ipcc
INTERGOVERNMENTAL PANEL ON climate change

WMO UNEP

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections

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

Meeting Report

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

This meeting was agreed in advance as part of the IPCC workplan, but this does not imply working group or panel endorsement or approval of the proceedings or any recommendations or conclusions contained herein.

Supporting material prepared for consideration by the Intergovernmental Panel on Climate Change. This material has not been subjected to formal IPCC review processes.

ipcc
INTERGOVERNMENTAL PANEL ON climate change

Working Group I (WG I) – The Physical Science Basis

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections

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

Meeting Report

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

This meeting was agreed in advance as part of the IPCC workplan, but this does not imply working group or panel endorsement or approval of the proceedings or any recommendations or conclusions contained herein.

Supporting material prepared for consideration by the Intergovernmental Panel on Climate Change.
This material has not been subjected to formal IPCC review processes.

WG I Technical Support Unit · c/o University of Bern
Zaehringerstrasse 25 · 3012 Bern · Switzerland
telephone +41 31 631 5616 · fax +41 31 631 5615 · email wg1@ipcc.unibe.ch · **www.ipcc-wg1.unibe.ch**

WMO UNEP

ISBN 978-92-9169-129-6

Published July 2010 by the IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland. Electronic copies of this report are available from the IPCC website (http://www.ipcc.ch/) and the IPCC WGI website (http://www.ipcc-wg1.unibe.ch/).

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections

25-27 January 2010
Boulder, Colorado, USA

WGI Co-Chairs

Thomas Stocker (Physics Institute, University of Bern, Switzerland)
Dahe Qin (China Meteorological Administration, China)

WGII Co-Chairs

Christopher Field (Carnegie Institution, Stanford University, USA)
Vicente Barros (Ciudad Universitaria, Argentina)

Scientific Steering Committee

Thomas Stocker (IPCC WGI Co-Chair, Physics Institute, University of Bern, Switzerland)
Christopher Field (IPCC WGII Co-Chair, Carnegie Institution, Stanford University, USA)
Matthew Collins (Hadley Centre for Climate Prediction and Research, United Kingdom)
Reto Knutti (ETH Zurich, Switzerland)
Linda Mearns (National Center for Atmospheric Research, USA)
Benjamin Santer (Lawrence Livermore National Laboratory, USA)
Dáithí Stone (University of Cape Town, South Africa)
Penny Whetton (Commonwealth Scientific and Industrial Research Organization, Australia)

Good Practice Guidance Paper Core Writing Team

Gabriel Abramowitz (University of New South Wales, Australia)
Matthew Collins (Hadley Centre for Climate Prediction and Research, United Kingdom)
Veronika Eyring (Deutsches Zentrum für Luft- und Raumfahrt, Germany)
Peter Gleckler (Lawrence Livermore National Laboratory, USA)
Bruce Hewitson (University of Cape Town, South Africa)
Reto Knutti (ETH Zurich, Switzerland)
Linda Mearns (National Center for Atmospheric Research, USA)

Local Support

Kyle Terran (Joint Office for Science Support, University Corporation for Atmospheric Research)

IPCC Working Group I Technical Support Unit

Pauline Midgley
Gian-Kasper Plattner (Coordinating Editor)
Melinda Tignor (Technical Editor)
Judith Boschung
Vincent Bex

IPCC Working Group II Technical Support Unit

Kristie Ebi
David Dokken

This Meeting Report should be cited as:

IPCC, 2010: *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, and P.M. Midgley (eds.)]. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland, pp. 117.

Preface

Climate model results provide the basis for projections of future climate change and increasing numbers of models are likely to contribute to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR5). The heterogeneity in the new generation of climate models and an increasing emphasis on estimates of uncertainty in the projections raise questions about how best to evaluate and combine model results in order to improve the reliability of projections. Therefore, an Expert Meeting on Assessing and Combining Multi Model Climate Projections was organized by IPCC's Working Group I with contributions from Working Group II (WGI & WGII). The meeting was held in Boulder, Colorado, USA, from 25 to 27 January 2010. It is important for the success of the IPCC AR5 that this discussion took place early in the AR5 assessment cycle. The Expert Meeting provided a platform to explore the possibility of establishing a common framework that is applicable across the two IPCC Working Groups.

The scientific core of this meeting report summarises the discussions and conclusions of the Expert Meeting. It contains a stand-alone Good Practice Guidance Paper for IPCC Lead Authors of AR5, as well as for scientists working in model intercomparison projects. The Guidance Paper briefly summarizes methods used in assessing the quality and reliability of climate model simulations and in combining results from multiple models, and provides recommendations for good practice in using multi-model ensembles. Applications include detection and attribution, model evaluation and global climate projections as well as regional projections relevant for impact and adaptation studies. This meeting report further includes the extended abstracts of the presentations and posters from the Expert Meeting as well as a non-comprehensive bibliography of relevant literature.

We extend our sincere thanks to the National Center for Atmospheric Research and the University Corporation for Atmospheric Research for hosting the meeting and for the excellent local arrangements. We are also very appreciative of the advice of the members of the Scientific Steering Committee who shaped the meeting programme as well as for their help in carrying it out. We would like to thank all participants who contributed to a very constructive and fruitful meeting where the exchange of views and knowledge resulted in more clarity on the issues involved and the current status of scientific understanding. The members of the core writing team put in many hours of effort following the meeting in order to produce the Good Practice Guidance Paper in a timely fashion, for which we are very grateful. The excellent work by the Technical Support Unit of WGI at all stages of the meeting organisation and report production is appreciated.

We are sure that the product of this Expert Meeting will provide useful service to the scientific community, in particular to the many AR5 Lead Authors in WGI and WGII who will assess the information that is derived from the wide range of climate model simulations.

Thomas Stocker

Prof. Thomas Stocker
Co-Chair, WGI

Qin Dahe

Prof. Qin Dahe
Co-Chair, WGI

Christopher Field

Prof. Christopher Field
Co-Chair, WGII

Prof. Vicente Barros
Co-Chair, WGII

Table of Contents

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)

Citation: Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P.J. Gleckler, B. Hewitson, and L. Mearns, 2010: Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections. In: *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, and P.M. Midgley (eds.)]. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland.

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 covariability 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 performance 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).

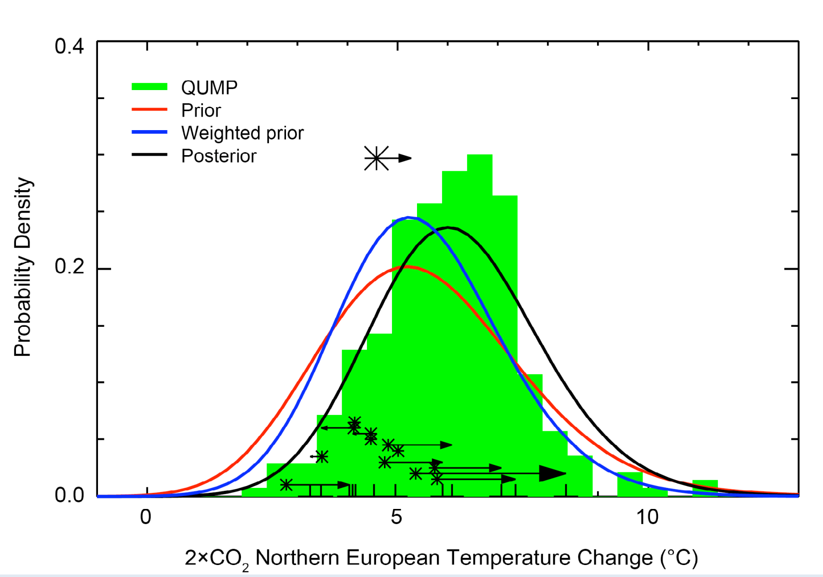
- The distinction between 'best effort' simulations (i.e., the results from the default version of a model submitted to a multi-model database) and perturbed physics ensembles is important and must be recognized. Perturbed physics ensembles can provide useful information about the spread of possible 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).

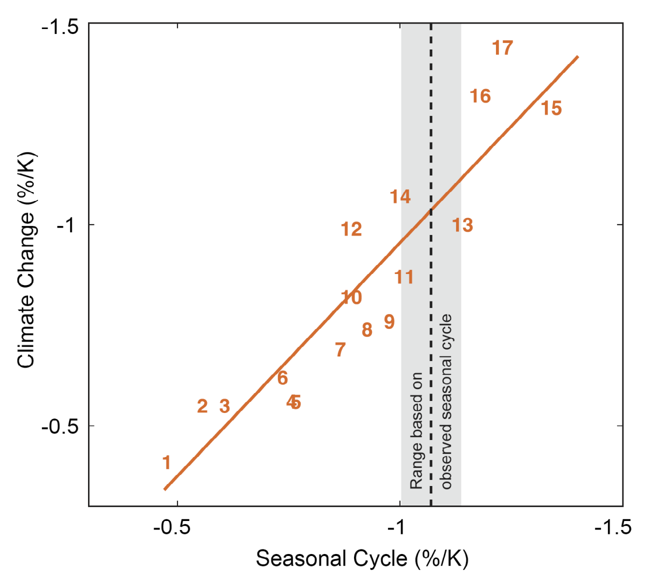
Box 3.2: Examples of Model Evaluation Through Relationships Between Present-Day Observables and Projected Future Changes

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.

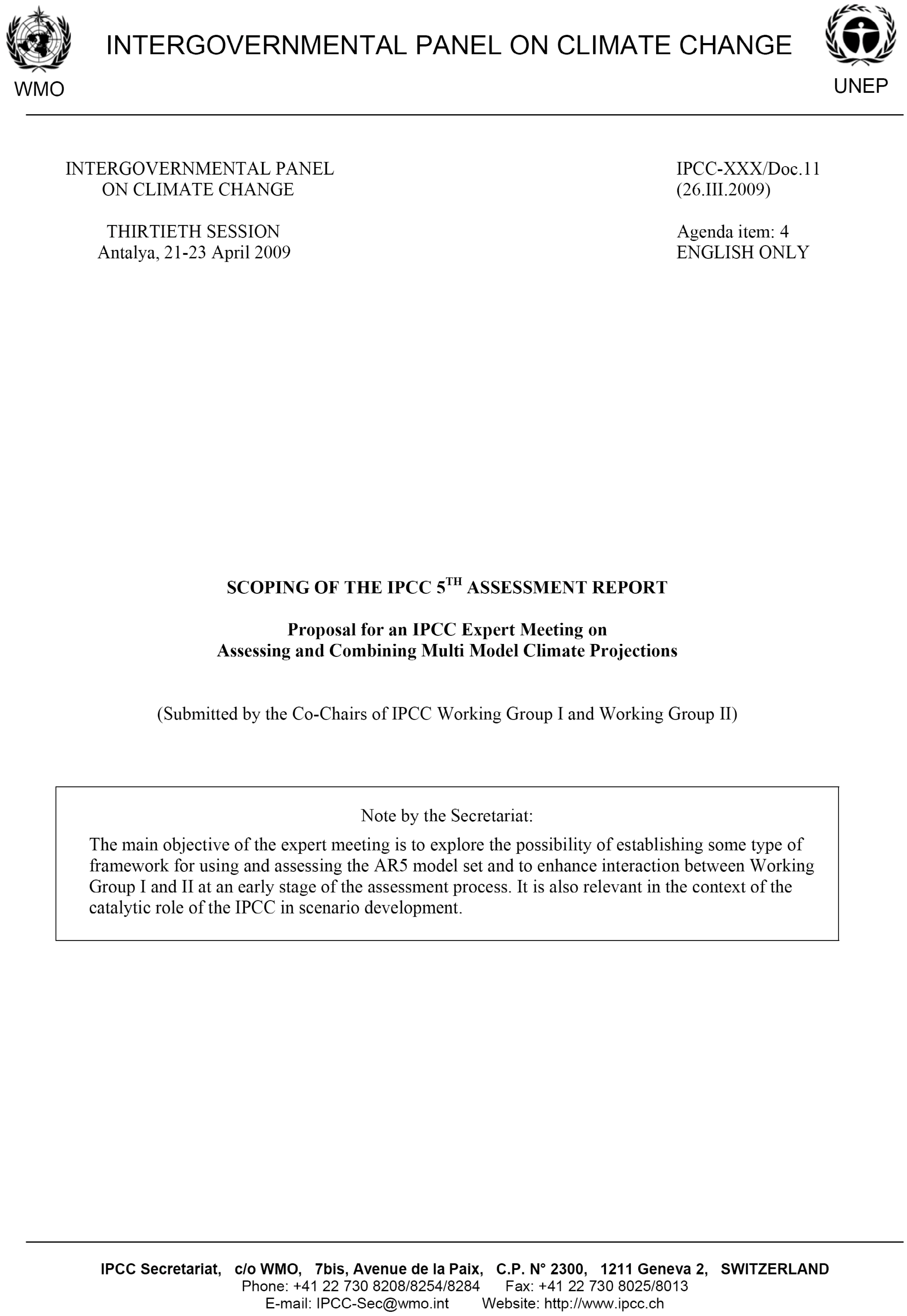
Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Phil. Trans. Royal Soc. A*, **365**, 2053-2075.

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.

Annex 1: Proposal



INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE

WMO UNEP

INTERGOVERNMENTAL PANEL
ON CLIMATE CHANGE

THIRTIETH SESSION
Antalya, 21-23 April 2009

IPCC-XXX/Doc.11
(26.III.2009)

Agenda item: 4
ENGLISH ONLY

SCOPING OF THE IPCC 5TH ASSESSMENT REPORT

Proposal for an IPCC Expert Meeting on
Assessing and Combining Multi Model Climate Projections

(Submitted by the Co-Chairs of IPCC Working Group I and Working Group II)

Note by the Secretariat:

The main objective of the expert meeting is to explore the possibility of establishing some type of framework for using and assessing the AR5 model set and to enhance interaction between Working Group I and II at an early stage of the assessment process. It is also relevant in the context of the catalytic role of the IPCC in scenario development.

Proposal for an IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections

Submitted by the Co-Chairs of IPCC Working Group I and Working Group II

Background

Climate model results provide the basis for IPCC projections of future climate change. Previous assessment reports included model evaluations but avoided weighting or ranking models. Projections and uncertainties were based on a 'one model, one vote' approach, despite the fact that they 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 the Coupled Model Intercomparison Phase 5 (WCRP CMIP5), a collaborative process in which the community has agreed on the type of simulations to be performed. The widespread participation in CMIP5 provides some perspective on model uncertainty. Nevertheless, these intercomparisons are not designed to yield formal error estimates and remain 'ensembles of opportunity.'

Since participation in the IPCC process is important for modelling centres, the number of models and model versions is likely to increase in CMIP5. Some groups may submit multiple versions of the same model with different parameter settings. The new generation of models is likely to be more heterogeneous than ever, as some but not all of the new models will include interactive representations of biogeochemical cycles, chemistry, ice sheets, land use or interactive vegetation. This makes a simple model average increasingly difficult to defend and to interpret. Many models are not independent and some are clearly more robust than others when compared with selected observations.

The reliability of projections could be improved if the models were weighted according to some measure of skill and if their interdependencies were taken into account. Indeed such methods using forecast verification were shown to be superior to simple averages in the area of weather forecasting. Since there is no verification for a climate forecast on timescales of decades to centuries, the skill or performance of the models needs to be defined, for example, by comparing simulated patterns of present day climate to observations. Such metrics are useful but not unique and often it is unclear how they relate to the forecast of interest. Defining a set of minimum criteria for a model to be 'credible' or agreeing on a metric of performance is therefore difficult and the criteria are likely to depend on the variable and timescale of interest. Combined with an estimated data volume exceeding 1000 Terabytes, the AR5 faces immense obstacles in trying to make sense of the deluge of model runs and data that it will produce.

Recent studies have started to address these issues by proposing ways to weight or rank models, based on process evaluation, agreement with present day observations, past climate or observed trends. While there is agreement that 'the end of model democracy' may be near, there is no consensus on how such a model selection or weighting process could be agreed upon. An IPCC expert meeting addressing these important questions will help to bring the community into a position to make better use of the new model results and will provide more robust and reliable projections of future climate, along with improved estimates of uncertainty. At the same time, the dialogue between WGI and WGII should be strengthened in order to determine what kind of model results from WGI can be provided to WGII and how that exchange can be organized efficiently, given the tight schedule of the AR5.

