IPCC Expert Meeting on
Assessing and Combining Multi Model Climate Projections

National Center for Atmospheric Research
Boulder, Colorado, USA
25-27 January 2010

Good Practice Guidance Paper on
Assessing and Combining Multi Model Climate Projections

Core Writing Team:
Reto Knutti (Switzerland), Gabriel Abramowitz (Australia), Matthew Collins (United Kingdom),
Veronika Eyring (Germany), Peter J. Gleckler (USA), Bruce Hewitson (South Africa), Linda Mearns (USA)

Edited by:
Thomas Stocker, Qin Dahe, Gian-Kasper Plattner,
Melinda Tignor, Pauline Midgley

Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections

Core Writing Team: Reto Knutti (Switzerland), Gabriel Abramowitz (Australia), Matthew Collins (United Kingdom), Veronika Eyring (Germany), Peter J. Gleckler (USA), Bruce Hewitson (South Africa), Linda Mearns (USA)

Executive Summary

Climate model simulations provide a cornerstone for climate change assessments. This paper summarizes the discussions and conclusions of the Intergovernmental Panel on Climate Change (IPCC) Expert Meeting on Assessing and Combining Multi Model Climate Projections, which was held in Boulder, USA on 25-27 January 2010. It seeks to briefly summarize methods used in assessing the quality and reliability of climate model simulations and in combining results from multiple models. It is intended as a guide for future IPCC Lead Authors as well as scientists using results from model intercomparison projects. This paper provides recommendations for good practice in using multi-model ensembles for detection and attribution, model evaluation and global climate projections as well as regional projections relevant for impact and adaptation studies. It illustrates the potential for, and limitations of, combining multiple models for selected applications. Criteria for decision making concerning model quality and performance metrics, model weighting and averaging are recommended. This paper does not, however, provide specific recommendations regarding which performance metrics to use, since this will need to be decided for each application separately.

1. Key Terminology

Many of the definitions below reflect the broad usage of these terms in climate science. While some terms are occasionally used interchangeably, the definitions presented here attempt to provide clear distinctions between them, while still encompassing the wide range of meanings encountered by meeting participants.

Model evaluation: The process of comparing model output with observations (or another model) either qualitatively using *diagnostics* or quantitatively using *performance metrics*. During model development, it is also common to compare new models with previous versions to assess relative improvements.

Diagnostic: A quantity derived from model output, often used for comparison with observations, or intercomparison of the output from different models. Examples include spatial maps, time-series and frequency distributions. More specific examples would be the trend in global mean temperature over a certain time period, or the climate sensitivity of a model.

Performance metric: A quantitative measure of agreement between a simulated and observed quantity which can be used to assess the performance of individual models. A performance metric may target a specific process to quantify how well that process is represented in a model. The term *metric* is used in different ways in climate science, for example for metrics such as radiative forcing or global warming potential. In IPCC (2007) it is defined as a consistent measurement of a characteristic of an object or activity that is otherwise difficult to quantify. More generally, it is a synonym for 'measure' or 'standard of measurement'. It often also refers more specifically to a measure of the difference (or distance) between two models or a model and observations. A *performance metric* is a statistical measure of agreement between a simulated and observed quantity (or co-variability between quantities) which can be used to assign a quantitative measure of performance ('grade') to individual models. Generally a performance metric is a quantity derived from a *diagnostic*. A performance metric can target specific processes, i.e., measure agreement between a model simulation and observations (or possibly output from a process model such as a Large Eddy Simulation) to quantify how well a specific process is represented in a model. Constructing quantitative performance metrics for a range of observationally-based diagnostics allows visualization of several aspects of a model's performance. Synthesis of a model's perform-

ance in this way can facilitate identification of missing or inadequately modelled processes in individual models, is useful for the assessment of a generation of community-wide collections of models (in the case of systematic biases), or can be used for a quantitative assessment of model improvements (e.g., by comparing results from Phases 3 and 5 of the Coupled Model Intercomparison Project CMIP3 and CMIP5).

Model quality metric, model quality index: A measure designed to infer the skill or appropriateness of a model for a specific purpose, obtained by combining performance metrics that are considered to be important for a particular application. It defines a measure of the quality or 'goodness' of a model, given the purposes for which the model is to be used, and is based on relevant *performance metrics* including one or more variables. In combination with a formal statistical framework, such a metric can be used to define model weights in a multi-model (or perturbed-physics) context. A model quality index may take into account model construction, spatio-temporal resolution, or inclusion of certain components (e.g., carbon cycle) in an ad-hoc and possibly subjective way, e.g., to identify subsets of models.

Ensemble: A group of comparable model simulations. The ensemble can be used to gain a more accurate estimate of a model property through the provision of a larger sample size, e.g., of a climatological mean of the frequency of some rare event. Variation of the results across the ensemble members gives an estimate of uncertainty. Ensembles made with the same model but different initial conditions only characterise the uncertainty associated with internal climate variability, whereas multi-model ensembles including simulations by several models also include the impact of model differences. Nevertheless, the multi-model ensemble is not designed to sample uncertainties in a systematic way and can be considered an ensemble of opportunity. Perturbed-physics parameter ensembles are ensembles in which model parameters are varied in a systematic manner, aiming to produce a more systematic estimate of single-model uncertainty than is possible with traditional multi-model ensembles.

