative to 1980-1999 from the Bayesian method by Furrer et al. (2007b; 2007a), for the A1B scenario and for 4, 10 and 21 models. b) Width of probability density function (2.5 - 97.5%) of temperature change in different regions (December to February, A1B, 2081-2100 vs. 1981-2000) as a function of number of models included, based on the method by Smith et al. (2009), the most recent version of the method originally proposed by Tebaldi et al. (2005). The analysis was repeated many times with different subsets of models and the results then averaged. A fit (solid) of the form $1/\sqrt{(N)}$ where $N$ is the number of models is given for illustration for $N>6$. Because the models are assumed to be independent, the uncertainty of the projections is reduced for a larger number of models in all cases shown.

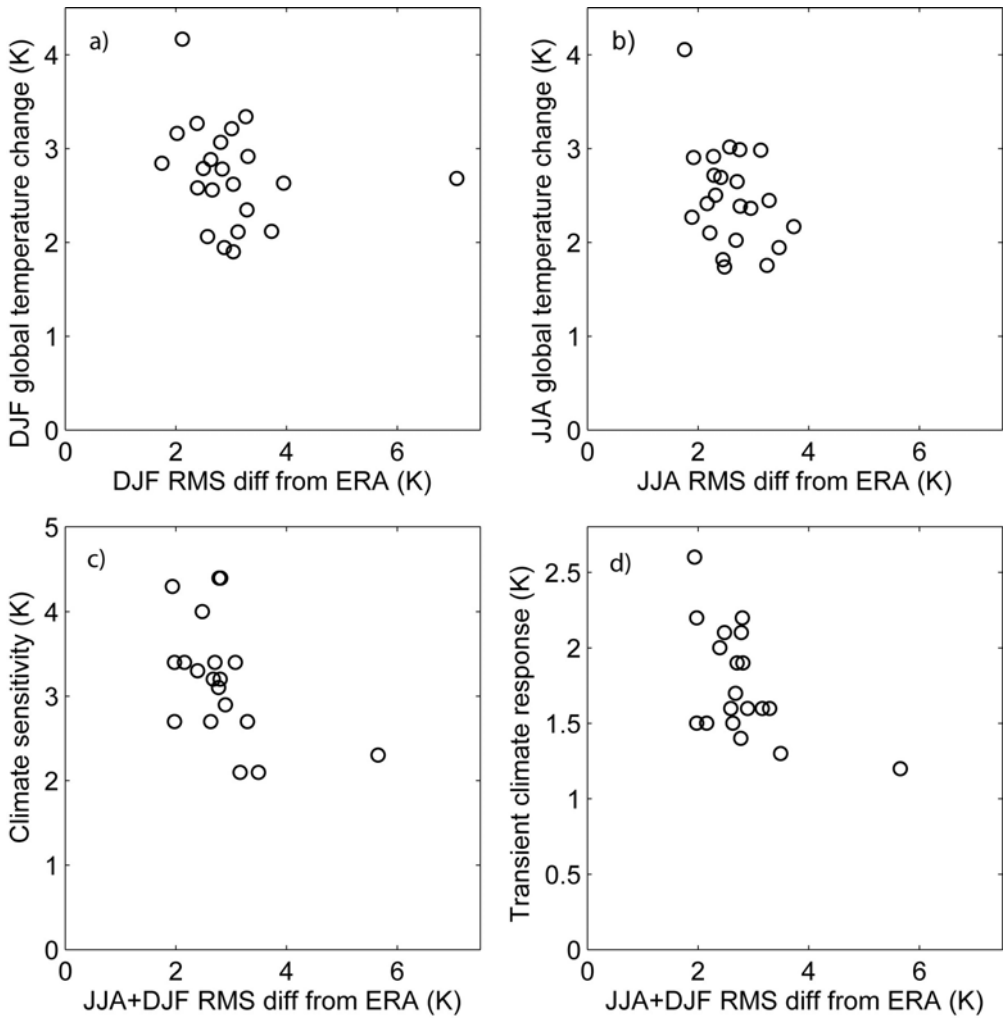


Figure 6: Scatter plots of root mean square error from ERA surface temperature for different seasons (DJF: December to February, JJA: June to August) versus predicted DJF warming (panel a, difference of 2080-2099 from 1980-1999 in the A1B scenario), JJA warming (b, same period and scenario), climate sensitivity (equilibrium global surface warming for $2xCO_2$, panel c) and transient climate response (global surface warming at the time of $CO_2$ doubling in a 1%/yr $CO_2$ increase scenario, panel d). Each circle marks one model. Correlations are near zero in all cases, indicating that the climatology of surface temperature is weakly related to the predicted warming.