Objectives

The main objective of the expert meeting is to see if it is possible to establish some type of framework for using and assessing the AR5 model set. Components of this effort are

- To stimulate discussion on metrics to evaluate climate models;
- To learn from other communities where model skill based on forecast verification is used;
- To assess the potential of different model weighting, ranking and selection schemes for not equally credible models for their use in IPCC AR5;
- To determine whether minimum model performance requirements for inclusion in AR5 should and can be defined.

Expected outcome

The expert meeting will provide tentative best practices in selecting and combining results from multiple models for IPCC AR5; in short the beginning of a quantitative framework for analysis and assessment of the models. Specific aims of the meeting will be to maximize the robustness and policy relevance of the projections provided in the presence of model error, projection uncertainty, observational uncertainties and a heterogeneous set of models. Interactions between WGI and WGII will be ensured by the participation of a number of representatives from WGII with broad expertise on impacts and user needs.

Initial organising committee (a broader scientific steering committee will be formed)
Prof. Reto Knutti (ETH Zurich, Switzerland)
Dr. Benjamin Santer (Lawrence Livermore National Lab, USA)
Dr. Penny Whetton (CSIRO, Australia)
Dr. Mat Collins (Met Office Hadley Centre, UK)
Dr. Daithi Stone (University of Cape Town, South Africa)
Dr. Claudia Tebaldi (Climate Central/NCAR, USA)
Dr. Karl Taylor (Program for Climate Model Diagnosis and Intercomparison, LLNL, USA)

Timing: late 2009/early 2010, after the AR5 Scoping Meeting (dependent on date decided for IPCC 31st Plenary)

Duration: 2.5 to 3 days

Location: possible host organizations have been identified and are being explored with the relevant IPCC national focal point

Participants: About 40 participants in total, with broad international representation. It is proposed that 16 journeys for experts from developing countries and economies in transition including WGI and WGII Vice-Chairs are allocated as part of the line item “expert meetings related to the AR5” in the already agreed IPCC Trust Fund budget for 2009.

Expertise: Climate model development, model evaluation, statistical methods, uncertainty quantification.

Annex 2: Programme

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections

National Center for Atmospheric Research, Mesa Laboratory, Boulder, Colorado, USA

25-27 January 2010

PROGRAMME

Monday, 25 January 2010

Shuttles depart hotels for venue (see schedule for details)

08:30	**Registration *(Seminar Room Foyer)***
09:00	**Welcome Address: *Alice Madden, Climate Change Coordinator, Colorado Governor's Office***
09:10	**Welcome Address: *Eric Barron, Director, National Center for Atmospheric Research***
09:20	**Welcome and Opening (Qin/Stocker/Field)**
PLENARY SESSION I (Chair: Qin Dahe)	
09:30	**Introduction, Background and Challenges (Thomas Stocker)**
09:50	**Keynote Presentation: *Challenges in Combining Projections from Multiple Climate Models* (Reto Knutti)** [20 min presentation + 5 min discussion]
10:15	**Keynote Presentation: *An Overview of Approaches to Future Projections Based on Multi-Model Ensembles* (Claudia Tebaldi)** [20 min presentation + 5 min discussion]
10:40	**Break *(Seminar Room Foyer)***
11:10	**Keynote Presentation: *Thoughts on the Use of Multi-Model Ensembles* (Isaac Held)** [20 min presentation + 5 min discussion]
11:35	**Keynote Presentation: *The Difficulties Involved in Identifying the "Best" Model from a Large, Multi-Model Archive* (Ben Santer)** [20 min presentation + 5 min discussion]
12:00	**General Discussion**
12:30	**Lunch *(NCAR Cafeteria)***
13:30	**Introduction of Break-Out Groups and Good Practice Guidance Paper (Stocker and Field) *(Seminar Room)***

BREAK-OUT GROUPS - PART A: Topical Discussions	
14:00	**BOG1: Extracting Information from Global Model Projections [Chair: Dáithí Stone & Rapporteur: Mat Collins]** ***(Damon Room)*** **BOG2: Extracting Model Information for Regional Projections and Impacts [Chair: Penny Whetton & Rapporteur: Wolfgang Cramer]** ***(Directors Room)*** **BOG3: Feasibility and Implications of Model Ranking [Chair: Jerry Meehl & Rapporteur: Peter Stott]** ***(Seminar Room)***
16:00	**Break** ***(Seminar Room Foyer)***
16:30	**Reports from Break-Out Groups (BOG Chairs)** ***(Seminar Room)***
POSTER SESSION I	
17:00	**Poster Session** ***(Seminar Room Foyer)***
18:30	**Adjourn**
18:30	**Welcome Reception** ***(NCAR Cafeteria)***

19:45 Shuttles depart venue for hotels

Tuesday, 26 January 2010

Shuttles depart hotels for venue (see schedule for details)

08:30	**Summary Day 1; Introduction Day 2 (Plattner)** ***(Seminar Room)***
PLENARY SESSION II (Chair: Chris Field)	
08:35	**Keynote Presentation:** ***Linking Detection and Attribution with Probabilistic Climate Forecasting*** **(Myles Allen)** [20 min presentation + 5 min discussion]
09:00	**Keynote Presentation:** ***Probabilistic Projections of Climate Change at Global and Regional Scales*** **(David Sexton)** [20 min presentation + 5 min discussion]
09:25	**Keynote Presentation:** ***Extracting Information from Regional Multi-Model Climate Change Projections*** **(Bruce Hewitson)** [20 min presentation + 5 min discussion]
09:50	**Break** ***(Seminar Room Foyer)***
10:20	**Keynote Presentation:** ***Representing Multi-Model Climate Projection Uncertainties in Modelling Impact Risks and Adaptation Options: Recent Advances in Europe*** **(Tim Carter)** [20 min presentation + 5 min discussion]
10:45	**Keynote Presentation:** ***Bringing it All Together: Are We Going in the Right Direction for Providing Users with Better Information About Future Climate to Support Decision-Making?*** **(Linda Mearns)** [20 min presentation + 5 min discussion]
11:10	**General Discussion**
11:30	**Discussion on Good Practice Guidance Paper: Purpose and Structure (Thomas Stocker)**
12:30	**Lunch** ***(NCAR Cafeteria)***
POSTER SESSION II	
13:30	**Poster Session** ***(Seminar Room Foyer)***
15:00	**Break** ***(Seminar Room Foyer)***
BREAK-OUT GROUPS - PART B: Structure of Good Practice Guidance Paper	
15:30	**BOG1: Methods [Chair: Dáithí Stone & Rapporteur: Mat Collins]** ***(Damon Room)*** **BOG2: Spatial and Temporal Applications and Specific User Needs [Chair: Penny Whetton & Rapporteur: Wolfgang Cramer]** ***(Directors Room)*** **BOG3: Feasibility and Implications [Chair: Jerry Meehl & Rapporteur: Peter Stott]** ***(Seminar Room)***
17:30	**Reports from Break-Out Groups (BOG Chairs)** ***(Seminar Room)***
18:00	**Adjourn**

18:15 ***Shuttles depart venue for hotels***

Wednesday, 27 January 2010

Shuttles depart hotels for venue (see schedule for details)

08:30	Summary Day 2; Introduction Day 3 (Plattner) *(Seminar Room)*
BREAK-OUT GROUPS - PART B Continued: Drafting of Bullets/Outline for Good Practice Guidance Paper	
08:35	BOG1: Methods [Chair: Dáithí Stone & Rapporteur: Mat Collins] *(Damon Room)* BOG2: Spatial and Temporal Applications and Specific User Needs [Chair: Penny Whetton & Rapporteur: Wolfgang Cramer] *(Directors Room)* BOG3: Feasibility and Implications [Chair: Jerry Meehl & Rapporteur: Peter Stott] *(Seminar Room)*
10:00	Reports from Break-Out Groups (BOG Chairs) *(Seminar Room)*
10:30	Break (Seminar Room Foyer)
PLENARY SESSION III: Good Practice Guidance Paper (Chair: Thomas Stocker)	
11:00	Plenary Approval of Executive Summary of Guidance Paper
12:45	Closing Remarks and Next Steps (Qin/Stocker/Field)
13:00	End of Meeting

13:15 *Shuttles depart venue for hotels*

Annex 3: Extended Abstracts

** Speaker*
† Submitting participant(s)

Model Independence Weights for Multi-Model Ensembles

Gab Abramowitz[1] and Craig Bishop[2]

[1] *Climate Change Research Centre, University of New South Wales, Australia*
[2] *Navel Research Laboratory, USA*

This talk will present a weighting strategy that accounts for model independence. In this case, independence is defined as correlation between model errors. The 24 CMIP3 models in the PCMDI database are used together with 30 years of 5° x 5° monthly surface air temperature from the HadCRUT3 observed dataset. Seven different weighting strategies are compared:

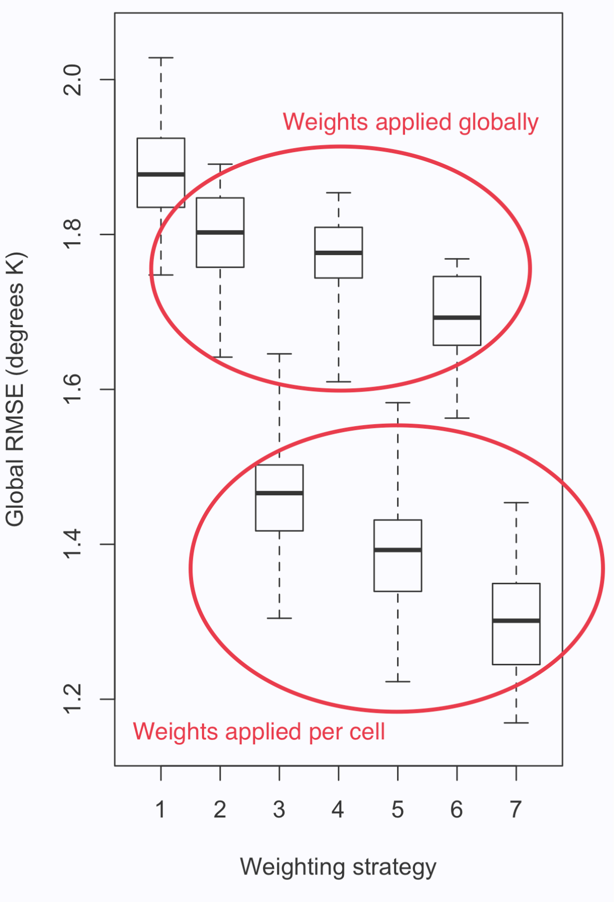


Figure 1. 1) Multi-model mean; 2) Globally bias-corrected multi-model mean; 3) Per-cell bias-corrected multi-model mean; 4) Global bias-corrected performance weights (minimising error variance); 5) Per-cell bias-corrected performance weights; 6) Global bias-corrected performance and independence weights; and 7) Per-cell bias-corrected performance and independence weights

Of the 30 available years, 29 are used to derive model weights (and bias corrections) and the remaining year used to test them. The experiment is repeated for all 30 possible testing years to produce the out-of-sample results in the box and whisker plots shown here. These results suggest that gains from independence weighting are *at least as large* as those from performance weighting, and tend to have an additive effect. We will discuss the benefits and potential downfalls of bias-correcting and weighting on a per-gridcell basis, give a brief overview of the derivation of the weights and detail a potential IPCC strategy for considering the effective independent contribution from individual models for any given variable.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Linking Detection and Attribution with Probabilistic Climate Forecasting

Myles Allen

Department of Physics, University of Oxford, United Kingdom

Some of the earliest studies attempting to quantify uncertainty in climate forecasts emerged directly from the detection and attribution literature of the 1990s, notably the optimal fingerprinting approach of Hasselmann (1993,1997), Santer et al. (1995) and Hegerl et al. (1996). Leroy (1998) and Allen and Tett (1999) observed that optimal fingerprinting could be cast as a linear regression problem in which it is assumed that climate models simulate the patterns of the climate response to various external drivers correctly, and observations are used to estimate the magnitude of that response. A subsequent generalisation by Huntingford et al. (2006) allows for some uncertainty in the patterns of response, but is still based on the principle that models provide much more reliable information regarding response patterns than response magnitudes.

The physical justification for this principle is strong: the spatial pattern of response to, for example, greenhouse forcing is driven by the differences in heat capacity between land and ocean and the location of the continents, which are not model-dependent. Likewise, the temporal pattern of response depends primarily on the time-history of greenhouse forcing and only secondarily on the time-scales of the response. In contrast, the magnitude of the response depends on the transient climate response, or TCR. This in turn depends on the atmospheric feedbacks that control the equilibrium climate sensitivity and on the efficiency of ocean heat uptake, both of which are uncertain.

Hence, in the context of multi-model ensembles, optimal fingerprinting is equivalent to generating a large "pseudo-ensemble" simply by taking the mean pattern of response to a given external forcing as simulated by a small ensemble and scaling it up and down by an arbitrary parameter representing uncertainty in response magnitude. It is important that responses to short-term (e.g., volcanic) and long-term (e.g., most anthropogenic) forcings are estimated separately using a multiple regression, since uncertainty in the time-constants of the climate system (primarily linked to ocean heat uptake) mean that errors in response magnitude may be very different on different timescales. Ideally, the response to anthropogenic aerosol forcing should also be estimated separately from the response to greenhouse forcing: although both operate on similar timescales, some potential sources of uncertainty in the aerosol response do not affect the greenhouse response, and vice versa. Hence a pre-requisite for this approach are separate simulations of the responses to individual forcings, either separately or in combinations.

The goodness-of-fit between individual members of this pseudo-ensemble are then evaluated with a standard weighted sum of squares, with the expected model-data differences due to internal climate variability, observation error and (in some studies) model pattern uncertainty providing the matrix of weights or metric. The range of, for example, the warming attributable to anthropogenic greenhouse gas increases over the past 50 years across the members of this pseudo-ensemble that fit the data better than would be expected by chance in, say, 90% of cases provides a confidence interval on this quantity. This approach is the primary information source for attribution statements in the IPCC Third and Fourth Assessments.

Applying the same scaling factors to model-simulated responses to future forcing provides a natural method of deriving confidence intervals on future climate change. This approach was used by Allen et al. (2000), Stott and Kettleborough (2002) and, for regional changes, by Stott et al. (2006), and has been referred to as the ASK approach. The crucial assumption (which is also implicit in attribution studies) is that fractional errors in model-simulated responses persist over time, so a model that underestimates the past response to a given forcing by, for example, 30% may be expected to continue to do so in the future. This assumption is supported by comparing model results for scenarios under which forcing is sustained into the future, such as A1B (Stott et al., 2006), but Allen et al. (2000) note that it would be less reliable for stabilisation scenarios.

The ASK approach can provide ranges of uncertainty in forecast climate that may, for variables that are poorly constrained by observations, be much wider than the range of available model simulations. This was clearly an advantage when very few models were available, and will continue to be necessary as long as the spread of model simulations is thought to underestimate the full range of uncertainty. ASK therefore provides a complementary approach to more recent methods of probabilistic forecasting such as weighted or un-weighted perturbed-physics or multi-model ensembles. There are, however, some important points of principle in which ASK as traditionally implemented differs from most ensemble-based approaches, which need to be addressed if results are to be compared cleanly.

Consistent with the attribution literature, ASK provides classical ("frequentist") confidence intervals – that is, ranges over which models match observations better than a given threshold for goodness-of-fit. In contrast, most ensemble-based approaches provide Bayesian posterior probability intervals – ranges within which a given percentage of the weighted ensemble is found to lie. These are only comparable if ensemble members are distributed uniformly across the observable quantities that are used to constrain them and uncertainties in these quantities are approximately Gaussian (the so-called Jeffreys' prior condition). If the constraints provided by the observations are weak and models tend to cluster near the best-fitting model (as would be expected if all modelling groups are aiming to simulate observations as well as possible), these conditions are not satisfied, so ranges provided by ASK are not directly comparable to ranges provided by other approaches. Worse, ranges on forecast anthropogenic warming will then not be consistent with ranges on past anthropogenic warming, leading to the absurd conclusion that we are less uncertain about the future than we are about the recent past (Frame et al., 2007).