Multi-model mean (un-weighted): An average of simulations in a multi-model ensemble, treating all models equally. Depending on the application, if more than one realization from a given model is available (differing only in initial conditions), all realizations for a given model might be averaged together before averaging with other models.

Multi-model mean (weighted): An average across all simulations in a multi-model dataset that does not treat all models equally. Model 'weights' are generally derived from some measure of a model's ability to simulate the observed climate (i.e., a *model quality metric/index*), based on how processes are implemented or based on expert judgment. Weights may also incorporate information about model independence. In climate model projections, as in any other application, the determination of weights should be a reflection of an explicitly defined statistical model or framework.

2. Background and Methods

Climate model results provide the basis for projections of future climate change. Previous assessment reports included model evaluation but avoided weighting or ranking models. Projections and uncertainties were based mostly on a 'one model, one vote' approach, despite the fact that models differed in terms of resolution, processes included, forcings and agreement with observations. Projections in the IPCC's Fifth Assessment Report (AR5) will be based largely on CMIP5 of the World Climate Research Programme (WCRP), a collaborative process in which the research and modelling community has agreed on the type of simulations to be performed. While many different types of climate models exist, the following discussion focuses on the global dynamical models included in the CMIP project.

Uncertainties in climate modelling arise from uncertainties in initial conditions, boundary conditions (e.g., a radiative forcing scenario), observational uncertainties, uncertainties in model parameters and structural uncertainties resulting from the fact that some processes in the climate system are not fully understood or are impossible to resolve due to computational constraints. The widespread participation in CMIP provides some perspective on model uncertainty. Nevertheless, intercomparisons that facilitate systematic multi-model evaluation are not designed to yield formal error estimates, and are in essence 'ensembles of opportunity'. The spread of a multiple model ensemble is therefore rarely a direct measure of uncertainty, particularly given that models are unlikely to be independent, but the spread can help to characterize uncertainty. This involves understanding how the variation across an ensemble was generated, making assumptions about the appropriate statistical framework, and choosing appropriate model quality metrics. Such topics are only beginning to be addressed by the research community (e.g., Randall et al., 2007; Tebaldi and Knutti, 2007; Gleckler et al., 2008; Knutti, 2008;

Reichler and Kim, 2008; Waugh and Eyring, 2008; Pierce et al., 2009; Santer et al., 2009; Annan and Hargreaves, 2010; Knutti, 2010; Knutti et al., 2010).

Compared to CMIP3, the number of models and model versions may increase in CMIP5. Some groups may submit multiple models or versions of the same model with different parameter settings and with different model components included. For example, some but not all of the new models will include interactive representations of biogeochemical cycles (carbon and nitrogen), gas-phase chemistry, aerosols, ice sheets, land use, dynamic vegetation, or a full representation of the stratosphere. The new generation of models is therefore likely to be more heterogeneous than in earlier model intercomparisons, which makes a simple model average increasingly difficult to defend and to interpret. In addition, some studies may wish to make use of model output from earlier CMIP phases or other non-CMIP sources.

The reliability of projections might be improved if models are weighted according to some measure of skill and if their interdependencies are taken into account, or if only subsets of models are considered. Indeed such methods using forecast verification have been shown to be superior to simple averages in the area of weather and seasonal forecasting (Stephenson et al., 2005). Since there is little opportunity to verify climate forecasts on timescales of decades to centuries (except for a realization of the 20th century), the skill or performance of the models needs to be defined, for example, by comparing simulated patterns of present-day climate to observations. Such performance metrics are useful but not unique, and often it is unclear how they relate to the projection of interest. Defining a set of criteria for a model to be 'credible' or agreeing on a quality metric is therefore difficult. However, it should be noted that there have been de facto model selections for a long time, in that simulations from earlier model versions are largely discarded when new versions are developed. For example, results produced for the Third Assessment Report of the IPCC were not directly included in the projections chapters of the Fourth Assessment Report unless an older model was used again in CMIP3. If we indeed do not clearly know how to evaluate and select models for improving the reliability of projections, then discarding older results out of hand is a questionable practice. This may again become relevant when deciding on the use of results from the AR4 CMIP3 dataset along with CMIP5 in AR5.

Understanding results based on model ensembles requires an understanding of the method of generation of

the ensemble and the statistical framework used to interpret it. Methods of generation may include sampling of uncertain initial model states, parameter values or structural differences. Statistical frameworks in published methods using ensembles to quantify uncertainty may assume (perhaps implicitly):

a. that each ensemble member is sampled from a distribution centered around the truth ('truth plus error' view) (e.g., Tebaldi et al., 2005; Greene et al., 2006; Furrer et al., 2007; Smith et al., 2009). In this case, perfect independent models in an ensemble would be random draws from a distribution centered on observations.

Alternatively, a method may assume:

b. that each of the members is considered to be 'exchangeable' with the other members and with the real system (e.g., Murphy et al., 2007; Perkins et al., 2007; Jackson et al., 2008; Annan and Hargreaves, 2010). In this case, observations are viewed as a single random draw from an imagined distribution of the space of all possible but equally credible climate models and all possible outcomes of Earth's chaotic processes. A 'perfect' independent model in this case is also a random draw from the same distribution, and so is 'indistinguishable' from the observations in the statistical model.