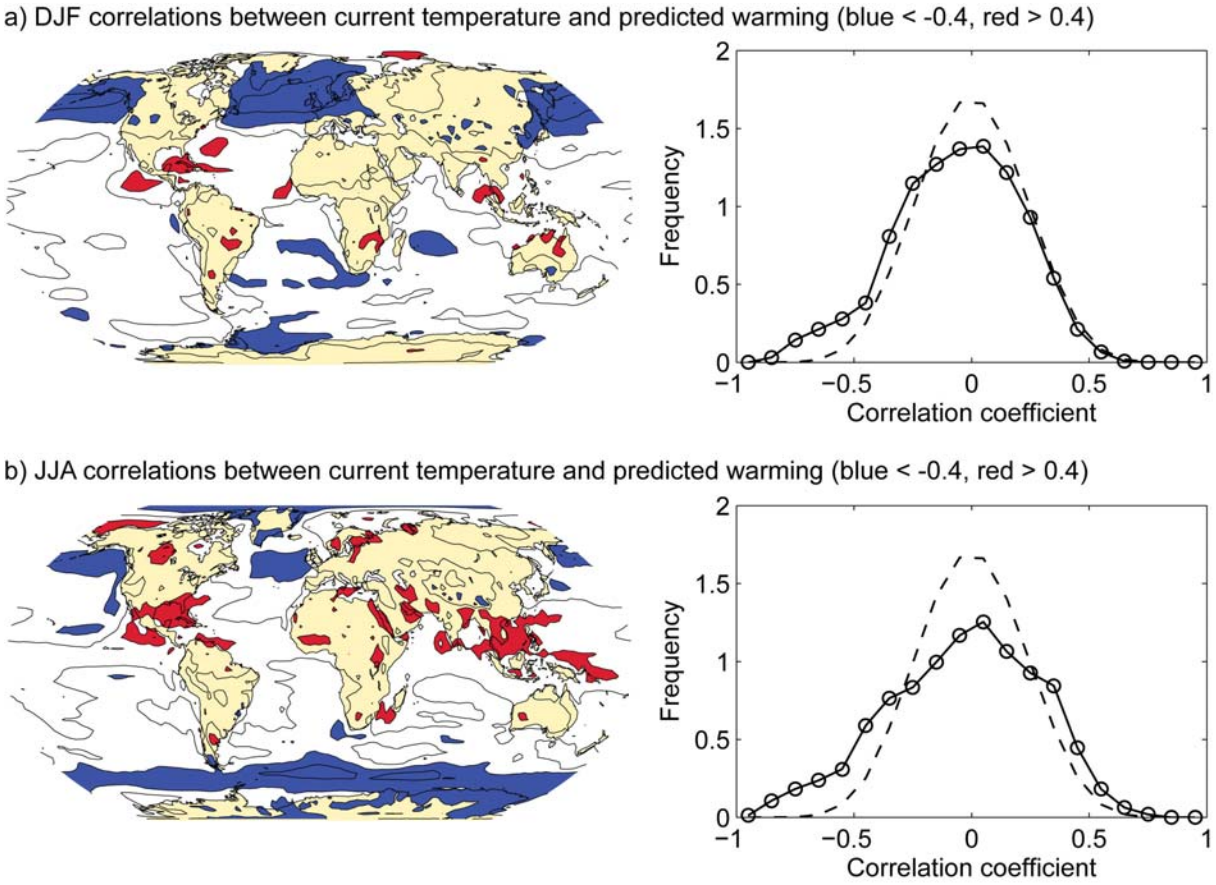


Figure 7: Correlations between mean temperature 1980-1999 and predicted warming 2080-2099 in the A1B scenario at each grid point, for December to February (DJF) and June to August (JJA). Contour intervals are 0.2, correlations smaller than -0.4 and larger than +0.4 are shown in blue and red, respectively. Panels on the right show the distribution of correlation values for the CMIP3 models (circles) and the distribution that would result from normally distributed random numbers. Most correlations are insignificant and are expected to appear by chance because the number of models is small.