A fundamental issue here is that the standard uncertainty qualifiers used by Working Group 1 ("likely", "very likely" etc.) are used to refer both to classical confidence intervals and Bayesian posteriors. The nominal definition of these qualifiers is unambiguously Bayesian, but in many, perhaps most, instances they are used to refer to classical confidence intervals or the results of hypothesis tests. This ambiguity of usage within IPCC has already attracted criticism among statisticians (Spiegelhalter, 2008). A simple solution would be to restrict the use of "likely" etc. to cases in which a confidence interval can be derived or hypothesis test performed (which refer, appropriately, to likelihoods of goodness-of-fit) and to use the more explicitly Bayesian language recommended by Moss and Schneider (2000) for Bayesian posterior probabilities.

Using Multi-Model Ensembles to Estimate Distributions of Local 2m Temperature Scenarios

Rasmus E. Benestad

Norwegian Meteorological Institute, Norway

An advantage of empirical-statistical downscaling (ESD) over dynamical downscaling with regional climate models (RCMs) is that ESD requires much less computer resources, and hence is more easily applied to both large sets of global climate model (GCM) simulations as well as long time series. Furthermore, ESD can incorporate ways to evaluate the GCMs in terms of their ability of reproducing typical observed coherent spatial patterns of variability around the location of interest. Here, ESD work has involved the use of so-called 'common EOFs' to ensure that the same spatial structures associated with local 2m temperature variations are also extracted from the GCM results and used to make projections for the future and the past. In other words, ESD can both give an indication of the GCM's ability to simulate the regional climate variability, as well as providing estimates for how the local climate variable is projected to change. An arbitrary set of local skill scores has been used to weight the local T(2m) projections in weighted means, but more recent analysis has also used unweighted ensembles to make probabilistic projections for simulated temperature. The ESD analysis has been applied to around 1000 locations over the entire world (and incorporated into GoogleEarth). One benefit of this exercise is that it provides a consistent set-up for all continents, using the same methodology for all locations. Hence, it provides a valuable set of benchmark values against which RCM simulations can be compared.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Representing Multi-Model Climate Projection Uncertainties in Modelling Impact Risks and Adaptation Options: Recent Advances in Europe

Timothy R. Carter

Finnish Environment Institute (SYKE), Finland

The basic approach for assessing impacts of future anthropogenic climate change has changed little since the results of equilibrium general circulation model (GCM) simulations were first adopted by impact analysts in the early 1980s. This has typically involved constructing one or more scenarios of future climate to represent uncertainties in climate projections from the set of available model-based simulations. Since climate represents only one of many possible sources of uncertainty in estimates of future impacts, constraints of computer power, time or availability of climate projections have limited the number of climate scenarios that can be adopted in a study, particularly if impact models are complex and require detailed inputs. Hence, results of impact studies can be strongly influenced by the scenario "ensemble of opportunity" selected, especially if there are large differences in climate projections. This can pose problems when interpreting the results of scenario-based impact studies. Moreover, inconsistencies in the climate scenarios adopted in different impact studies has bedevilled efforts at comparative assessment across studies, regions and sectors by IPCC Working Group II and other bodies.

In recent years improved computer power combined with a more diverse set of climate models has facilitated the production of ever-larger numbers of multi-model climate projections, representing a wide range of sources of uncertainty in future climate. Using these multi-model ensembles, along with systematic model parameter uncertainty analysis and statistical and dynamical techniques for regionalising GCM projections, climate researchers are increasingly moving towards the representation of changes in future regional climate in terms of probabilities. Indeed, a situation can be envisaged in the next few years where probabilistic projections of regional climate changes for a given scenario of future emissions could become routine, using a consistent methodology to represent uncertainties on the basis of available model-based and observational information. For example, there is no reason why a methodology similar to that employed in developing climate projections for the United Kingdom (UKCP09) (Murphy et al., 2009) could not be applied to regions worldwide.

Multi-model climate projections and enhanced computer power also provide an opportunity to undertake multiple simulations with impact models and to examine a wider range of uncertainties in the outcomes. New techniques are being developed to combine uncertainties in climate projections with uncertainties in impacts, and some forms of adaptation can also be analysed using a modelling framework. Moreover, rather than conducting "what if" scenario-based impact studies, there could be added value in expressing future impacts in terms of the risks of exceeding pre-defined thresholds of impact (Jones, 2000). Probabilistic projections of climate offer the prospect of developing more systematic approaches to the determination of impact risks.

We report here some exploratory analysis of impact risks that has been conducted in the European Union funded ENSEMBLES project, using new multi-model climate projections for Europe (van der Linden and Mitchell, 2009). Impacts have been estimated using both conventional scenario-based analysis with a focus on analysis of extreme events (Figure 1, right hand pathways) as well as probabilistic climate projections in combination with impact response surfaces (Figure 1, left hand pathway).

Examples are presented to show how probabilistic projections of climate can be depicted for different regions of Europe, and how these have been superimposed on impact response surfaces to provide estimates of the risks of runoff exceedance, crop yield shortfall, nitrate leaching and permafrost loss (Morse et al., 2009). Multi-model projections have also been used to characterise the likelihood of future property damage from strong winds. Some examples include consideration of potential future adaptation as well as impact risks avoided through mitigation, using the ENSEMBLES E1 "peak and decline" emissions scenario.

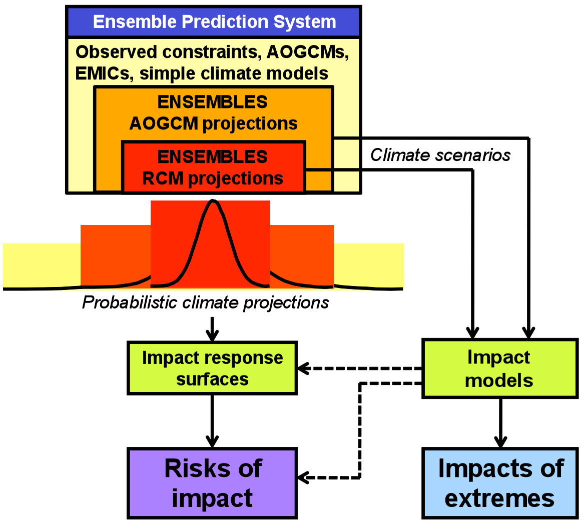

Ensemble Prediction System
Observed constraints, AOGCMs, EMICs, simple climate models
ENSEMBLES AOGCM projections
ENSEMBLES RCM projections
Climate scenarios
Probabilistic climate projections
Impact response surfaces
Impact models
Risks of impact
Impacts of extremes

Figure 1. Two methods of impact assessment using outputs from the Ensemble Prediction System: either using model-based scenarios applied to impact models for estimating impacts of extreme events or using probabilistic projections combined with impact response surfaces for evaluating impact risks (Morse et al., 2009).

References

Jones, R.N., 2000: Managing uncertainty in climate change projections – issues for impact assessment. *Climatic Change*, **45**, 403–419.

Morse, A., C. Prentice, and T. Carter, 2009: Assessments of climate change impacts [Research Theme 6], Chapter 9 in: van der Linden P. and J.F.B. Mitchell (eds.), op cit. ref. 3, pp. 107-129.

Murphy, J.M., D.M.H. Sexton, G.J. Jenkins, et al., 2009: UK Climate Projections Science Report: Climate change projections. Met Office Hadley Centre, Exeter, UK, 190 pp.

van der Linden, P. and J.F.B. Mitchell (eds.), 2009: ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, Exeter, UK, 160 pp.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Weighting Models Based on Several RCM-Specific Metrics: Exploring the Concept

Ole Bøssing Christensen[1], Jens Hesselbjerg Christensen[1], Filippo Giorgi[2], Markku Rummukainen[3]

[1]*Danish Meteorological Institute, Denmark*
[2]*The Abdus Salam International Centre for Theoretical Physics, Italy*
[3]*Rossby Centre, Sweden*

Introduction

An important new development within the European FP6 project ENSEMBLES has been to explore performance-based weighting of regional climate models. Until now an assumption of equal weight has been implicitly applied. At the same time, it is evident that different RCMs give results of a varying degree of realism, e.g. related to extremes. It is not straightforward to construct, assign and combine metrics of model performance. Rather, there is a considerable degree of subjectivity both to the definition of the metrics and to the way these are combined into weights. This does not mean that weighting, however exploratory, would not be meaningful. Rather, it stresses that the assumptions and choices behind the weights need to be recognized and taken into account. Here we discuss the applicability of combining a set of six specifically designed RCM weights to produce one model index producing combined climate change information from the range of RCMs used within ENSEMBLES.

An important principle for the regional climate modelling efforts within ENSEMBLES was to design and calibrate procedures for use in constructing probabilistic regional climate scenarios. In the prior PRUDENCE project it was realised that the formulation of the RCM pays an almost equal role compared to that of the GCM, at least for summer conditions, when the interior of the model is largely decoupled from the large scale boundary conditions originating from the driving GCM, which are imposed on the RCM throughout the integration (Déqué et al. 2007).

In order to address the uncertainty due to RCM formulation, it then becomes central to choose metrics, which are independent of the driving GCM, yet tracing those climate characteristics, which are important also for the estimation of climate change. The choices made in ENSEMBLES are neither optimal nor fully comprehensive. On the other hand, we think that we have captured some essential climate variables, which both build on the fact that the RCMs being used operate at a higher resolution than the driving GCMs, and at the same time highlight performance believed to benefit from increased resolution.

Six metrics for RCM validation

A set of metrics to generate weights based on model performance when compared to observations, was agreed upon in ENSEMBLES. These metrics produce weights of individual RCMs, and are based on the following set of metrics for reanalysis-driven RCM simulations. In ENSEMBLES the ECMWF ERA-40 reanalysis was used.

ƒ1: Large scale circulation based on a weather regime classification

ƒ2: Meso-scale signal based on seasonal temperature and precipitation analysis;

ƒ3: Probability density distribution match of daily and monthly temperature and precipitation analysis;

ƒ4: Extremes in terms of re-occurrence periods for temperature and precipitation;

ƒ5: Trend analysis for temperature

ƒ6: Representation of the annual cycle in temperature and precipitation

-see ENSEMBLES deliverable D3.2.2 for definitions. Several forthcoming papers in a special issue of Climate Research will deal with these weights.

The suggested metrics cover a range of the so-called added-value measures for dynamical downscaling (meso-scale information, fine scale processes to better capture the PDF as well as extreme events) along with some of the minimal requirements for a model to be assessed as credible (annual cycle of precipitation and temperature, large scale agreement with driving model to capture flow regimes and long time temperature trends).

When appropriate, these metrics are defined for different seasons and both for sub-regions and for the European continent as a whole. The combination of the performance of a particular RCM, over the sub-regions and seasons can be combined into a single weight for each RCM. This is done by a multiplication of the weights f1, f2,…,f6 raised to different powers ni, where all the individual weights have a value between 0 and 1. ni can be chosen as any positive number to weigh the various metrics differently (a value of 0 implies equal weighting). The intention behind this approach is that a high total weight requires high scores in all metrics considered.

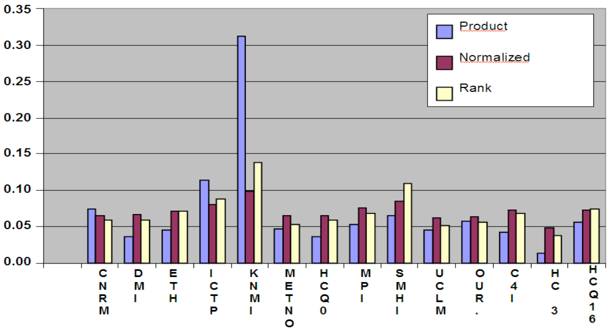


Figure 1. Total weight of ENSEMBLES RCMs calculated from 6 individual weights as a) a product, b) product of factors with a normalised spread, and c) from a product of ranks.

Resulting weighting

Once the individual weights relative to the different metrics are calculated, an overall weight can be calculated for the ENSEMBLES models accessible at http://ensemblesrt3.dmi.dk. Fig. 1 shows total model weights for Europe associated with different ways of calculating the total. There is a significant spread in model weights. The overall largest weight is associated with the model of the KNMI, mostly due to high f2 and f4, (not shown) which has the largest variability across models. Similarly, the model METOHC Q3 shows the lowest weight, mostly due to the contributions from the same functions.

In order to produce probabilistic regional climate scenarios, also the quality of the GCMs used as driving models for the regional climate change runs ought to be considered. The relevant institutes involved in GCM modelling within ENSEMBLES have reached the conclusion that objective weights cannot be constructed and we presently recommend equal weights despite obvious shortcomings. However, we would like to emphasize that downscaling of a poorly performing GCM over Europe will not increase the credibility of the particular GCM results in concern.

Discussion and summary

We have developed a methodology for producing model weights based on the aggregated information of different metrics of model performance. We produced our overall weight from the product of individual performance metrics. This implies a very stringent test, as a well-performing model needs to have relatively high weights in all metrics. Some weights exhibit a much higher inter-model spread than others. For example, the KNMI model has the largest weight mostly because of one particular metric. This effect can be ameliorated by normalizing the weights by the total inter-model spread, or by comparing ranks instead of score values.

It is important to emphasize that our weights are relative; they do not measure the absolute performance of a model, but the relative performance compared to other ensemble models. This subjectivity of the approach calls for an evaluation of the sensitivity of these weights to the criteria used to derive them. This could be accomplished for example by testing the sensitivity to the inclusion of different metrics, the weighting of the different metrics and the way the individual metrics are combined. Such a sensitivity analysis should be performed in order to assess the uncertainty related to the use of a given weighting procedure.

References

Christensen, J.H., T.R. Carter, M. Rummukainen, and G. Amanatidis 2007: Evaluating the performance and utility of regional climate models: the PRUDENCE project. *Clim. Ch.*, **81**, 1-6, doi:10.1007/s10584-006-9211-6.

Christensen, J.H., and O.B. Christensen, 2007: A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Clim. Ch.*, **81**, 7-30, doi:10.1007/s 10584-006-9210-7.

Déqué, M., D.P. Rowell, D. Lüthi, F. Giorgi, J.H. Christensen, B. Rockel, D. Jacob, E. Kjellström, M. de Castro, and B. van den Hurk, 2007: An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections. *Clim. Ch.*, **81**, 53-70, doi:10.1007/s 10584-006-9228-x.

Perturbed Physics Ensembles and Key Metrics of Climate Change for AR5

Mat Collins, Ben Booth, Glen Harris, James Murphy, and David Sexton

Met Office Hadley Centre, United Kingdom

In a companion abstract we describe a method for producing probabilistic projections of climate change conditioned on emissions scenarios (Sexton et al.). The method is based on the Bayesian technique outlined in Rougier (2007) and combines information from HadCM3 perturbed physics ensembles, the CMIP3 multi-model archive, observations and our understanding of climate change to produce projections at global-model scales. In this poster, further aspects of the projections are discussed:

1. A comparison between perturbed physics ensembles (PPEs) and CMIP3 multi-model ensembles (MMEs) in terms of global measures of errors in 2-dimensional time-averaged fields. This component including an investigation of the reasons for the MME ensemble mean always being closer to observations than any individual ensemble member.

2. A comparison of global climate feedbacks and global radiative forcings between PPEs and MMEs.

3. Examples of relationships between errors in models and the magnitude of global feedbacks and the need for multivariate observational constraints.

4. Examples of PDFs for key new and old metrics of climate change; climate sensitivity, transient climate response, Giorgi-region temperature and precip changes, probabilities of crossing policy-relevant temperature and CO_2 targets, etc.

References

Collins, M., B.B.B. Booth, B. Bhaskaran, G. Harris, J.M. Murphy, D.M.H. Sexton, and M.J. Webb, A comparison of perturbed physics and multi-model ensembles: Model errors, feedbacks and forcings. *Climate Dynamics*, submitted.

Collins, M., G.R. Harris, D.M.H. Sexton, J.M. Murphy, and B.B.B. Booth, 2010: Probabilistic projections of climate change for key variables. In prep.

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. *Natural Hazards and Earth System Sciences*, submitted.

Murphy, J.M., D.M.H. Sexton, G. Jenkins, P. Boorman, B. Booth, K. Brown, R. Clark, M. Collins, G. Harris, and E. Kendon, 2009: UKCP09 Climate change projections. ISBN 978-1-906360-02-3.