With the assumption of statistical model (a), uncertainties in predictions should tend to zero as more models are included, whereas with (b), we anticipate uncertainties to converge to a value related to the size of the distribution of all outcomes (Lopez et al., 2006; Knutti et al., 2010). While both approaches are common in published literature, the relationship between the method of ensemble generation and statistical model is rarely explicitly stated.

The second main distinction in published methods is whether all models are treated equally or whether they are weighted based on their performance (see Knutti, 2010 for an overview). Recent studies have begun to explore the value of weighting the model projections based on their performance measured by process evaluation, agreement with present-day observations, past climate or observed trends, with the goal of improving the multi-model mean projection and more accurately quantifying uncertainties (Schmittner et al., 2005; Connolley and Bracegirdle, 2007; Murphy et al., 2007; Waugh and Eyring, 2008). Model quality information has also been

used in recent multi-model detection and attribution studies (Pierce et al., 2009; Santer et al., 2009). Several studies have pointed out difficulties in weighting models and in interpreting model spread in general. Formal statistical methods can be powerful tools to synthesize model results, but there is also a danger of overconfidence if the models are lacking important processes and if model error, uncertainties in observations, and the robustness of statistical assumptions are not properly assessed (Tebaldi and Knutti, 2007; Knutti et al., 2010). A robust approach to assigning weights to individual model projections of climate change has yet to be identified. Extensive research is needed to develop justifiable methods for constructing indices that can be used for weighting model projections for a particular purpose. Studies should employ formal statistical frameworks rather than using ad hoc techniques. It is expected that progress in this area will likely depend on the variable, spatial and temporal scale of interest. Finally, it should be noted that few studies have addressed the issue of structural model inadequacies, i.e., errors which are common to all general circulation models (GCMs).

User needs frequently also include assessments of regional climate information. However, there is a danger of over-interpretation or inappropriate application of climate information, such as using a single GCM grid cell to represent a point locality. There is therefore a general need for guidance of a wide community of users for multi-model GCM climate projection information plus regional climate models, downscaling procedures and other means to provide climate information for assessments. Difficulties arise because results of regional models are affected both by the driving global model as well as the regional model. There have been efforts in combining global and regional model results from past research programs (e.g., PRUDENCE) and continue in the present with ongoing GCM and Regional Climate Models (RCM) simulations programs (Mearns et al., 2009). The relationship between the driving GCM and the resulting simulation with RCMs provides interesting opportunities for new approaches to quantify uncertainties. Empirical-statistical downscaling (ESD) is computationally cheaper than RCMs, and hence more practical for downscaling large ensembles and long time intervals (Benestad, 2005) although ESD suffers from possible out-of-sample issues.

3. Recommendations

In the following, a series of recommendations towards 'best practices' in 'Assessing and Combining Multi-model

Climate Projections' agreed on by the meeting participants are provided. Most of the recommendations are based on literature and experience with GCMs but apply similarly to emerging ensembles of regional models (e.g., ENSEMBLES, NARCCAP). Some recommendations even apply to ensembles of other types of numerical models.

The participants of the IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections are not in a position to provide a 'recipe' to assess the literature and results from the CMIP3/5 simulations. Here, an attempt is made to give good practice guidelines for both scientific studies and authors of IPCC chapters. While the points are generic, their applicability will depend on the question of interest, the spatial and temporal scale of the analysis and the availability of other sources of information.

3.1 Recommendations for Ensembles

When analyzing results from multi-model ensembles, the following points should be taken into account:

- Forming and interpreting ensembles for a particular purpose requires an understanding of the variations between model simulations and model set-up (e.g., internal variability, parameter perturbations, structural differences, see Section 2), and clarity about the assumptions, e.g., about model independence,

exchangeability, and the statistical model that is being used or assumed (Box 3.1).

- The distinction between 'best effort' simulations (i.e., the results from the default version of a model submitted to a multi-model database) and perturbed physics ensembles is important and must be recognized. Perturbed physics ensembles can provide useful information about the spread of possible future climate change and can address model diversity in ways that best effort runs are unable to do. However, combining perturbed physics and best effort results from different models is not straightforward. An additional complexity arises from the fact that different model configurations may be used for different experiments (e.g., a modelling group may not use the same model version for decadal prediction experiments as it does for century scale simulations).

- In many cases it may be appropriate to consider simulations from CMIP3 and combine CMIP3 and CMIP5 recognizing differences in specifications (e.g., differences in forcing scenarios). IPCC assessments should consider the large amount of scientific work on CMIP3, in particular in cases where lack of time prevents an in depth analysis of CMIP5. It is also useful to track model improvement through different generations of models.

Box 3.1: Examples of Projections Derived Using Complex Multivariate Statistical Techniques which Express Projections as Probability Density Functions

Because of the relative paucity of simple observational constraints (Box 3.2) and because of the requirement to produce projections for multiple variables that are physically consistent within the model context, complex statistical techniques have been employed. The majority of these are based on a Bayesian approach in which prior distributions of model simulations are weighted by their ability to reproduce present day climatological variables and trends to produce posterior predictive distributions of climate variables (see Box 3.1, Figure 1). Numerous examples of such Bayesian approaches employing output from the multi-model archives are found in the literature (e.g., Giorgi and Mearns 2003; Tebaldi et al., 2005; Greene et al., 2006; Lopez et al., 2006; Furrer et al., 2007). Differences in the projected PDFs for the same climate variables produced by the different techniques indicate sensitivity to the specification and implementation of the Bayesian statistical framework which has still to be resolved (Tebaldi and Knutti, 2007).