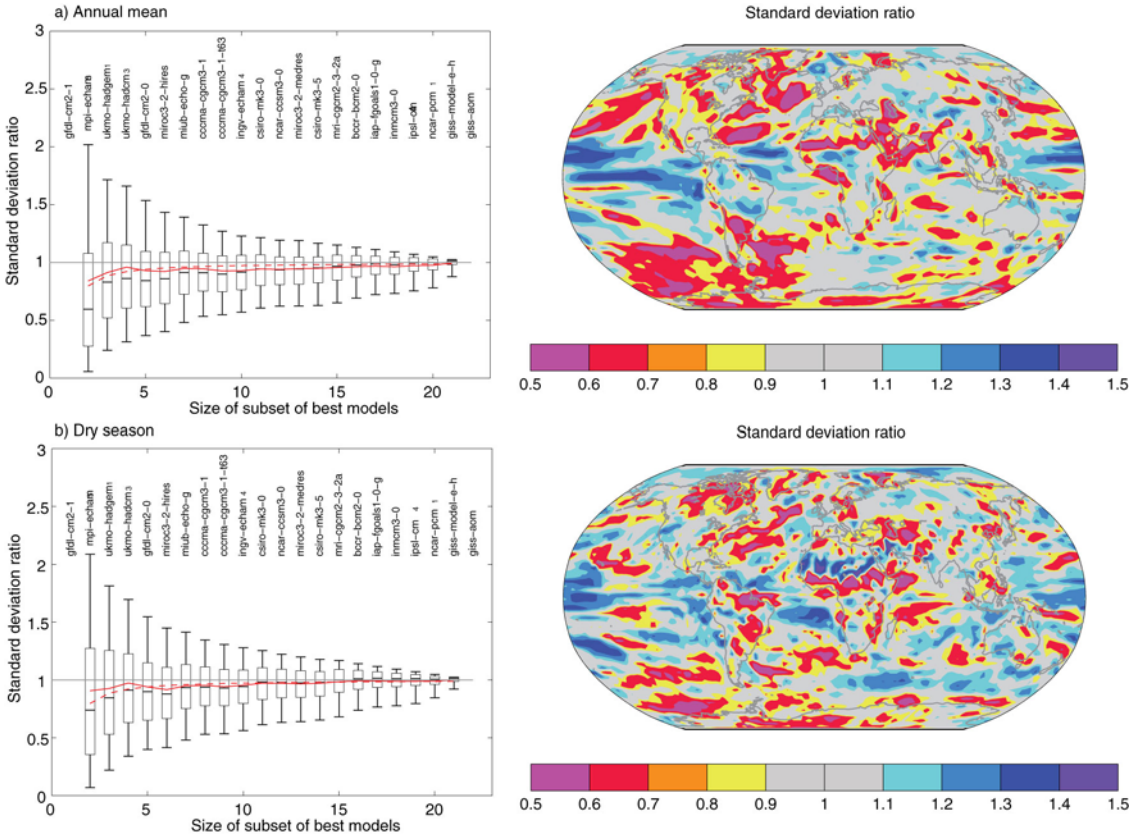


Figure 8: Ratio *R* of the standard deviation of a subset of *N* 'good' models (for present day climatology) to the standard deviation of all models, for different sizes of the subset and annual mean precipitation trend (panel a, in percent per degree global temperature change over the period 1900-2100) and the dry season (panel b, three driest consecutive months at each grid point). For example, for *N*=3, *R* is the standard deviation of gfdl-cm2-1, mpi-echam5 and ukmo-hadgem1, divided by the standard deviation of all models. *R* is calculated at every grid point and summarized in the bar plot. Ratios *R* less than unity indicate that the subset of good models has a smaller spread in the predicted rainfall trend. Panels on the left show the box plots of *R* for all grid points and different subsets of models. The box marks the median and interquartile range, the line marks the 5 to 95% range. The model names of the ranking used (updated from Reichler and Kim 2008) is given in the figure. The red solid line indicates the mean of *R* over all grid points. The red dashed line is what would result from normal random numbers, because the standard deviation is negatively biased for small samples (see text). Panels on the right show the ratio *R* for the eleven best vs. all models. While the spread for a subset of good models decreases in some places, the effect is small and there is little benefit of selecting a subset.