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

Sexton, D.M.H., B.B.B. Booth, M. Collins, G. Harris, and J.M. Murphy, Keynote: Projections of climate change at global and regional scales. This meeting.

Uncertainty Propagation from Climate Change Projections to Impacts Assessments: Water Resource Assessments in South America

Hideo Shiogama[1], †Seita Emori[1, 2], Naota Hanasaki[1], Manabu Abe[1], Yuji Masutomi[3], Kiyoshi Takahashi[1], and Toru Nozawa[1]

[1] *National Institute for Environmental Studies, Japan*
[2] *Center for Climate System Research, University of Tokyo, Japan*
[3] *Center for Environmental Science in Saitama, Japan*

Generally, climate change impact assessments are based on climate change projections from coupled Atmosphere Ocean General Circulation Models (AOGCMs). Therefore, uncertainties of climate change projections propagate to impact assessments, and affect subsequent policy decisions. In this study, we applied a statistical method, Maximum Covariance Analysis (Singular Value Decomposition analysis), to analyze the uncertainty propagation. Most of the impact assessment studies investigate the uncertainty propagation from local climate changes to local impact assessments. However, large scale patterns of climate change can significantly influence local impact assessments. Performances of AOGCM simulations of local present climate are often evaluated to constrain uncertainties of climate change projections and impact assessments. However, uncertainties of large scale patterns of climate change projections cannot be constrained by local climate metrics. Therefore, we examined a covariance matrix between inter-model uncertainties of local impact assessments and those of large scale climate change projections.

In this study, we examined a water resource impact assessment in South America. This impact assessment was performed by Shiogama et al. (2009). Input data of the water resource model were changes in temperature (ΔT) and precipitation (ΔP) of 14 AOGCMs under the SRES A2 emission scenario. Outputs were 14 simulations of changes in the annual mean runoff (ΔR). We decomposed the inter-model covariance matrix between ΔR in South America and the combination of ΔT and ΔP in the world. Before the decomposition, ΔR, ΔT and ΔP were normalized by the global mean temperature change of each AOGCM. The top panels of Figure 1 show the 1st modes of ΔR, ΔT and ΔP. The 1st ΔR mode was found to have a north-south pattern. This ΔR mode correlates with an El Niño like warming pattern, suggesting that models with stronger El Niño like warming tend to have this north-south ΔR pattern. The bottom panels of Figure 1 show the 2nd modes. The 2nd ΔR mode was found to have an east-west pattern. This pattern of ΔR is associated with the patterns of more warming in the Northern Hemisphere and less warming in the Southern Hemisphere, and a northward movement of the Intertropical Convergence Zone.

This statistical analysis technique enables us to determine what kind of uncertainties of large scale climate change projections lead to uncertainties of local impact assessments. Furthermore, we can apply this technique to examine whether the uncertainties of impact assessment significantly correlate with biases of the present climate simulations. We computed regression patterns between the expansion coefficient of the ΔR modes and the present climate simulations. It was found that the 1st and 2nd ΔR modes were associated with intensities of the Walker circulation and the Hadley circulation in the present climate simulations, respectively. We also showed that the ensemble mean assessment of ΔR had significant biases in the present climate simulations. It is suggested that a standard deviation of +1 of the expansion coefficients of the 1st and 2nd ΔR modes gives minimum biases. In the Amazon region, decreases of runoff were more reliable assessments than increases of runoff, although the ensemble mean impact assessment indicated wetting. This finding has great implications for the carbon cycle feedback, although we did not evaluate uncertainties of carbon cycle models or water resource models.

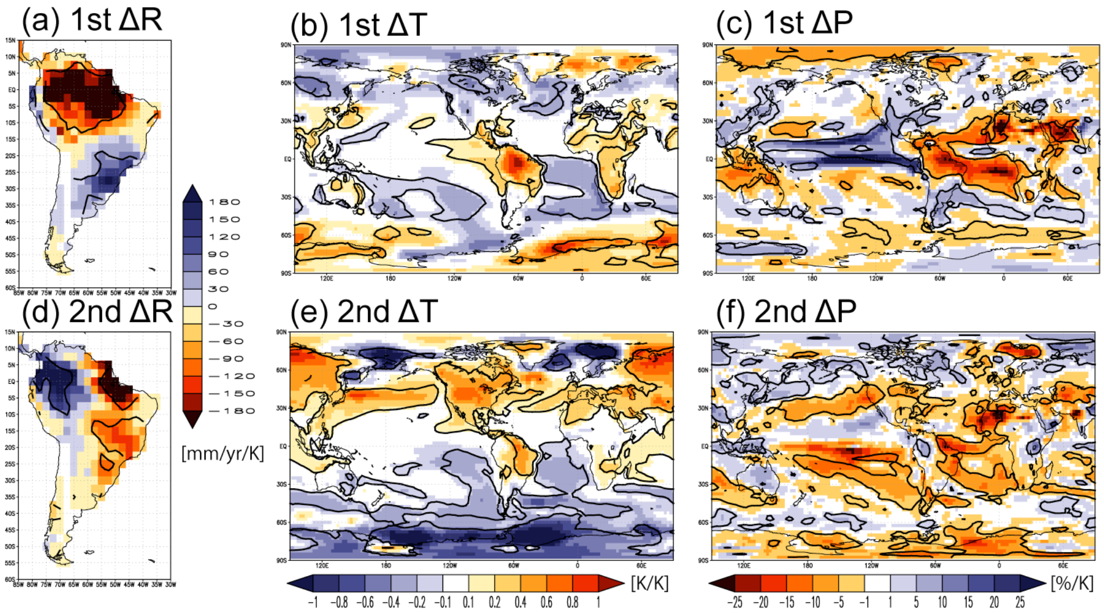


Figure 1. Top panels show the 1st modes of (a) ΔR [mm/yr/K], (b) ΔT [K/K] and (c) ΔP [%/K]. Contours indicate statistical significant heterogeneous correlations at ±10% levels of *t*-test. Bottom panels are the same as the top panels, but for the 2nd modes.

References

Shiogama, H., N. Hanasaki, Y. Masutomi, T. Nagashima, T. Ogura, K. Takahashi, Y. Hijioka, T. Takemura, T. Nozawa and S. Emori, 2009: Emission scenario dependencies in climate change assessments of the hydrological cycle. *Clim. Change Lett.*, accepted.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Quantitative Performance Metrics for Earth System Model Simulations

Veronika Eyring[1] and Pierre Friedlingstein[2,3]

[1]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Germany
[2]Institut Pierre-Simon Laplace (IPSL), Laboratoire des Sciences du Climat et de l'Environnement, France
[3]University of Exeter, United Kingdom

1. Introduction

To determine a safe level of greenhouse gas emissions, physical and biogeochemical feedback processes on various time scales have to be understood and projected with quantified uncertainties. A prerequisite to understanding and improving the representation of climate and biogeochemical feedbacks in state-of-the-art Earth system models (ESMs) is the systematic evaluation of the modeled processes through comparisons with observations. Extensive knowledge and experience has been acquired in a variety of international Model Intercomparison Projects (MIPs) that focus on the evaluation and quantification of processes for most if not all components of an ESM. However, an integrated evaluation to assess the performance of ESMs as a whole has so far been lacking. Only an integrated approach, applying a common strategy across the different Earth system components, will yield a realistic quantitative assessment of our ability to represent the various physical and biogeochemical climate feedbacks involved and will allow to explore the value of weighting multi-model climate projections.

2. Framework for ESM Evaluation

The ESM evaluation strategy proposed here is a response to this need. We follow the concept of WCRP's SPARC Chemistry-Climate Model Validation (CCMVal) activity (Eyring et al., 2005). We identify a set of core climate and biogeochemical feedbacks and processes, associated with one or more model diagnostics and with relevant datasets that can be used for the ESM evaluation. This process-based evalution is proposed in addition to the model evalution that focuses on long-term trends and variability in well-observed Essential Climate Variables (ECVs) (GCOS, 2008). Driven by the availability of observations and pre-existing knowledge in modeling of ESM components, diagnostics are developed in three key areas for climate projections: physical climate feedbacks, global carbon cycle feedbacks, and atmospheric composition feedbacks.

Model performance metrics (a statistical measure of agreement between models and observations) can be further developed to allow a quantitative assessment of the performance for all ESM components in an integrated way. This approach will also enable the documentation of model improvements, for different versions of individual models and for different generations of community-wide models used in international assessments. At the same time, the diagnostics themselves should develop as experience is gained and as new measurements become available.

3. Examples: CCMVal and ILAMB

The approach described above for ESMs has been succesfully applied to 13 chemistry-climate models (CCMs) participating in CCMVal in support of the 2006 WMO/UNEP Scientific Assessment of Ozone Depletion (WMO, 2007). The starting point was the study by Eyring et al. (2006), who evaluated a subset of key processes important for stratospheric ozone in the CCMs. Waugh and Eyring (2008) applied quantitative performance metrics to the same diagnositcs and models and assigned a quantitative measure of performance ("grade") to each model-observations comparison. Theses grades were then used to assign relative weights to the CCM projections of 21st century total ozone previously presented without any weights. For the limited set of processes, diagnostics and according performance metrics that was used in this study there were generally only small differences between weighted and unweighted multi-model mean and variances of total ozone projections, suggesting that the multi-model mean was a robust quantity in CCMVal-1 simulations. This study raises several issues with the grading and weighting that need further examination. However, it provides a framework and benchmarks that enables quantification of model improvements and assignment of relative weights to the model projections.

A more recent example comes from the International Land Model Benchmarking (ILAMB) under the AIMES programme of IGBP. ILAMB builds on previous independent evaluations of water, energy and carbon fluxes (e.g., Randerson et al., 2009) to evaluate the Dynamic Global Vegetation Models (DGVMs) that are now used in ESMs. Combination of satellite, atmospheric, and surface datasets, spanning from the recent decades to the full 20th century allow to evaluate the models on seasonal, interannual and centenial time scale. Of particular interest in the context of future climate-carbon cycle feedback is the development of metrics based on, among others, the interannual growth rate of atmospheric CO_2, a direct measure of the land tropical ecosystems to climate variability (Cadule et al., 2009).

4. Summary and way ahead

A framework for a process-oriented evaluation of Earth System Models (ESMs) with a focus on the models' ability to project Earth System feedbacks and change throughout the 21st century and beyond is currently been developed. Each feedback is associated with the key processes that detemine it along with diagnostics and observational datasets that can be used for the evaluation. The challenging approach of applying the framework to state-of-the-art ESMs (e.g., those participating in CMIP5) is beyond the scope of national or local activities and relies on broad support from the modeling and observational community. Global data sets relating to key properties of the atmosphere, ocean, land and cryosphere will need to be assembled so that outputs from individual component models, and from coupled ESMs, can be confronted with these data in a consistent and quantitative way. A synoptic view on different ESMs will allow the community to identify common gaps in modeling quality and to highlight particularly well or poorly modeled processes. This will allow attributing specific model behaviour to specific process parameterisations and paving the way for a systematic model improvement in the future or for new observation strategies needed to better constrain ESMs. The integrated results of this approach are expected to lead to long-term improvements of ESMs and thereby to enhanced confidence in climate projections.

References

Cadule, P., et al., 2009: Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements. *Glob. Biogeochem. Cycl.*, in press.

Eyring, V., et al., 2005: A strategy for process-oriented validation of coupled chemistry-climate models. *Bull. Am. Meteorol. Soc.*, **86**, 1117-1133.

Eyring, V., et al., 2006: Assessment of temperature, trace species and ozone in chemistry-climate model simulations of the recent past. *J. Geophys. Res.*, **11**(D22308), doi:10.1029/2006 JD007327.

GCOS, 2007: Systematic Observation Reuirements for Satellite-based Products for Climate – Supplemental Details to the GCOS Implementation Plan. GCOS 1007, WMO/TD No. 1338.

Randerson, J., et al., 2009: Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models. *Glob. Change. Biolog.*, **15**, 2462-2484.

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

WMO/UNEP, 2007: *Scientific Assessment of Ozone Depletion: 2006*, World Meteorological Organization, Global Ozone Research and Monitoring Project, Report No. 50, Geneva, Switzerland.

Marine Biogeochemical Model Comparison Efforts: Tools for Assessing Model Skill

Marjorie A.M. Friedrichs

Virginia Institute of Marine Sciences, College of William & Mary, USA

Although application of marine biogeochemical models to the study of carbon cycling and global climate is becoming increasingly common, objective quantification of the relative performance of these models is rare. Recently, a community-wide model intercomparison exercise (Friedrichs et al., 2006; 2007) was conducted in order to assess the relative skill of twelve lower trophic level marine ecosystem models of varying complexity. In order to isolate the effects of the different biogeochemical parameterizations, each of the twelve models was run within an identical physical framework. The key component of this exercise was the use of a consistent variational adjoint implementation in which chlorophyll, nitrate, export, and primary productivity data were assimilated. This insured that the different model structures were being objectively compared, rather than the degree of tuning. Finally, identical metrics were used to assess model skill. In particular, model skill was assessed by comparison to the skill of the mean of the observations. Experiments were performed in which no data were assimilated (Expt. 1), data were assimilated from individual sites (Expt. 2) and from two sites simultaneously (Expt. 3). A cross-validation experiment (Expt. 4) was also conducted whereby data were assimilated from one site, and the resulting optimal parameters were used to generate a simulation for the second site.

Results from this marine biogeochemical comparison exercise revealed that when a single pelagic regime was considered, the simplest models could be tuned to fit the data as well as those with multiple phytoplankton functional groups. However, models with multiple phytoplankton functional groups produced lower misfits when the models were required to simulate both regimes using identical parameter values. The cross-validation experiments revealed that as long as only a few key biogeochemical parameters were optimized, the models with greater phytoplankton complexity were more portable. In contrast, models with multiple zooplankton compartments did not generally outperform models with single zooplankton compartments, even when zooplankton biomass data were assimilated. The results of this model comparison effort highlight the importance of using formal parameter optimization techniques, as well as using identical physical circulation fields and consistent metrics for model skill assessment.

A second ongoing biogeochemical model inter-comparison effort has concentrated on the ability of models to estimate depth-integrated primary productivity (PP). Estimates of PP are routinely generated from satellite ocean color based models and are also now available from biogeochemical ocean general circulation models. Calibration and validation of these PP models are not straightforward, however, and comparative studies show large differences between model estimates. Friedrichs et al. (2009) compare PP estimates obtained from 30 different models to a tropical Pacific PP database consisting of ~1000 14C measurements spanning more than a decade (1983-1996). Model skill was compared using both Taylor Diagrams (Taylor, 2001) and the recently introduced Target Diagrams (Jolliff et al., 2009), which more clearly emphasize model bias (Figure 1.) Target diagrams are particularly useful when large numbers of models are being compared, and bias is a large component of total model error. In addition, the inclusion of two reference isolines on the Target Diagram (the observational uncertainty and the root-mean-squared difference (RMSD) of the mean of the observations) makes these plots a particularly useful tool for model skill assessment.