Recent approaches have also employed perturbed physics ensembles in which perturbations are made to the parameters in a single modelling structure (e.g., Murphy et al., 2007; Murphy et al., 2009). In this case it is possible to illustrate a statistical framework to produce PDFs of future change (e.g., Rougier, 2007). Assume that we can express a climate model output, y, as a function, f, of its input parameters, x; $y = f(x) + \varepsilon$ where $y = (y_h, y_f)$ is composed of historical and future simulation variables, and ε is the error term that accounts for uncertainty in observations, from the use of emulators (see below), and from structural uncertainty as inferred from other models, then it is possible to sample the input space x by varying parameters in the model and constrain that input space

according to the likelihood of each model version computed by comparing the simulation of historical climate with that observed. Multiple observational variables may be used in the likelihood weighting and joint projections are possible as the physics of the relationships between variables (temperature and precipitation for example) are preserved through the link to the model parameter space. The implementation of such techniques is however a challenge involving techniques such as emulators which approximate the behaviour of the full climate model given a set of input parameters, as is the estimation of structural uncertainty not accounted for by parameter perturbations (Murphy et al., 2007; Murphy et al., 2009).
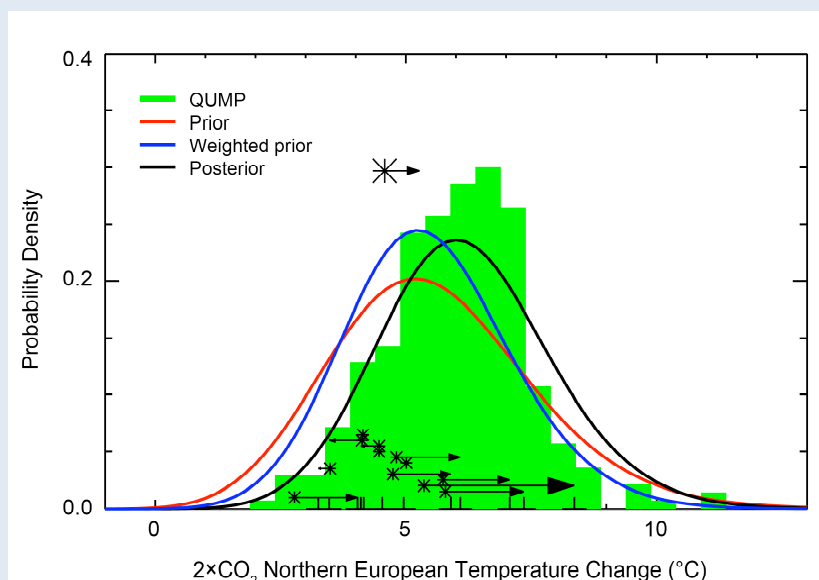
Box 3.1, Figure 1. Equilibrium probability density functions for winter surface temperature change for Northern Europe following a doubling of the atmospheric CO_2 concentration. The green histogram (labelled QUMP) is calculated from the temperature difference between 2 x CO_2 and 1 x CO_2 equilibrium simulations with 280 versions of HadSM3. The red curve (labelled prior) is obtained from a much larger sample of responses of the HadSM3 model constructed using a statistical emulator and is the prior distribution for this climate variable. The blue curve (labelled weighted prior) shows the effect of applying observational constraints to the prior distribution. The asterisks show the positions of the best emulated values of the CMIP3 multi-model members and the arrows quantify the discrepancy between these best emulations and the actual multi-model responses. These discrepancies are used to shift the HadSM3 weighted prior distribution, and also broaden the final posterior distribution (black curve). Tick marks on the x-axis indicate the response of the CMIP3 slab models used in the discrepancy calculation. From Harris et al. (2010).

- Consideration needs to be given to cases where the number of ensemble members or simulations differs between contributing models. The single model's ensemble size should not inappropriately determine the weight given to any individual model in the multi-model ensemble. In some cases ensemble members may need to be averaged first before combining different models, while in other cases only one member may be used for each model.

- Ensemble members may not represent estimates of the climate system behaviour (trajectory) entirely independent of one another. This is likely true of members that simply represent different versions of

the same model or use the same initial conditions. But even different models may share components and choices of parameterizations of processes and may have been calibrated using the same data sets. There is currently no 'best practice' approach to the characterization and combination of inter-dependent ensemble members, in fact there is no straight-forward or unique way to characterize model dependence.

3.2 Recommendations for Model Evaluation and Performance Metrics

A few studies have identified a relationship between skill in simulating certain aspects of the observed climate and

the spread of projections (see Box 3.2). If significant advancements are made in identifying such useful relationships, they may provide a pathway for attempting to quantify the uncertainty in individual processes and projections.