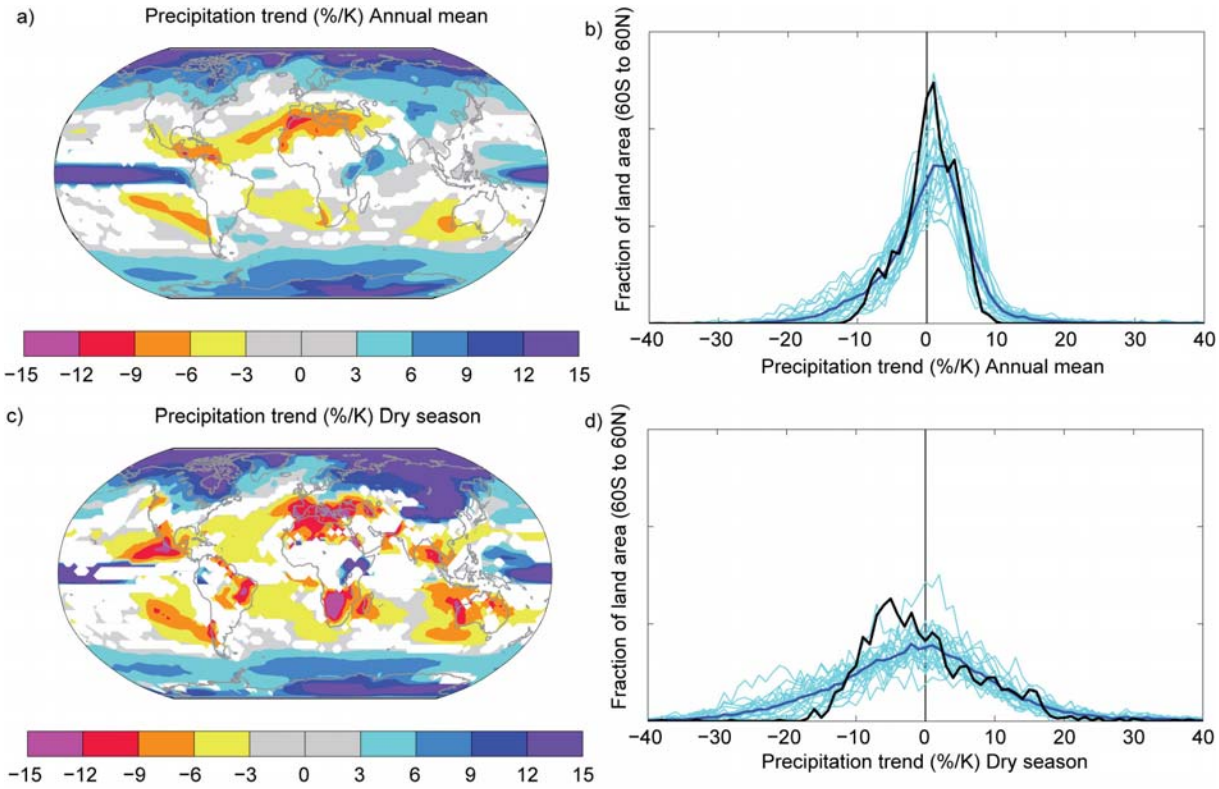


Figure 9: a/c) Precipitation trend (in percent per degree global temperature change over the period 1900-2100, relative to the base period 1900-1950) for the annual mean (a) and the dry season (three driest consecutive months at each grid point, panel c). White is used where fewer than 16 of 22 models agree on the sign of the change (see Solomon et al. (2008) for details). b/d) Distribution of the fraction of land area between 60°S and 60°N that shows a certain drying or wettening. Light blue lines indicate each CMIP3 model, the average of the light blue lines is given in dark blue. The black distribution is the result for the multi-model mean shown in the left panels. The expected precipitation change in the multimodel mean is about 30% smaller than in any single model.