Primary findings of this PP model intercomparison effort (Figure 1) include: (1) skill varied significantly between models, but performance was not a function of model complexity or type (i.e. satellite-based model vs. circulation/biogeochemical model), (2) nearly all models underestimated the observed variance of PP, specifically yielding too few low PP values, and (3) roughly half of

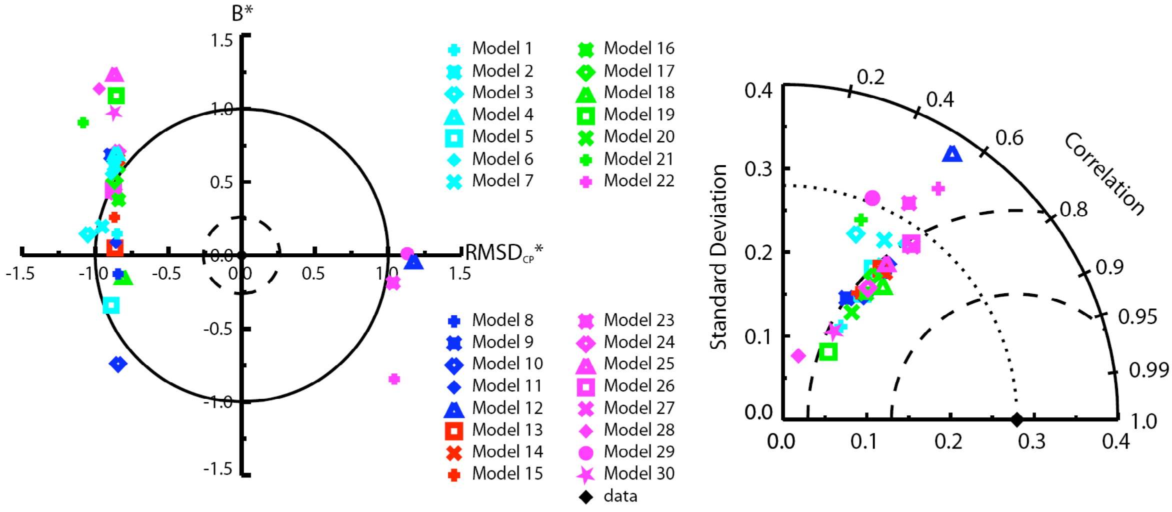


Figure 1. (a) Target diagram for log(PP) displaying Bias normalized by the standard deviation of the PP data (B*) and normalized centered-pattern RMSD (RMSD$_{CP}$*) for the 30 participating models relative to the tropical Pacific database. Concentric circles represent isolines of normalized total RMSD (RMSD*): the inner dashed circle represents the normalized observational PP uncertainty, and the outer solid circle represents the RMSD* of the productivity data. (b) Taylor diagram of log(PP). The distance from the origin is the standard deviation of the modeled productivities. The azimuth angle represents the correlation between the observations and the modeled productivities, and the distance between each model symbol and the data (black diamond) is the RMSD$_{CP}$. Dashed lines are isolines of RMSD$_{CP}$ = 0.25 and RMSD$_{CP}$ = 0.15. Dotted line represents the standard deviation of the data. Adapted from Friedrichs et al., 2009.

the models performed better than the mean of the pbservations. Additional analyses revealed that more than half of the total RMSD associated with the satellite-based PP models might be accounted for by uncertainties in the input variables and/or the PP data. Finally, average RMSD between in situ PP data and PP estimates from the satellite-based productivity models were 58% lower than analogous values computed in a previous PP model comparison seven years ago. The success of these types of comparison exercises is illustrated by the continual modification and improvement of the participating models and the resulting increase in model skill.

References:

Friedrichs, M.A.M., M.-E. Carr, R. Barber, M. Scardi, D. Antoine, R.A. Armstrong, I. Asanuma, M.J. Behrenfeld, E.T. Buitenhuis, F. Chai, J.R. Christian, A.M. Ciotti, S.C. Doney, M. Dowell, J. Dunne, B. Gentili, W. Gregg, N. Hoepffner, J. Ishizaka, T. Kameda, I. Lima, J. Marra, F. Mélin, J.K. Moore, A. Morel, R.T. O'Malley, J. O'Reilly, V.S. Saba, M. Schmeltz, T.J. Smyth, J. Tjiputra, K. Waters, T.K. Westberry, and A. Winguth, 2009: Assessing the uncertainties of model estimates of primary productivity in the tropical Pacific Ocean. *J. Mar. Sys.,* **76**, doi:10.1016/j.marsys. 2008.05.010.

Friedrichs, M.A.M., J.A. Dusenberry, L.A. Anderson, R.A. Armstrong, F. Chai, J.R. Christian, S.C. Doney, J. Dunne, M. Fujii, R. Hood, D.J. McGillicuddy, J.K. Moore, M. Schartau, Y.H. Spitz, and J.D. Wiggert, 2007: Assessment of skill and portability in regional marine biogeochemical models: Role of multiple phytoplankton groups. *J. Geophys. Res.*, **112**, C08001, doi10:1029/2006JC003852.

Friedrichs, M.A.M., R. Hood, and J. Wiggert, 2006: Ecosystem model complexity versus physical forcing: Quantification of their relative impact with assimilated Arabian Sea data. *Deep-Sea Res. II*, **53**, 576-600.

Jolliff, J., J.C. Kindle, I. Shulman, B. Penta, M.A.M. Friedrichs, R. Helber, R. A. Arnone, 2009: Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. *J. Mar. Sys.,* doi:10.1016/j.marsys.2008.05.014.

Taylor, K.E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183-7192.

Upgrades to the REA Method for Producing Probabilistic Climate Change Projections

Ying Xu[1], Xuejie Gao[1, 2], Filippo Giorgi[2]

[1]National Climate Center, China Meteorological Administration, China
[2]International Centre for Theoretical Physics, Italy

We present an augmented version of the Reliability Ensemble Averaging (REA) method designed to generate probabilistic climate change information from ensembles of climate model simulations. Compared to the original version, the augmented one includes consideration of multiple variables and statistics in the calculation of the performance-based weights. In addition, the model convergence criterion previously employed is removed. The method is applied to the calculation of changes in mean and variability for temperature and precipitation over different sub-regions of East Asia based on the recently completed CMIP3 multi-model ensemble. Comparison of the new and old REA methods, along with the simple averaging procedure, and the use of different combinations of performance metrics shows that at fine sub-regional scales the choice of weighting is relevant. This is mostly because the models show a substantial spread in performance for the simulation of precipitation statistics, a result that supports the use of model weighting as a useful option to account for wide ranges of quality of models. The REA method, and in particular the upgraded one, provides a simple and flexible framework for assessing the uncertainty related to the aggregation of results from ensembles of models in order to produce climate change information at the regional scale.

The Limited Contribution of Model Uncertainty to the Uncertainty in Observationally-Constrained Estimates of Anthropogenic Warming

Nathan Gillett[1] and Peter Stott[2]

[1] Canadian Centre for Climate Modelling and Analysis, Environment Canada,Canada
[2] Met Office Hadley Centre, United Kingdom

Output from multiple climate models is often used in studies which attempt to constrain past and future anthropogenic warming using observational constraints. Recent studies have used output from multiple models first to obtain a less noisy estimate of the anthropogenic response, since it is anticipated that different models will exhibit different errors in their response patterns. Second, these studies have used inter-model differences to account for model uncertainty in estimates of the uncertainty in anthropogenic warming, using the Error in Variables approach. Here we show that explicitly accounting for model uncertainty in this way only marginally inflates the uncertainty in estimates of anthropogenic warming. We suggest that this is because inter-model differences in the magnitude of the anthropogenic response are disregarded in the analysis, since the magnitude is constrained by observations; and overall uncertainty in observationally-constrained anthropogenic warming is dominated by internal variability in the observations, with model uncertainty in the pattern of the response making only a small contribution.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

A World Climate Research Programme (WCRP) Panel Tasked to Identify and Promote Performance Metrics for Climate Models

Peter Gleckler[1], Beth Ebert[2], Veronika Eyring[3], Robert Pincus[4], Karl Taylor[1], and Richard Wood[5]

[1]*Program for Climate Model Diagnosis and Intercomparison, USA*
[2]*Centre for Australian Weather and Climate Research, Australia*
[3]*Deutsches Zentrum fuer Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Germany*
[4]*Earth System Research Laboratory, Boulder, USA*
[5]*Met Office, United Kingdom*

The Working Group on Numerical Experimentation (WGNE), jointly established by the WCRP Joint Scientific Committee and the WMO Commission for Atmospheric Sciences, has the responsibility of fostering the development of atmospheric circulation models for use in weather, climate, water and environmental prediction on all time scales and diagnosing and resolving shortcomings. The WGNE has established a panel to identify a well-defined set of performance metrics for climate models to objectively gauge the strengths and weaknesses of different models and to track improvement as models are further developed. The panel members have been selected according to their relevant scientific contributions and membership or active liaison efforts in key WCRP activities (P. Gleckler, WGNE; B. Ebert, Joint Working Group on Forecast Verification Research, JWGFVR; V. Eyring, the Working Group for Coupled Models, WGCM, and Stratospheric Processes and their Role in Climate project, SPARC; R. Pincus, Global Energy and Water Cycle Experiment, GEWEX; K. Taylor, WGCM; and R. Wood, Working Group on Ocean Model Development, WGOMD).

The WGNE metrics panel is working to coordinate the development of a hierarchy of climate model performance metrics. At the most basic level will be a limited set of traditional "broad-brush" statistical measures that gauge model quality against well observed quantities. This will be further developed into a more extended set of metrics targeted towards quantifying skill in simulating key processes. Identifying this more extensive set of metrics will require the WGNE metrics panel to engage with other WCRP activities that are currently developing more specialized metrics, e.g., for key modes of variability and important climate processes associated with the atmosphere, land, ocean and sea-ice. The panel will define its suite of performance metrics for well-established WCRP benchmark experiments such as the Coupled Model Intercomparison Project (CMIP) historically forced "20th Century" experiment, prescribed SST (AMIP) simulations, and the WGNE transpose-AMIP. An expansion to include biogeochemical processes for CMIP ESM experiments is envisaged but will likely require additional panel member to consider carbon cycle and atmospheric composition climate feedbacks.

Over the course of the next year, the WGNE metrics panel will identify its standard set of performance metrics that are: based on comparison with carefully selected observations; easy to calculate, reproduce and interpret; established in the peer-reviewed literature; covering a diverse suite of climate characteristics; emphasizing large- to global-scale measures of mean climate (and limited variability) for the atmosphere, oceans, land surface, and sea-ice. The panel will oversee the development of software and collection of observational datasets to calculate these metrics and make them publicly available. The resulting capability will be applied to the next phase of CMIP (i.e., CMIP5) and results will be compared with earlier model versions. These results will be made publicly available as CMIP5 simulations are submitted to the archive, i.e., as the research phase of CMIP5 is just beginning. One goal of this activity is to ensure that any new climate model simulations introduced in the scientific literature or made available to the research community will be tested against an expected set of routine benchmarks.

The quest for a defensible approach to weight projections from individual models in a multi-model ensemble remains elusive, and is beyond the purview of the WGNE metrics panel. However, the panel may choose to identify a set of minimum performance standards, or make broad recommendations regarding the weighting of model projections. Furthermore, depending on the success of research related to multi-model projections in the coming years (e.g., exploring the relationship between well-observed features of climate to key physical feedbacks), it may be appropriate at a later date to incorporate climate-change related performance metrics into the WGNE metrics hierarchy. Possible actions related to the model weighting challenge, such as these, remain an active area of panel discussions.

Assessment of TCR and Ocean Heat Uptake Efficiency

Jonathan Gregory

University of Reading, United Kingdom

Assessment of reliability of AOGCMs for use in projections is mainly based on the realism of their simulation of present-day mean climate. However, the huge improvement in present-day climate simulation resulting from climate model development over the last twenty years has not been accompanied by a commensurate reduction of model systematic uncertainty in climate projections, suggesting that response to forcing is determined by processes which are not strongly constrained by observations of mean climate. If so, assessment of models ought to be based on their simulation of time-dependent past climate change (which means of course that realistic simulation of the past would no longer be an independent check on the reliability of models). Gregory and Forster (2008) demonstrated a linear relationship between radiative forcing and global-mean surface air temperature change during recent decades in observations and the HadCM3 AOGCM. The slope of this relationship (the climate resistance, in W m^{-2} K^{-1}) results from climate feedback and ocean heat uptake efficiency, and it determines the transient climate response to CO_2 increase, which might thus be constrained from observations. We explore the robustness of the constant climate resistance (or TCR) as a way of making climate projections from radiative forcing. We look at whether the ocean heat uptake efficiency is scenario-independent and how well it can be constrained from observations. Heat uptake is particularly important for projections of sea-level rise due to thermal expansion, which depends also on the expansion efficiency of heat (a measure of how much thermal expansion results from a given addition to heat content), another quantity which might also be assessed from observations.

ENSO and Tropical Pacific Metrics for Coupled GCMs

Eric Guilyardi[1], Andrew Wittenberg[2], and the ENSO metrics work group of the CLIVAR Pacific Panel

[1]LOCEAN/IPSL, France and NCAS-Climate, University of Reading, United Kingdom
[2]NOAA/GFDL, USA

The wide diversity of El Niño simulations in coupled GCMs contributes to large uncertainties in projections of future tropical climate variability and its global impacts (Meehl et al., 2007; Vecchi and Wittenberg, 2009; Collins et al., 2009). This shortcoming – a major issue in the IPCC AR4 – has helped motivate a new chapter in the upcoming AR5 report, dedicated to ENSO and other modes of climate variability.

Uncertainty in the future of ENSO arises not only from diverse model biases, but also from the diverse and inconsistent metrics used to evaluate ENSO from study to study. To better coordinate future studies, the CLIVAR Pacific Panel asked a group of ENSO experts to propose a set of standard ENSO metrics, to aid in diagnosing and understanding inter-model differences and assessing simulation quality.

Here we present these proposed metrics, which span aspects of the tropical Pacific mean state, annual cycle, and ENSO (Guilyardi et al., 2009). Examples are given of "user profiles," in which some metrics are emphasized depending on the judgement or interests of a particular investigator. Applying the metrics and user profiles to "weight" the various AR4 models' ENSO amplitude responses to elevated CO_2, we find that the future ENSO amplitude projections depend strongly on the chosen user profile. We suggest that given our current state of understanding of ENSO, an important first application for community ENSO metrics may be to identify models that don't pass key performance thresholds.

References

Collins M., S.-I. An, W. Cai, A. Ganachaud, E. Guilyardi, F-F Jin, M. Jochum, M. Lengaigne, S. Power, A. Timmermann, G. Vecchi, and A. Wittenberg, 2009 : The impact of global warming on the tropical Pacific and El Niño. *Nature Geosciences*, submitted

Guilyardi E., A. Wittenberg, A. Fedorov, M. Collins, C. Wang, A. Capotondi, G.J. van Oldenborgh, and T. Stockdale, 2009: Understanding El Niño in ocean-atmosphere general circulation models: progress and challenges. *Bull. Amer. Met. Soc*, **90**, 325-340.

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. *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* [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.

Vecchi, G.A., and A.T. Wittenberg, 2009: El Niño and our future climate: Where do we stand? *Wiley Interdisciplinary Reviews: Climate Change*, in press.

A Strategy to Improve Projections of Arctic Climate Change

Alex Hall

Department of Atmospheric Sciences, University of California Los Angeles, USA

Here we describe our recent efforts to constrain the simulated response of the Arctic to anthropogenic forcing with observations. In the AR4 models, we show that the spread in simulated Arctic climate response is determined by the longwave component of an Arctic-specific climate feedback parameter. This negative longwave feedback is in turn controlled by the strength of the wintertime temperature inversion in the current climate, with strong (weak) inversions giving large (small) negative feedback. A comparison with reanalysis and satellite data reveals that the atmospheric temperature inversion is unrealistically strong in most models, indicating that these models simulate excessive negative longwave feedback in the Arctic. Further analysis of the observed and simulated relationships between sea ice concentration and inversion strength shows that many models mishandle the effects of atmosphere-to-ocean heat fluxes through the ice pack on the atmospheric boundary layer, generating systematic errors in inversion strength. The conclusion of this research is that model development efforts to improve the models' Arctic response to anthropogenic forcing should focus on the polar atmospheric boundary layer.

Use of Multimodel Results for Detection and Attribution

Gabriele Hegerl

University of Edinburgh, United Kingdom

Data from large climate model ensembles have proven very useful for detection and attribution. Several examples of using multimodel information are given:

- The causes of European climate variability over the last 5 centuries have been investigated using three individual simulations with coupled climate models that differ somewhat in their forcing. The effect of external forcing on winter and spring temperature can be detected, and the best estimate indicates that a lot of the winter warming since the Little Ice Age was due to anthropogenic forcing. Solar forcing may have influenced summer temperatures, and both summer and winter show significant short term response to volcanic eruptions. The level of agreement and disagreement between the models helps to understand the role of forcings on seasonal temperatures. Forcing uncertainty plays a role in results for summer, where a model with aerosol forcing seems to indicate much less overall warming than one without, and the effect of external forcing is marginally detectable on summer temperatures only before 1900.

- The multimodel archive can be used for a simple approach to attribute changes in the number of warm nights in recent decades to external forcing. The results indicate that changes in the number of warm nights are detectable over many regions worldwide both individually for regions and on average. The variability between individual ensemble members is used as a conservative estimate of internal climate variability.