No general all-purpose metric (either single or multi-parameter) has been found that unambiguously identifies a 'best' model; multiple studies have shown that different metrics produce different rankings of models (e.g., Gleckler et al., 2008). A realistic representation of processes, especially those related to feedbacks in the climate system, is linked to the credibility of model projections and thus could form the basis for performance metrics designed to gauge projection reliability. The participants of the Expert Meeting recommend consideration of the following points for both scientific papers and IPCC assessments:

• Process-based performance metrics might be derived from the analysis of multi-model simulations and/or from process studies performed in projects that complement CMIP (e.g., from detailed evaluation and analysis of physical processes and feedbacks carried out in a single column modelling framework by the Cloud Feedback Model Intercomparison Project (CFMIP) or the Global Energy and Water Cycle Experiment Cloud Systems Studies (GEWEX GCSS)).

• Researchers are encouraged to consider the different standardized model performance metrics currently being developed (e.g., WCRP's Working Group on Numerical Experimentation (WGNE) / Working Group on Coupled Modelling (WGCM) metrics panel, El Niño Southern Oscillation (ENSO) metrics activity, Climate Variability and Predictability (CLIVAR) decadal variability metrics activity, the European Commission's ENSEMBLES project, Chemistry-Climate Model Validation activity (CCMVal)). These metrics should be considered for assembly in a central repository.

• A performance metric is most powerful if it is relatively simple but statistically robust, if the results are not strongly dependent on the detailed specifications of the metric and other choices external to the model (e.g., the forcing) and if the results can be understood in terms of known processes (e.g., Frame et al., 2006). There are however few instances of diagnostics and performance metrics in the literature where the large intermodel variations in the past are well correlated with comparably large intermodel variations in the model projections (Hall and Qu, 2006; Eyring et al., 2007; Boe et al., 2009) and to date a set of diagnostics and performance metrics that can strongly reduce uncertainties in global climate sensitivity has yet to be identified (see Box 3.2).

Box 3.2: Examples of Model Evaluation Through Relationships Between Present-Day Observables and Projected Future Changes

Correlations between model simulated historical trends, variability or the current mean climate state (being used frequently for model evaluation) on the one hand, and future projections for observable climate variables on the other hand, are often predominantly weak. For example, the climate response in the 21st century does not seem to depend in an obvious way on the simulated pattern of current temperature (Knutti et al., 2010). This may be partly because temperature observations are already used in the process of model calibration, but also because models simulate similar temperature patterns and changes for different reasons. While relationships across multiple models between the mean climate state and predicted changes are often weak, there is evidence in models and strong physical grounds for believing that the amplitudes of the large-scale temperature response to greenhouse gas and aerosol forcing within one model in the past represent a robust guide to their likely amplitudes in the future. Such relations are used to produce probabilistic temperature projections by relating past greenhouse gas attributable warming to warming over the next decades (Allen et al., 2000; Forest et al., 2002; Frame et al., 2006; Stott et al., 2006). The comparison of multi-model ensembles with forecast ranges from such fingerprint scaling methods, observationally-constrained forecasts based on intermediate-complexity models or comprehensively perturbed physics experiments is an important step in assessing the reliability of the ensemble spread as a measure of forecast uncertainty.

An alternative assessment of model performance is the examination of the representation of key climate feedback processes on various spatial and temporal scales (e.g., monthly, annual, decadal, centennial). There are, however,

only few instances in the literature where the large intermodel variations in the past are well correlated with comparably large intermodel variations in the model projections.

Hall and Qu (2006) used the current seasonal cycle to constrain snow albedo feedback in future climate change. They found that the large intermodel variations in the seasonal cycle of the albedo feedback are strongly correlated with comparably large intermodel variations in albedo feedback strength on climate change timescales (Box 3.2, Figure 1). Models mostly fall outside the range of the estimate derived from the observed seasonal cycle, suggesting that many models have an unrealistic snow albedo feedback. Because of the tight correlation between simulated feedback strength in the seasonal cycle and climate change, eliminating the model errors in the seasonal cycle should lead to a reduction in the spread of albedo feedback strength in climate change. A performance metric based on this diagnostic could potentially be of value to narrow the range of climate projections in a weighted multi-model mean.