- Detection of zonal precipitation trends could only be achieved using the multimodel archive, with a large numbers of simulations helping to cancel the large variability generated within the climate system. However, the size of the signal was substantially larger in observations than simulations. It is unclear what the contribution to this discrepancy is from observational uncertainty, from forcing uncertainty (for example, the role of aerosols in subtropical drying) and from inter-model differences. Addressing intermodel differences may require more creative approaches than simple model averaging for a best estimate of the forced change.

In conclusion, the use of ensembles which differ in model characteristics have proven very helpful for detection and attribution despite difficulties in interpretation posed by their inhomogeneity. Ensembles of models that also use a different realization of forcing can help to span the uncertainty in forced climate change more fully than model simulations with identical forcing and have clearly proven useful, but care has to be applied when using them. Attribution methods can help to investigate the role of individual forcings and address the possibility of spurious agreement between models simulations and observations

References

Hegerl, G.C., J Luterbacher, F. Gonzalez-Ruoco, S.F.B Tett, and E. Xoplaki, 2010: Influence of external forcing on European temperatures, nearly submitted.

Zhang, X., F.W. Zwiers, G.C. Hegerl, N. Gillett, H. Lambert, S. Solomon, P. Stott, and T. Nozawa, 2007: Detection of Human Influence on 20th Century Precipitation Trends. *Nature*, **468(**448), 461-466.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Thoughts on the Use of Multi-Model Ensembles

Isaac Held

NOAA Geophysical Fluid Dynamics Laboratory, USA

It is suggested that the focus of IPCC should not be on designing methods for the use of multi-model information, but that it should ideally have standard tests for assessing whether methods suggested in the literature are superior to simple averages. The importance of "predicting the future of models" is emphasized as the key ingredient for this purpose, and more generally for designing metrics for climate models. The value of overall performance metrics versus metrics tailored for specific applications is also discussed.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Extracting Information from Regional Multi-Model Climate Change Projections

Bruce Hewitson[1] and Jens H. Christensen[2]

[1]*Department of Environmental & Geographical Sciences, University of Cape Town, South Africa*
[2]*Danish Climate Centre, Danish Meteorological Institute, Denmark*

At the root of responding to climate change is the fact that impacts occur on regional and local scales, requiring robust messages of regional climate change. It is well recognized that the resolution of global climate models (GCMs) has not been adequate to meet the regional scale information needs. Consequently both the IPCC TAR and AR4 included chapters on developing regional messages of change, but with only partial success, and which still drew heavily on GCM results augmented by limited regional downscaling results. An important element for this was the need for a homogenised assessment procedure across WG-I, thus the relatively late availability of the WCRP CMIP3 multi-model database prevented updated large scale downscaling efforts to be ready in time for the inclusion in AR4. For the same reason, results such as recently achieved within the ENSEMBLES project focusing on SRES scenario A1B will appear out of date by the time AR5 is completed.

Four information sources were identified in the AR4 Ch11 for deriving regional messages of change; "AOGCM simulations; downscaling of AOGCM-simulated data using techniques to enhance regional detail; physical understanding of the processes governing regional responses; and recent historical climate change." Of these, the use of downscaling techniques has recently made notable advances (although all four information sources remain poorly integrated). The developments of multi-model regional experiments such as initiated within ENSEMBLES, and in particular expanded by planned activities that broadly fall within the WCRP CORDEX initiative, offer a valuable new opportunity to draw on regional resolution data for all terrestrial regions in a way that was not broadly possible at the time of previous IPCC assessment reports.

Deriving information from such multi-method regional downscaling experiments in part mimics the challenges of assessing the data from GCMs, such as is found in the WCRP CMIP3 multi-model database, but in part has its own unique challenges. Recognizing that data is not information, three primary issues need to be addressed in relation to translating output data from a given model: what is the relative regional skill of each contributing model, what are the relative signals of natural and forced variability represented in the models, and understanding the limits of spatial detail that can possibly be represented. Current approaches are partially successful in addressing these, albeit more strongly focused on specific grid-cell biases. Note that the definition of skill is not always well defined at the regional to local level. Following this are the methodological challenges to reaching a first order message of regional change, and of representing the envelope of possible climate response. Approaches range from simple averages and ranges, through to more sophisticated methods to assess probability envelopes (for example, Bayesian techniques). Still poorly resolved is the question of how well the models span the true probability space.

More unique to the regional downscaling methods are questions of assessing the real-world local scale variance from the grid cell average information of models in order to support the climate Vulnerability, Impact and Adaptation (VIA) communities which typically are used to point-wise station observation information. . Central to this is the lack of detailed long term observations in many parts of the world. Similar is the challenge of integrating statistical downscaling results with that of dynamical climate model output. As one increases in spatial resolution, the nuances and subtleties of the task increase, especially when taken with a view to informing the activities of the VIA communities. While progress is being made on the above issues, perhaps the biggest potential for enhancing information from the multi-model output is through new approaches to identify signal versus noise (natural variability), and integrating this with an understanding of the changes in the large scale and regional-scale processes that drive the local climate.

This talk uses examples from past and current work to illustrate the complexity of the challenge, as well as the potential for developing stronger regional-scale messages of climate change relevant to end-users.

Enhancing Regional Climate Change Messages Through Assessing Multi-Model Regional Circulation Changes

Bruce Hewitson

Department of Environmental & Geographical Sciences, University of Cape Town, South Africa

The development of regional climate change messages has traditionally focused on grid cell data from climate models, and especially through downscaling global models with either dynamical or statistical methods. The changes in the regional atmospheric processes, especially on the daily time scale of weather events, remains weakly explored, yet offers a valuable additional source of information to assess the robustness and value of local downscaled projections. Changes in the driving processes can occur through changes in seasonal timing, frequency, and intensity, as well as in the occurrence of atmospheric states new to the region. Consistency between these changes and the local surface variables that are more traditionally evaluated can potentially give new insight into the projections. Self Organizing Maps (SOMs) are used to characterize the continuum of weather events based on reanalysis data of atmospheric circulation variables. Multi-model climate change projections of these variables, using data from the CMIP3 archive, are mapped to the SOM to allow for the assessment of changes in occurrence frequency and changes in the mean state of the characteristic weather events in response to future global climate change. The consistency between these changes and the local downscaled surface variable changes are assessed for physical consistency to evaluate the robustness of projected local climate change.

Metrics and Likelihood

Charles Jackson

University of Texas at Austin, USA

A discussion is presented of the rationale for making use of metrics to weight multi-model projections of climate. The multi-model "ensemble of opportunity" that has been helpful in documenting spread among plausible simulations of the climate system, assumes ensemble members are independent and broadly representative of climate model development uncertainty. Each member of this ensemble is viewed as being equally plausible. The AR5 ensemble will include many more samples from individual models and the independence among modeling efforts is becoming less clear. The question that will be discussed below is whether it is appropriate to apply metrics within likelihood weights of AR5 ensemble runs as a way of improving the objectivity in multi-model evaluations of change.

Metrics quantify a measure of skill a model has to reproduce what is observed. They are used to aid climate model development, with a well-recognized caveat that it is often difficult to improve one metric without degrading other metrics. Uncertainties arising from climate model biases and natural variability make it impossible to select a single preferred model configuration. With a common set of observations as targets for model development, we should expect, statistically speaking, the largest proportion of new models (as selected by independent climate experts) would be configurations that contain many smaller compensating errors and smaller proportion of model configurations with a few large compensating errors. Assuming modeling errors are Gaussian, the probability density of these acceptable models is an exponentially decreasing function of the metrics that are used to evaluate the models as normalized by judgments of acceptability. If the model development process were truly independent and a great many models were available to select from, there would be no need to apply weighting as the likelihood of any given model configuration would be apparent and appropriately accounted for within the distribution of model results. Each climate model would, in effect, be pre-weighted by the use of model metrics within the climate model development process.

Samples from the AR5 ensemble will not be independent nor are they likely to span the space of possibilities. The AR5 ensemble will contain biases. It is possible to partially bias-correct this distribution by weighting the contribution of individual ensemble members by an exponential of a set of normalized metrics. This requires consideration of the choice of mix of metrics and scientifically grounded judgments of model acceptability. The best practices for making these decisions are not yet clear.

Some additional discussion is warranted with regards to the interpretation and usefulness of metric-based likelihood probability measures of projected climate. Without an improvement in our understanding of how metrics can be useful measures of skill in projecting future climate, it is difficult to claim that the maximum likelihood is also the most probable climate trajectory. What we can say is that the maximum likelihood solution represents models that are consistent, in most respects, with what has been observed. Model configurations that reside within the lower probability tails of model acceptability should still be viewed as being equally valid to other individual models.

Consider the distribution of errors that exist for the NCAR Community Atmosphere Model that arise from differences in reasonable choices of values assigned to six parameters important to clouds and convection (Jackson et al., 2008; Jackson, 2009). For this task, a likelihood function was defined that quantifies the level of agreement between the model and observations within AMIP-style experiments spanning years 1990 to 2001. The likelihood function that quantifies the relative acceptability of proposed parameter sets was based on a normalized multivariate measure of model error that was inflated to accommodate estimates of model bias. This sampling procedure mimics how independent model development efforts might objectively select different model configurations given the same set of observations and model evaluation criteria. The sampling procedure is carried out as a global optimization such that overall modeling errors are reduced over many experiments.

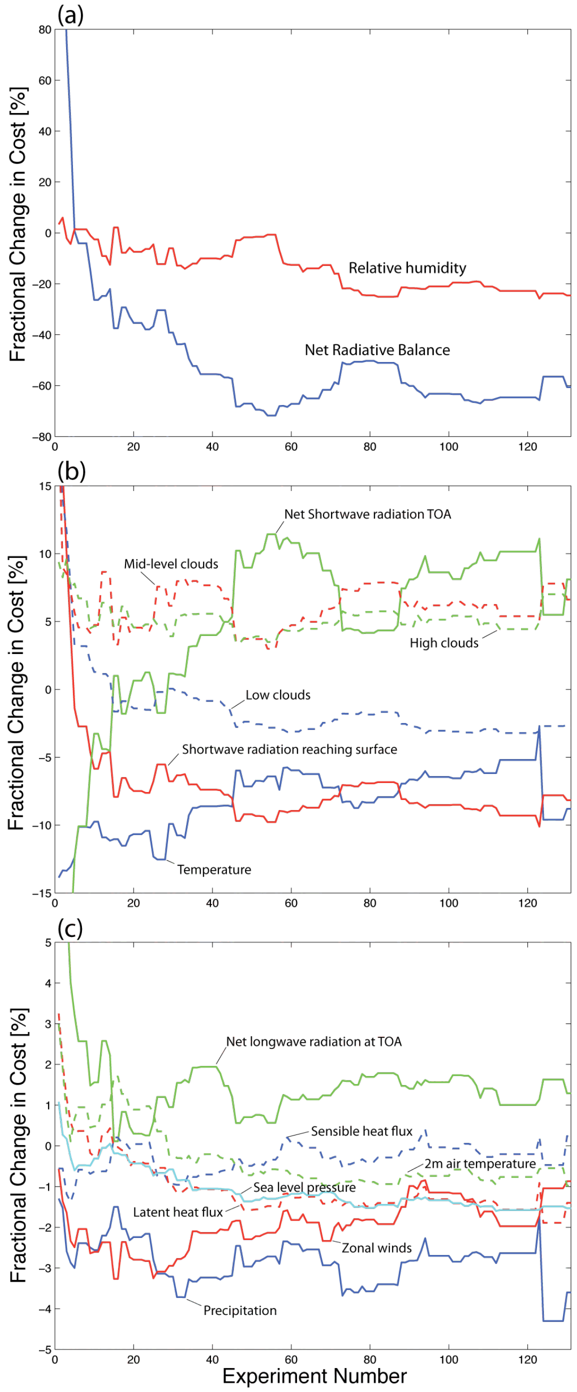


Figure 1. Fractional change in modeling errors that occur arise from selecting different combinations of values for six parameters important to clouds and convection within the NCAR Community Atmosphere Model. Reproduced from Jackson (2009).

Apparent within Figure 1 are fluctuations in modeling errors (i.e., the "cost") that demonstrate the ambiguity in identifying a single optimal model. Models selected after experiment generation 40 contained very similar overall modeling errors, although one can clearly see that the particular choices varied widely in their skill in reproducing individual fields.

It is also hopeful to see within Figure 1 that that many of the fields improved together, implying that the model is successful at representing some of the relationships that govern how different physical quantities relate. One can see this point by noting that fields that dominate changes in modeling errors (in panels a and b) are reduced concomitantly with fields that changed little over the course of sampling (panel c). This skill appears to have its limits with the model depiction of clouds. In particular the improvements that occurred with surface variables, came at the expense of top of the atmosphere shortwave radiative fluxes and clouds.

References

Jackson, C.S., 2009: Use of Bayesian Inference and Data to Improve Simulations of Multi-physics Climate Phenomena. *Journal of Physics, Conference Series 180*, doi:10.1088/1742-6596/180/1/012029

Jackson, C.S., M.K. Sen, G. Huerta, Y. Deng, and K.P. Bowman, 2008: Error Reduction and Convergence in Climate Prediction. *Journal of Climate*, **21**(24), 6698-6709. doi:10.1175/2008JCLI2112.1.

Application-Based Model Discrimination (Arctic Case)

Vladimir Kattsov

Voeikov Main Geophysical Observatory, Russia

Analysis of different aspects (means and variability) of the Arctic surface climate and the Arctic Ocean TH structure shows significant biases and inter-model scatters. Using projection ensembles is essential, but simple across-model averages are not the best use of the model data. On the other hand, the problem of model discrimination is not trivial: new metrics and diverse approaches are needed to select models for ensemble projections. "Objective" or "universal" discrimination of models is hardly possible: the selection is application (i.e., region, variable, task, etc.) dependent. Model discrimination is likely to be a serious challenge for the IPCC AR5. For a number of reasons (poor observations, poor understanding of processes, etc.), the Arctic is a particular challenge in this context. Details can be found in a number of papers (with the author's participation) devoted to evaluation of the CMIP3 models:

References

Overland, J.E., M. Wang, N.A. Bond, J.E. Walsh, V.M. Kattsov, and W.L. Chapman, 2009: Considerations in the Selection of Global Climate Models for Regional Climate Projections: The Arctic as a Case Study. *J. Climate,* (submitted)

Alekseev, G.V., A.I. Danilov, V.M. Kattsov, S.I. Kuzmina, and N.E. Ivanov, 2009: Changes in the climate and sea ice of the Northern hemisphere in the 20th and 21st centuries from data of observation and modeling. *Izvestia of Russian Academy of Sciences: Physics of Atmosphere and Ocean*, **45**(6), 723–735.

Pavlova, T.V., V.M. Kattsov, Ye.D. Nadyozhina, P.V. Sporyshev, V.A. Govorkova, 2007: Terrestrial cryosphere evolution through the 20th and 21st centuries as simulated with the new generation of global climate models. *Earth Cryosphere*, **11**(2), 3-13.

Sorteberg, A., V. Kattsov, J.E. Walsh, T. Pavlova, 2007: The Arctic Surface Energy Budget as Simulated with the IPCC AR4 AOGCMs. *Climate Dynamics,* doi:10.1007/s00382-006-0222-9

Kattsov, V.M., J.E. Walsh, W.L. Chapman, V.A. Govorkova, T.V. Pavlova, and X. Zhang, 2007: Simulation and Projection of Arctic Freshwater Budget Components by the IPCC AR4 Global Climate Models. *J. Hydrometeorology*, **8**, 571-589.

Wang M., Overland J.E., Kattsov V., Walsh J.E., Zhang X., Pavlova T., 2007: Intrinsic versus forced variation in coupled climate model simulations over the Arctic during the 20th Century. *J.Climate*, **20**, 1084-1098.

Kattsov, V.M., G.A. Alekseev, T.V. Pavlova, P.V. Sporyshev, R.V. Bekryaev, V.A. Govorkova, 2007: Modeling the evolution of the World Ocean ice cover in the 20th and 21st centuries. *Izvestia of Russian Academy of Sciences: Physics of Atmosphere and Ocean*, **43**(2), 165–181.

Kattsov, V.M., P.V. Sporyshev, 2006: Timing of global warming in IPCC AR4 AOGCM simulations. *Geophys. Res. Letters*, **33**, L23707, doi:10.10 29.2006GL027476.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Challenges in Combining Projections from Multiple Climate Models

Reto Knutti

Institute for Atmospheric and Climate Science, ETH, Switzerland

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 regional climate change. International climate change assessments like IPCC rely heavily on these models and often provide model ranges as uncertainties and equal-weighted averages as best-guess results, the latter 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. This is based on the result that a multi-model average of present-day climate generally out-performs any individual model. However, there are several challenges in averaging models and interpreting spread from such ensembles of opportunity.

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 when each institution is only developing one or two model versions. The multi model ensemble should probably be interpreted as a set of 'best guess' models from different institutions, all carefully tuned to the same datasets, rather than a set of models representing the uncertainties that are known to exist or trying to push the extremes of plausible model response.

Model skill in simulating present day climate conditions is often weakly related to the magnitude of predicted change (Knutti et al., 2009). It is thus unclear how the skill of these models should be evaluated, i.e. what metric should be used to define whether a model is 'good' or 'bad', and 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. Metrics of skill are also likely to depend on the question and quantity of interest.

In many probabilistic methods, the models are assumed to be independent and distributed around the truth, which implies that the uncertainty of the central tendency of the ensemble decreases as the number of models increases. Because all models are based on similar assumptions and share common limitations, this behavior is unlikely to be meaningful at least for a large number of models. Indeed the averaging of models and the correlation structure suggest that the effective number of independent models is much smaller than the number of models in the ensemble, and that model biases are often correlated (Jun et al., 2008a; Jun et al., 2008b; Knutti et al., 2009).

The fundamental problem in estimating uncertainty for climate projections is that the projections relate to a state of the system never observed and far into the future. Skill cannot be quantified in a frequentist sense like for example in a weather forecast, where thousands of hindcasts can be evaluated against observed weather. Credibility for climate projections therefore needs to be established indirectly by evaluating the models on their representation of present day climate, its variability, anthropogenic trends or in paleoclimate applications (Knutti, 2008a; Knutti, 2008b). A rarely discussed problem is that by doing so, the same data is often used to develop parameterizations in models, to calibrate models, to evaluate them, and in some cases to weight them when combining multiple models.

The bottom line is that despite of a massive increase computational capacity and despite of (or maybe because of) an increase in model complexity, the model spread in future projections is often not decreasing. Even on the largest scale, e.g. for climate sensitivity, the range covered by models has remained virtually unchanged for three decades. Probabilistic projections based on Bayesian methods that determine weights for each model strongly depend on the assumptions made for the likelihood, i.e., the metric chosen to define model performance (Tebaldi and Knutti, 2007). Future model intercomparisons and methods to quantify uncertainties will face additional challenges when combining perturbed physics ensembles (a single model run with multiple parameters sets) and structurally different models, and when trying to incorporate structural error, i.e. the fact that many models tend to have common biases. Whether and how to weight models in multi model projections seems unclear at this stage (Knutti, 2010). Some recent studies have proposed ways to do so while others have shown that the pitfalls may be larger than the potential benefits.

Finally, the sheer amount of data that will result from the next CMIP5 coordinated experiments presents almost insurmountable problems to distribute, analyze, visualize and communicate the results effectively in the very short time that will be available before the IPCC chapters need to be written.

References

Jun M., R. Knutti, and D. W. Nychka, 2008a: Spatial Analysis to Quantify Numerical Model Bias and Dependence: How Many Climate Models Are There? *J. Am. Stat. Assoc.*, **103**, 934-947.

Jun M., R. Knutti, and D. W. Nychka, 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? *Philos. T. R. Soc. A.*, **366**, 4647-4664.

Knutti, R., 2008b: Why are climate models reproducing the observed global surface warming so well? *Geophys. Res. Lett.*, **35**, L18704, doi:10.1029/ 2008GL034932.

Knutti, R., 2010: The end of model democracy? *Clim. Change*, submitted.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2009: Challenges in combining projections from multiple models. *J. Clim.*, in press.

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.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Probabilistic Regional Climate Change Projections Using Bayesian Model Averaging

Won-Tae Kwon, Hee-Jeong Baek, and E-Hyung Park

National Institute of Meteorological Research, KMA, Korea

Recently Bayesian approaches have been applied to model evaluation and multi-model ensemble averaging for weather and climate predictions. We employed similar method used by Min et al. (2007) for regional-scale climate change projections using IPCC AR4 data set. The objective of this study is to calculate a probabilistic projection over East Asia using Bayesian Averaging Model (BAM) and to estimate uncertainties of regional climate projections.

The BMA technique is applied to the twenty-first century temperature changes simulated by the 18 AOGCMs of IPCC AR4 to produce probabilistic predictions of regional temperature over East Asia. Monthly surface temperature data over land are derived from the Climate Research Unit for the period of 1950-1999, interpolated into 2.5 x 2.5° grid. Model training is based on long-term temporal components (Legendre degrees from LP1 to LP3) to eliminate the noise on shorter time-scales. The BMA based upon the Bayes Factor (BF) approach takes the likelihood ratio which is an exponential function of a generalized Mahalanobis distance between observation and model simulation for the period of 1950-1999. Hence, it filters out low-skilled models more effectively than a mean-square error based approach. This approach provides a way of observationally constrained prediction of PDFs by using weighting factors which are obtained through evaluating models for the last 50 years of the twentieth century.

The results show that both PDFs from weighted and unweighted methods indicate broadening of modal structure during the second half of 21st century and that mean, 5th and 95th-percentile values of BF method are larger than the values from the unweighted method, as presented in the figure. This suggests that observationally constrained probabilistic climate change predictions using BMA are feasible and can provide more information than the unweighted ensemble. Comprehensive measure of model skills based either on space–time vectors of temperature or on multiple variables (e.g., temperature and sea level pressure) might be useful to produce more robust weighting factors and hence more reliable probabilistic predictions of regional climate changes.

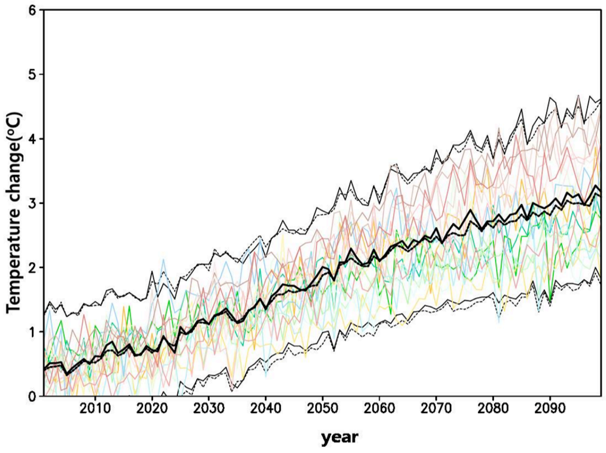


Figure 1. Multi-model average (thick) and its 5-95% percentile (thin lines) of mean annual 2 m air temperature predictions over East Asia for 2001-2099 with BF (solid) and AEM (dashed) under SRES A1B scenarios. Colored solid lines represent predictions of 18 participating models.

Effect of Chemistry-Aerosol-Climate Coupling on Predictions of Future Climate and Future Levels of Tropospheric Ozone and Aerosols

Hong Liao[1], Ying Zhang[1], Wei-Ting Chen[2], Frank Raes[4], and John H. Seinfeld[2,3]

[1]*LAPC, Institute of Atmospheric Physics, Chinese Academy of Sciences, China*
[2]*Department of Environmental Science and Engineering, California Institute of Technology, USA*
[3]*Department of Chemical Engineering, California Institute of Technology, USA*
[4]*European Commission, Joint Research Center, Italy*

We explore the extent to which chemistry-aerosol-climate coupling influences predictions of future ozone and aerosols as well as future climate using the Goddard Institute for Space Studies (GISS) general circulation model II' with on-line simulation of tropospheric ozone-NO_x-hydrocarbon chemistry and sulfate, nitrate, ammonium, black carbon, primary organic carbon, and secondary organic carbon aerosols. Based on IPCC scenario A2, year 2100 ozone, aerosols, and climate simulated with full chemistry-aerosol-climate coupling are compared with those simulated from a stepwise approach. In the stepwise method year 2100 ozone and aerosols are first simulated using present-day climate and year 2100 emissions (denoted as simulation CHEM2100sw) and year 2100 climate is then predicted using offline monthly fields of O_3 and aerosols from CHEM2100sw (denoted as simulation CLIM2100sw). The fully coupled chemistry-aerosol-climate simulation predicts a 15% lower global burden of O_3 for year 2100 than the simulation CHEM2100sw which does not account for future changes in climate. Relative to CHEM2100sw, year 2100 column burdens of all aerosols in the fully coupled simulation exhibit reductions of 10-20 mg m^{-2} in DJF and up to 10 mg m^{-2} in JJA in mid to high latitudes in the Northern Hemisphere, reductions of up to 20 mg m^{-2} over the eastern United States, northeastern China, and Europe in DJF, and increases of 30-50 mg m^{-2} over populated and biomass burning areas in JJA. As a result, relative to year 2100 climate simulated from CLIM2100sw, full chemistry-aerosol-climate coupling leads to a stronger net global warming by greenhouse gases, tropospheric ozone and aerosols in year 2100, with a global and annual mean surface air temperature higher by 0.42 K. For simulation of year 2100 aerosols, we conclude that it is important to consider the positive feedback between future aerosol direct radiative forcing and future aerosol concentrations; increased aerosol concentrations lead to reductions in convection and precipitation (or wet deposition of aerosols), further increasing lower tropospheric aerosol concentrations.

Downscaling IPCC-AR4 Climate Change Scenarios for Adaptation Strategies in Mexico

Victor Magaña and David Zermeño

Centro de Ciencias de la Atmósfera, Universidad Nacional Autónoma de Mexico, Mexico

The need to develop regional scale climate change projections has led to explore dynamical and statistical techniques to have an ensemble of scenarios. Unfortunately, the availability of dynamically downscaled scenarios for Mexico is limited to two or three models. Such scenarios have been used to examine processes that result in a particular signal in temperature or precipitation change. The use of a statistical model to downscale the IPCC-AR4 scenarios is an adequate option to construct an ensemble of climate change projections of relatively high spatial resolution (50 km x 50 km). We have used the Climate Predictability Tool (CPT) to downscale most IPCC-AR4 models (around 25) and their various realizations. Results show that the CPT is capable of reducing systematic errors in the models to produce adequate climate change projections, under the assumption that the relationships between large scale and mesoscale patterns remain valid under a warmer climate.

The downscaled projections have been used in conjunction with vulnerability projections to estimate risk under climate change for various sectors such as water management, agriculture, and others. One of the main challenges is the projection of extreme meteorological events. By means of a weather generator (LARS) modulated by the downscaled climate change projections, we have mapped regions where probabilities of more intense extreme events (e.g., heat waves) are large. These scenarios of extreme events have been combined with projections of land use change to estimate the risk for health. Results suggest that a redefinition of urban growth and land use management is necessary in some parts of Mexico if risk for the populations under heat waves is to be reduced.

The availability of climate change scenarios of higher spatial resolution allows decision makers to estimate probabilities of change that surpass critical threshold values risk, such as those that make agriculture unviable. Even more, scenarios of risk for certain regions, sectors and groups of the population are leading to propose better strategies of land use management to compensate the effects of regional changes of climate change plus heat island effects. New experiments, with regional climate models, of climate change that include projections of land use change for the coming decades are being prepared for adaptation purposes.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Changes in Rainfall Extremes in South America as Derived from the ETZ CPTEC Regional Climate by the Downscaling of the HadCM3 Global Model

Jose A. Marengo

CCST/INPE, Brazil

Using the Eta-CPTEC model 4° km, this study analyzes the distribution of extremes of precipitation in South America for the period 2010-2100, under the SRES A1B scenario. The lateral boundary conditions used to drive the Eta-CPTEC regional model are supplied by versions of the Met Office Hadley Centre coupled climate model HadCM3.

The HadCM3 ensemble, from which the lateral boundary conditions are taken, was designed to quantify uncertainty in projections of climate change derived from uncertainty in parameter settings within the model, as per the second method described above. Through expert elicitation, key uncertain parameters were identified, primarily in the atmosphere but also in the land surface, and their plausible ranges defined. These parameters were modified within their plausible ranges to form a large (300+ member) ensemble, run with a computationally-efficient slab ocean. From this ensemble, a subset of 16 model variants, each with a different combination of parameter settings, was selected according to performance in the realistic simulation of the current climate, while still sampling parameter space widely. Together with the standard HadCM3 model, the 16 model variants were run in fully coupled transient mode, forced with SRES A1B emissions scenario-generated CO_2 concentrations (IPCC, 2000) to the end of the 21st century. As a first step, an analysis of the present day climate simulation for mean climate and extremes (mean and variability) is performed using previous studies on present time model runs with the same Eta/CPTEC-HadCM3. This allows for a more comprehensive identification and possible interpretation of systematic model biases. The results to be presented correspond to 1 model variant.

Putting it All Together: Are We Going in the Right Direction for Providing Users with Better Information About Future Climate to Support Decision-Making?

Linda O. Mearns

National Center for Atmospheric Research, USA

'You can't always get what you want ….. but if you try sometimes …. you just might find … you can get what you need.' – The Rolling Stones

The Intergovernmental Panel on Climate Change has made numerous statements in its reports over the years about the goal of providing policy relevant (but not policy prescriptive) information regarding climate change. However, it is only recently that this statement has received serious, concrete consideration, grasping the urgency for policy decisions. Noteworthy is the fact that in the Third Assessment Report, there was a chapter in Working Group I on Climate Scenario Development (for use in impacts/adaptation studies, Mearns et al., 2001). A statement commonly made after the Fourth Assessment Report was released was that the question at the center of the climate change issue had shifted from 'is climate change occurring?' to 'what are we going to do about it?' Certainly the steadily increasing strength of the statements about the detection and attribution of climate change that appeared in the four assessment reports contributed significantly to this shift. Yet it was also recognized that the role of physical climate science should not be marginalized but perhaps reshaped to better address the needs of decision makers. By now many climate scientists have had some interactions with various so-called stakeholders. One of the common perceptions of what is wanted by decision makers, particularly those concerned with adaptation planning, is 'accurate predictions' of future climate on very high spatial resolutions (e.g., for a small river catchment in Colorado for 2050). It is assumed this is desired because a particular approach to decision making is assumed, the so-called predict-then-act mode, wherein there is some consensus on 'true' probabilities about future climate change, and these probabilities are used at the front end of a process cascade that leads to management decisions based on those probabilities. This is assuming that climate change is a problem that falls neatly into a classical risk management framework wherein classical decision analysis techniques can be used to optimize decision making (Lempert et al., 2004). Another approach more readily embraces the reality that climate change is an issue imbued with deep (or extreme) uncertainty, that we may never be able to come up with truly robust probabilistic information about climate change on fine spatial scales twenty years in advance. This strategy, often termed robust decision making, focuses on starting with the types of decisions resource managers face and developing strategies that work well across a wide range of future climate conditions. As a part of this approach one also would want to identify adaptive strategies that can be modified to achieve better performance as one learns more about future climate change (Morgan et al., 2009). These strategies work much better under conditions of deep uncertainty than traditional decision analysis approaches. But what does all this mean for our work at this meeting? What is the effect of moving to robust decision making and adaptive decision making on the motivations and goals of climate scientists to provide useful information about future climate change? For one thing, it is important that climate scientists grasp that irrevocable decisions regarding adaptation, for example, are not going to be made tomorrow based on, say, the regional projections from the CMIP3 dataset, regardless of how that information is summarized (e.g., as scenarios or probabilities). We are all still in a learning mode. So one of the main purposes of producing information about future climate is so that we can all learn from it, not necessarily immediately make resource management decisions. Hulme and Dessai (2008) consider three different criteria for evaluating the success of national climate scenario efforts: predictive success, decision success, and learning success. They underscore that establishing predictive success is virtually impossible for long term climate change, whereas the other criteria are more verifiable and likely more important. Another important concept that emerges over is that of reducing uncertainty about future (regional) climate change. If one sees the main purpose of developing information about future climate as predicting future climate, then the goal of reducing uncertainty follows naturally. However, over the next decade, it may continue to be extremely difficult to reduce this uncertainty, or, perhaps more accurately, we don't yet know if we can. In a sense, the regional statements made about climate change (discussed by Hewitson) in the chapter on regional projections in the AR4 report did indicate some reduction in uncertainty, since such detailed statements had never been made before (e.g., decreased precipitation in the southwest US), but we also know that these statements are not immutable. While the overarching goal of science may be to reduce uncertainty in the sense that it is to increase our knowledge (reduce our ignorance) of how the world works, this is not the same thing as reducing uncertainty about future regional climate. We know that this specific metric (e.g., range of change in precipitation) could as easily expand as contract in the next decade. For these reasons, we should be more interested in reducing vulnerability of resource systems to climate change than in reducing uncertainty about the future climate itself. We are still learning about the relationship between these concepts. Finally I mention the danger of false certainty – do we know which is more dangerous for our future – deep uncertainty about climate change or false certainty about it? As Mark Twain stated, ''It ain't what you don't know that gets you in trouble. It's what you know for sure that just ain't so.''