Other examples include a relation between the seasonal cycle in temperature and climate sensitivity (Knutti et al., 2006) or the relation between past and future Arctic sea ice decline (Boe et al., 2009). Such relations across models are problematic if they occur by chance because the number of models is small, or if the correlation just reflects the simplicity of a parameterization common to many models rather than an intrinsic underlying process. More research of this kind is needed to fully explore the value of weighting model projections based on performance metrics showing strong relationships between present-day observables and projected future changes, or to use such relationships as transfer functions to produce projections from observations. It should be recognised however that attempts to constrain some key indicators of future change such as the climate sensitivity, have had to employ rather more complex algorithms in order to achieve strong correlations (Piani et al., 2005).
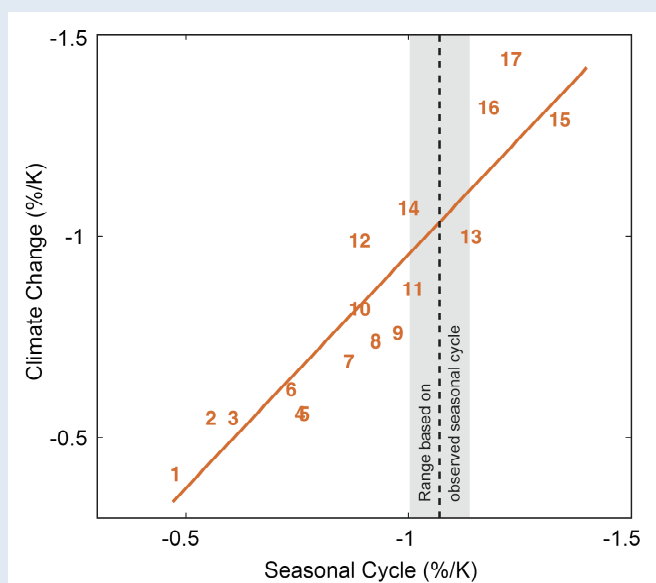
Box 3.2, Figure 1. Scatter plot of simulated ratios between changes in surface albedo, $\Delta\alpha_s$, and changes in surface air temperature, ΔT_s, during springtime, i.e., $\Delta\alpha_s/\Delta T_s$. These ratios are evaluated from transient climate change experiments with 17 AOGCMs (y-axis), and their seasonal cycle during the 20th century (x-axis). Specifically, the climate change $\Delta\alpha_s/\Delta T_s$ values are the reduction in springtime surface albedo averaged over Northern Hemisphere continents between the 20th and 22nd centuries divided by the increase in surface air temperature in the region over the same time period. Seasonal cycle $\Delta\alpha_s/\Delta T_s$ values are the difference between 20th-century mean April and May α_s averaged over Northern Hemisphere continents divided by the difference between April and May Ts averaged over the same area and time period. A least-squares fit regression line for the simulations (solid line) and the observed seasonal cycle $\Delta\alpha_s/\Delta T_s$ value based on ISCCP and ERA40 reanalysis (dashed vertical line) are also shown. From Hall and Qu (2006).

- Observational uncertainty and the effects of internal variability should be taken into account in model assessments. Uncertainties in the observations used for a metric should be sufficiently small to discriminate between models. Accounting for observational uncertainty can be done by including error estimates provided with the observational data set, or by using more than one data set to represent observations. We recognize however that many observational data sets do not supply formal error estimates and that modelers may not be best qualified for assessing observational errors.

- Scientists are encouraged to use all available methods cutting across the database of model results, i.e., they should consider evaluating models on different base states, different spatial and temporal scales and different types of simulations. Specifically, paleoclimate simulations can provide independent information for evaluating models, if the paleoclimate data has not been used in the model development process. Decadal prediction or evaluation on even shorter timescales can provide insight, but differences in model setups, scenarios and signal to noise ratios must be taken into account.

- A strong focus on specific performance metrics, in particular if they are based on single datasets, may lead to overconfidence and unjustified convergence, allow compensating errors in models to match certain benchmarks, and may prohibit sufficient diversity of models and methods crucial to characterize model spread and understand differences across models.

3.3 Recommendations for Model Selection, Averaging and Weighting

Using a variety of performance metrics, a number of studies have shown that a multi-model average often out-performs any individual model compared to observations. This has been demonstrated for mean climate (Gleckler et al., 2008; Reichler and Kim, 2008), but there are also examples for detection and attribution (Zhang et al., 2007) and statistics of variability (Pierce et al., 2009). Some systematic biases (i.e., evident in most or all models) can be readily identified in multi-model averages (Knutti et al., 2010).

There have been a number of attempts to identify more skillful vs. less skillful models with the goal to rank or weight models for climate change projections and for detection and attribution (see Section 2). The participants

of the Expert Meeting have identified the following points to be critical:

- For a given class of models and experiments appropriate to a particular study, it is important to document, as a first step, results from all models in the multi-model dataset, without ranking or weighting models.

- It is problematic to regard the behavior of a weighted model ensemble as a probability density function (PDF). The range spanned by the models, the sampling within that range and the statistical interpretation of the ensemble need to be considered (see Box 3.1).

- Weighting models in an ensemble is not an appropriate strategy for some studies. The mean of multiple models may not even be a plausible concept and may not share the characteristics that all underlying models contain. A weighted or averaged ensemble prediction may, for example, show decreased variability in the averaged variables relative to any of the contributing models if the variability minima and maxima are not collocated in time or space (Knutti et al., 2010).

- If a ranking or weighting is applied, both the quality metric and the statistical framework used to construct the ranking or weighting should be explicitly recognized. Examples of performance metrics that can be used for weighting are those that are likely to be important in determining future climate change (e.g., snow/ice albedo feedback, water vapor feedback, cloud feedback, carbon cycle feedback, ENSO, greenhouse gas attributable warming; see Box 3.2).

- Rankings or weightings could be used to select subsets of models, and to produce alternative multi-model statistics which can be compared to the original multi-model ensemble in order to assess robustness of the results with respect to assumptions in weighting. It is useful to test the statistical significance of the difference between models based on a given metric, so to avoid ranking models that are in fact statistically indistinguishable due to uncertainty in the evaluation, uncertainty whose source could be both in the model and in the observed data.

- There should be no minimum performance criteria for entry into the CMIP multi-model database.

Researchers may select a subset of models for a particular analysis but should document the reasons why.

- Testing methods in perfect model settings (i.e., one model is treated as observations with complete coverage and no observational uncertainty) is encouraged, e.g., withholding one member from a multi-model or perturbed physics ensemble, and using a given weighting strategy and the remaining ensemble members to predict the future climate simulated by the withheld model. If a weighting strategy does not perform better than an unweighted multi-model mean in a perfect-model setting, it should not be applied to the real world.

- Model agreement is not necessarily an indication of likelihood. Confidence in a result may increase if multiple models agree, in particular if the models incorporate relevant processes in different ways, or if the processes that determine the result are well understood. But some features shared by many models are a result of the models making similar assumptions and simplifications (e.g., sea surface temperature biases in coastal upwelling zones, CO_2 fertilization of the terrestrial biosphere). That is, models may not constitute wholly independent estimates. In such cases, agreement might also in part reflect a level of shared process representation or calibration on particular datasets and does not necessarily imply higher confidence.

3.4 Recommendations for Reproducibility

To ensure the reproducibility of results, the following points should be considered:

- All relevant climate model data provided by modelling groups should be made publicly available, e.g., at PCMDI or through the Earth System Grid (ESG, pointers from PCMDI website); observed datasets should also be made readily available, e.g., linked through the PCMDI website. Multi-model derived quantities (e.g., synthetic Microwave Sounding Unit (MSU) temperatures, post-processed ocean data, diagnostics of modes of variability) could be provided in a central repository.

- Algorithms need to be documented in sufficient detail to ensure reproducibility and to be available on request. Providing code is encouraged, but there was no consensus among all participants about whether to recommend providing all code to a public

repository. Arguments for providing code are full transparency of the analysis and that discrepancies and errors may be easier to identify. Arguments against making it mandatory to provide code are the fact that an independent verification of a method should redo the full analysis in order to avoid propagation of errors, and the lack of resources and infrastructure required to support such central repositories.

3.5 Recommendations for Regional Assessments

Most of the points discussed in previous sections apply also to regional and impacts studies. The participants of the meeting highlight the following recommendations for regional assessments, noting that many points apply to global projections as well. Although there is some repetition, this reflects that independent breakout groups at the Expert Meeting came up with similar recommendations:

- The following four factors should be considered in assessing the likely future climate change in a region (Christensen et al., 2007): historical change, process change (e.g. changes in the driving circulation), global climate change projected by GCMs and downscaled projected change. Particular climate projections should be assessed against the broader context of multiple sources (e.g., regional climate models, statistical downscaling) of regional information on climate change (including multi-model global simulations), recognizing that real and apparent contradictions may exist between information sources which need physical understanding. Consistency and comprehensiveness of the physical and dynamical basis of the projected climate response across models and methods should be evaluated.

- It should be recognized that additional forcings and feedbacks, which may not be fully represented in global models, may be important for regional climate change (e.g., land use change and the influence of atmospheric pollutants).

- When quantitative information is limited or missing, assessments may provide narratives of climate projections (storylines, quantitative or qualitative descriptions of illustrative possible realizations of climate change) in addition or as an alternative to maps, averages, ranges, scatter plots or formal statistical frameworks for the representation of uncertainty.

- Limits to the information content of climate model output for regional projections need to be communicated more clearly. The relative importance of uncertainties typically increase for small scales and impact relevant quantities due to limitations in model resolution, local feedbacks and forcings, low signal to noise ratio of observed trends, and possibly other confounding factors relevant for local impacts. Scientific papers and IPCC assessments should clearly identify these limitations.

- Impact assessments are made for multiple reasons, using different methodological approaches. Depending on purpose, some impact studies sample the uncertainty space more thoroughly than others. Some process or sensitivity studies may legitimately reach a specific conclusion using a single global climate model or downscaled product. For policy-relevant impact studies it is desirable to sample the uncertainty space by evaluating global and regional climate model ensembles and downscaling techniques. Multiple lines of evidence should always be considered.

- In particular for regional applications, some climate models may not be considered due to their poor performance for some regional metric or relevant process (e.g., for an Arctic climate impact assessment models need to appropriately simulate regional sea-ice extent). However, there are no simple rules or criteria to define this distinction, however. Whether or not a particular set of models should be considered is a different research-specific question in every special case. Selection criteria for model assessment should be based, among other factors, on availability of specific parameters, spatial and temporal resolution within the model and so need to be made transparent.

- The usefulness and applicability of downscaling methods strongly depends on the purpose of the assessment (e.g., for the analysis of extreme events or assessments in complex terrain). If only a subsample of the available global climate model uncertainty space is sampled for the downscaling, this should be stated explicitly.

- When comparing different impact assessments, IPCC authors need to carefully consider the different assumptions, climate and socio-economic baselines, time horizons and emission scenarios used. Many impact studies are affected by the relative similarity between different emission scenarios in the near term. Consideration of impact assessments based on the earlier emission scenarios (IPCC Special Report on Emission Scenarios, SRES) in the light of the new scenario structure (Representative Concentration Pathways, RCP) represents a considerable challenge. The length of time period considered in the assessment studies can significantly affect results.

3.6 Considerations for the WGI Atlas of Global and Regional Climate Projections

The WGI Atlas of Global and Regional Climate Projections in IPCC AR5 is intended to provide comprehensive information on a selected range of climate variables (e.g., temperature and precipitation) for a few selected time horizons for all regions and, to the extent possible, for the four basic RCP emission scenarios. All the information used in the Atlas will be based on material assessed in WGI Chapters 11, 12 or 14 (see http://www.ipcc-wg1.unibe.ch/AR5/chapteroutline.html).

There may, however, be disagreement between the downscaling literature and the content of the Atlas based on GCMs and this should be explained and resolved as far as possible. The limitations to the interpretation of the Atlas material should be explicit and prominently presented ahead of the projections themselves.

Options for information from multi-model simulations could include medians, percentile ranges of model outputs, scatter plots of temperature, precipitation and other variables, regions of high/low confidence, changes in variability and extremes, stability of teleconnections, decadal time-slices, and weighted and unweighted representations of any of the above. The information could include CMIP5 as well as CMIP3 simulations.

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.

Annan, J.D., and J.C. Hargreaves, 2010: Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.*, **37**, doi:10.1029/2009gl041994.

Benestad, R.E., 2005: Climate change scenarios for northern Europe from multi-model IPCC AR4 climate simulations. *Geophys. Res. Lett.*, **32**, doi:10.1029/2005gl023401.

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.

Christensen, J.H., B. Hewitson, A. Busuioc, A. Chen, X. Gao, I. Held, R. Jones, R.K. Kolli, W.-T. Kwon, R. Laprise, V. Magana Rueda, L. Mearns, C.G. Menendez, J. Raisanen, A. Rinke, A. Sarr, and P. Whetton, 2007: Regional Climate Projections. In: *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, Cambridge, United Kingdom, and New York, NY USA, 847-845.

Connolley, W.M., and T.J. Bracegirdle, 2007: An Antarctic assessment of IPCC AR4 coupled models. *Geophys. Res. Lett.*, **34**, doi:10.1029/2007gl031648.

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.*, **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.

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

Furrer, R., R. Knutti, S.R. Sain, D.W. Nychka, and G.A. Meehl, 2007: Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophys. Res. Lett.*, **34**, L06711, doi:10.1029/2006GL027754.

Giorgi, F., and L.O. Mearns, 2003: Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophys. Res. Lett.*, **30**, 1629, doi:10.1029/2003GL 017130.

Gleckler, P.J., K.E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, doi:10.1029/2007JD008972.

Greene, A.M., L. Goddard, and U. Lall, 2006: Probabilistic multimodel regional temperature change projections. *J. Clim.*, **19**, 4326-4346.

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.

Harris, G.R., M. Collins, D.M.H. Sexton, J.M. Murphy, and B.B.B. Booth, 2010: Probabilistic Projections for 21st Century European Climate. *Nat. Haz. and Earth Sys. Sci.*, (submitted).

IPCC, 2007: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S. D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 996pp.

Jackson, C.S., M.K. Sen, G. Huerta, Y. Deng, and K.P. Bowman, 2008: Error Reduction and Convergence in Climate Prediction. *J. Clim.*, **21**, 6698-6709.

Knutti, R., 2008: Should we believe model predictions of future climate change? *Phil. Trans. Royal Soc. A*, **366**, 4647-4664.

Knutti, R., 2010: The end of model democracy? *Clim. Change*, published online, doi:10.1007/s10584-010-9800-2 (in press).

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., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining projections from multiple models. *J. Clim.*, **23**, 2739-2756, doi: 10.1175/2009JCLI3361.1.

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.

Mearns, L.O., W.J. Gutowski, R. Jones, L.-Y. Leung, S. McGinnis, A.M.B. Nunes, and Y. Qian, 2009: A regional climate change assessment program for North America. *EOS*, **90**, 311-312.

Murphy, J., D. Sexton, G. Jenkins, P. Boorman, B. Booth, K. Brown, R. Clark, M. Collins, G. Harris, and E. Kendon, 2009: Climate change projections, ISBN 978-1-906360-02-3.

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. *Phil. Trans. Royal Soc. A*, **365**, 1993-2028.

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.

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.

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. In: *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, Cambridge, United Kingdom and New York, NY, USA, 589-662.

Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Am. Met. Soc.*, **89**, 303-311.

Rougier, J., 2007: Probabilistic inference for future climate using an ensemble of climate model evaluations. *Clim. Change*, **81**, 247-264.

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.

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, R.L., C. Tebaldi, D.W. Nychka, and L.O. Mearns, 2009: Bayesian modeling of uncertainty in ensembles of climate models. *J. Am. Stat. Assoc.*, **104**, 97-116, doi:10.1198/jasa.2009. 0007.

Stephenson, D.B., C.A.S. Coelho, F.J. Doblas-Reyes, and M. Balmaseda, 2005: Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus A*, **57**, 253-264.

Stott, P.A., J.A. Kettleborough, and M.R. Allen, 2006: Uncertainty in continental-scale temperature predictions. *Geophys. Res. Lett.*, **33**, L02708.

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., 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.

Waugh, D.W., and V. Eyring, 2008: Quantitative performance metrics for stratospheric-resolving chemistry-climate models. *Atm. Chem. Phys.*, **8**, 5699-5713.

Zhang, X.B., F.W. Zwiers, G.C. Hegerl, F.H. Lambert, N.P. Gillett, S. Solomon, P.A. Stott, and T. Nozawa, 2007: Detection of human influence on twentieth-century precipitation trends. *Nature*, **448**, 461-466.