References

Hume, M., and S. Dessai, 2008: Predicting, deciding, learning: can one evaluate the success of national climate scenarios? *Env. Res. Lett.,* **3**(045013), 7 pp.

Lempert, R., Nakicenvoic, N. Sarewitz, D., and M. Schlesinger, 2004: Characterizing climate change uncertainties for decision-makers. *Climatic Change*, **65**,1-9.

Mearns, L. O., M. Hulme et al., 2001: Climate change scenario development. In: *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change.* [Houghton, J.T., et al. (eds)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 739-768.

Morgan, G., H. Dowlatabadi, M. Henrion, et al., 2009: Best practice approaches for characterizing, communicating, and incorporating scientific uncertainty in climate decision making. US CCSP SAP 5.2 NOAA:Washington, D.C., 96 pp.

The Interdecadal Pacific Oscillation and Pacific Decadal Variability in CMIP3 Multi-Model Ensemble

Gerald A. Meehl, Aixue Hu and Julie Arblaster

National Center for Atmospheric Research, USA

The Interdecadal Pacific Oscillation (IPO) is the Pacific basin-wide manifestation of the Pacific Decadal Oscillation (PDO), which is defined for the north Pacific. Both are similar in pattern and involve decadal to multi-decadal fluctuations of sea surface temperatures (SSTs). In the positive phase, tropical Pacific SSTs have positive anomalies along with areas along the west coasts of North and South America, while there are negative SST anomalies in the northwest and southwest Pacific. The opposite pattern occurs in the negative phase. The IPO has been identified as being associated with decadal variability involving precipitation over southwestern North America and South Asia. The mid-1970s climate shift has been postulated to have been caused in part by a change in phase of the IPO, and a mechanism to produce the IPO has been identified in model simulations. Here, the CMIP3 multi-model ensemble is analyzed to identify characteristics involved with the IPO in terms of pattern, amplitude, and period across the models. Presumably, a credible simulation of the IPO in model simulations would be a necessary condition for skillful decadal predictions in the Pacific region.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Multi Model Ensembles, Metrics & Probabilities

Wendy S. Parker

Department of Philosophy, Ohio University, USA

A number of methodological questions must be addressed when carrying out multi model ensemble studies: On what basis should a model be included in or excluded from the study? Should projections from some models be given more weight than others? If so, how should they be weighted? Should ensemble results be transformed into probability distributions?

A minimum requirement for including a model in an ensemble study is the following: it is plausible that the model can provide at least some of the information that we seek in the study. Typically, this will be information about particular quantities under one or more emission scenarios. We may judge it plausible that a model can provide desired information about quantity of interest *A* (e.g., global mean surface temperature) but not about quantity of interest *B* (e.g., changes in regional precipitation). If so, then while the model may be included in an ensemble study that seeks information about both *A* and *B*, the model's results for *B* should be excluded from the analysis.

Judgments of plausibility should consider a number of factors, including: the construction of the model (e.g., whether/how it represents processes that significantly influence the quantities of interest), its spatiotemporal resolution, and its performance in simulating those aspects of past climate that seem particularly relevant (given the predictive goals at hand) in light of process-based thinking or for other reasons. These factors can be considered informally or more formally with the help of metrics.

A **metric of model performance** defines a measure of the difference between model output and one or more observational datasets (see Gleckler et al., 2008 for some examples). A **metric of model quality** defines a measure of the quality or "goodness" of a model, given the purposes for which the model is to be used, and may take into account all of the factors mentioned above: model construction, spatiotemporal resolution, and scores on relevant metrics of performance. A metric of model quality that is appropriate when evaluating whether a model is plausibly adequate for one purpose may not be particularly appropriate for evaluating whether a model is plausibly adequate for another.

Metrics may also be used to set more stringent standards for the inclusion of a model in an ensemble study. For instance, it might be required that a model not only meets some minimum plausibility standard but also scores within some specified distance of the best score achieved on a chosen metric of performance or quality. Populating an ensemble with only these "best" models, and not including other available models that are plausibly adequate for the purpose of interest, can have implications for the interpretation of ensemble results. In particular, it may give us reason to believe that the set of projections produced will fail to include some outcomes that remain plausible, given the limitations of current understanding. This in turn means that, if the results produced in the study are to be transformed into probability distributions (PDFs), not all of the probability mass should be distributed over the range of results produced.

Of course, even if all available models that meet a minimum plausibility requirement are included in a multi model study, it cannot be assumed that the range of results produced is likely to include the outcome that would be observed under the chosen emission scenario. Such a conclusion requires that we be able to assert that it is likely that at least one of the models included in the ensemble is adequate for projecting the quantities of interest with some specified level of accuracy, even if we cannot say which model (see Parker, under revision). For many predictive tasks of interest (e.g., predicting regional changes in climate later in the century), it seems we are currently unable to defend such an assertion about the models used in today's multi model studies. Again, this implies that, if multi model results are to be transformed into PDFs, then not all of the probability mass – and perhaps not even most of the probability mass – should be distributed over the range of results produced. This in turn implies that PDFs should not be produced by simply weighting results from different models according to the models' *relative* scores on some chosen metric of performance or quality.

Is important to keep in mind that the PDFs of interest are representations of epistemic uncertainty about what would occur under various emission scenarios; they are meant to convey our confidence (or degree of belief) that different outcomes would occur, *taking into account all of the available evidence.* This evidence includes much more than just the set of results produced in our latest multi model study. It includes (at least) results from other ensemble studies and background knowledge about the climate system. To take this evidence into account, we must engage in a complicated process of analysis and synthesis, involving judgments about the quality and degree of relevance of different pieces of evidence. This too speaks against the idea of producing PDFs by weighting results from a single ensemble study using a metric of performance or quality, since producing PDFs in this way will take into account only some of the available evidence (Parker in press). Metrics might still be used to identify some results as more plausible than others, but this relative plausibility cannot be transformed in a simple way into a probability.

Should PDFs be produced at all? If so, how? I would argue that a PDF should be offered only if both the approximate width and approximate shape of the PDF can be justified and only if the PDF can be considered robust in each of two senses, viz., it is not highly sensitive to contentious assumptions, and it is not expected to change substantially in the very near term, e.g., as models undergo incremental development. It would not be good, either for policy or for the credibility of climate science, to offer policymakers a PDF now with the knowledge that a very different PDF – implying very different probabilities for outcomes of interest – will probably be offered a couple of years later. If these criteria cannot be met, then uncertainty about a predictive outcome should be communicated in some other way, e.g. by offer a range in which one expects the outcome to fall or by indicating the expected sign of a change without assigning a magnitude (see Kandlikar et al., 2005).

If it is appropriate to offer PDFs for some predictive variables, then those PDFs should be arrived at after surveying all of the available evidence and after much critical discussion, taking care to avoid well-known pitfalls, such as double counting of evidence, as far as possible. Results from our latest multi model study will be an important piece of evidence, but only one piece.

References

Gleckler, P.J., K.E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *Journal of Geophysical Research,* **113**, D06104.

Kandlikar, M., J. Risbey, and S. Dessai, 2005: Representing and communicating deep uncertainty in climate change assessments. *Comptes Rendus Geoscience,* **337**, 443–455.

Parker, W.S., What does it mean when climate models agree? Examining the significance of robust predictions. *Philosophy of Science,* under revision.

Parker, W.S., Whose Probabilities? Predicting Climate Change with Ensembles of Models. *Proceedings of the Philosophy of Science Association*, in press.

The Influence of Model Skill on Regional Projections of Extreme Temperatures over Australia

Sarah E. Perkins

CSIRO Marine and Atmospheric Research, Australia

Model evaluation by means of comparing 20th Century simulations to observations is readily undertaken to gain some level of confidence of model reliability. Many studies including Watterson (1996), Taylor (2001), Delworth et al. (2006), Knutti et al. (2006), and Shukla et al. (2006) have devised evaluation metrics however all have based their analysis on monthly, seasonal or annual data. Given that climate on time scales of days has a direct impact on human health (Trigo et al., 2005, Woodruff et al., 2006) and human activities (e.g., agriculture; Luo et al., 2005), an assessment of the capacity of models to simulate climate on time scales of days is clearly valuable.

Perkins et al. (2007) introduced a metric which evaluated the CMIP3 GCMs on their ability to reproduce the observed probability density function (PDF) for daily minimum temperature (T_{MIN}), maximum temperature (T_{MAX}) and precipitation. The metric (S_{score}) measures the amount of overlap between the observed and modelled PDFs by summing the minimum values across the common bins. Evaluation was performed for twelve regions across Australia, each representing one or more different climatic types (Perkins et al., 2007). Perkins et al. (2007) demonstrated that the ensemble PDF closely resembled the observed PDF as models with lower S_{score} values were omitted. Overall, was concluded that while limited to one continent, some of the CMIP3 models show considerable skill at sub-continental scales, when assessed using daily data.

In order to investigate whether CMIP3 models with lower skill biased future projections of temperature extremes, Perkins et al. (2009) employed three measures of model evaluation calculated for daily T_{MIN} and T_{MAX}. The measures of skill included the difference between the observed and modelled mean, the S_{score} proposed by Perkins et al. (2007) and a new metric ($Tail_{skill}$) which focuses on the weighted difference between the top (bottom) 5% for T_{MAX} (T_{MIN}). The generalized extreme value (GEV) distribution was implemented (Kharin et al, 2007) to estimate and assess changes in the 20-year return value. Evaluation was performed for 1981-2000 and projections were considered for the SRES A2 scenario for 2081-2100. Models were chosen due to data availability for the 20c3m and A2 scenarios at the time data was obtained (6 models for T_{MAX} and 9 models for T_{MIN}). Models with multiple realizations were concatenated to form a single sample to avoid selective sampling of any one realization. Once evaluation was performed, models were divided into "weaker" and "stronger" ensembles by ranking their score for each evaluation metric. Projections were based on either an average (continental) or a range (regional) of the ensembles, and were compared to the all-model ensemble. Regional analysis was performed for regions 2, 3 and 10 outlined in Perkins et al. (2007). These regions include temperate, subtropical and tropical climates respectively.

Figure 1 shows the range in the projected T_{MAX} 20-year return values and the 90% bootstrapped confidence intervals calculated from 1000 samples for the all-model ensemble and the "weaker" and "stronger" ensembles for each measure of skill. The all-model ensemble for each region (first bar) shows a large range of projected temperatures. In each region, and irrespective of skill-score used, the projected T_{MAX} is always lower in the stronger models than the weaker models, and the 90% confidence levels for the two ensembles do not overlap in Regions 2 and 3. This suggests that the projected 20-year interval temperatures from the weak models are statistically significantly higher than those projected by the strong models, irrespective of whether ''strong'' is defined using the mean, PDF or tail skill measure. Results for T_{MIN} (not shown) inferred that the samples are not consistently significantly different at a 90% confidence level, but there is a systematic difference in that the weaker models always simulate larger amounts of increase in T_{MIN} than the stronger models. Continental results are presented in Perkins et al. (2009).

Figure 1 highlights a clear bias in that weaker models systematically and statistically significantly simulate a larger increase in T_{MAX} than the stronger models. The use of an all-model ensemble therefore tends to over-predict the amount of increase in both T_{MAX} and T_{MIN} in the 20-year return levels over Australia, at least for the climates presented here. It is likely that such results can be extrapolated to other regions across the continent. Results highlight the need to begin to exclude a given model from regional projections where it shows weaker skill, and that projections are less affected by the chosen measure of skill, compared to not evaluating and simply averaging across all models.

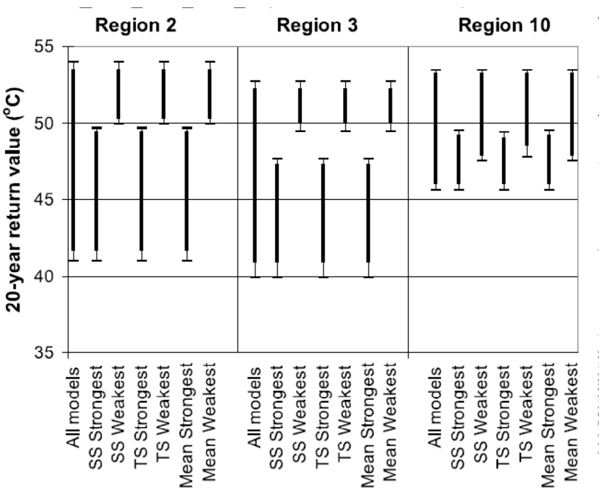


Figure 1. The 20-year return values of T_{MAX} and 90% confidence for the A2 emission scenario (2081-2100) for the all model ensemble, strongest and weakest three models in the S_{score} (SS), the strongest and weakest three models in the $Tail_{skill}$ (TS) ensemble and finally the strongest and weakest three models in the differenced mean skill based ensemble for Regions 2, 3 and 10 outlined in Perkins et al. (2007).

References

Delworth, T., et al., 2006: GFDL's CM2 global coupled climate models – Part 1: Formulation and simulation characteristics. *J. Climate*, **19**, 643-674.

Kharin, V.V., F.W. Zwiers, X. Zhang, and G.C. Hegerl, 2007: Changes in temperature and precipitation extremes in the IPCC ensemble of global couple model simulations. *J. Climate,* **20**, 1419-1444.

Knutti, R., G.A. Meehl, M.R. Allen, and D.A. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Climate,* **19**, 4224–4233.

Luo, Q., R.N. Jones, M. Williams, B. Bryan, and W. Bellotti, 2005: Probabilistic distributions of regional climate change and their application in risk analysis of wheat production. *Climate Research,* **29**, 41-52.

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. Climate,* **20**, 4356- 4376.

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**, L06710, doi: 10.1029/2009GL037293.

Taylor, K.E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.* **106**(D7), 7183–7192.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter, and D. Paolino, 2006: Climate model fidelity and projections of climate change. *Geophys. Res. Lett.,* **33**, L07702, doi:10.1029/2005GL025579.

Trigo, R., R. Garia-Herrera, J. Diaz, I. Trigo, and M. Valente, 2005: How exceptional was the early August 2003 heatwave in France? *Geophys. Res. Lett.,* **32**, 1071-1074.

Woodruff, R.E., T. McMichael, and C. Butler, 2006: Action on climate change: the health risks of procrastinating. *Aust. NZ. J. Publ. Heal.,* **30**, 567-571.

Watterson, I.G., 1996: Non-dimensional measures of climate model performance. *Geophys. Res. Lett.,* **31**, L24123, DOI:10.1029/2004GL021276.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Selecting Global Climate Models for Regional Climate Change Studies

David W. Pierce[1], Tim P. Barnett[1], Benjamin D. Santer[2], Peter J. Gleckler[2]

[1] Division of Climate, Atmospheric Sciences, and Physical Oceanography, Scripps Institution of Oceanography, USA
[2] Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, USA

Introduction

Global climate models are increasingly being downscaled to address regional climate issues. But which of the 24 models in the IPCC AR4 report should be used for a particular regional study? And how should different models be combined to give the most reliable results? We address this question using as a demonstration case a recent detection and attribution (D&A) study of changes in the hydrological cycle of the western U.S. (Barnett et al., 2008 [B08 hereafter]).

A common approach is simply to average over all models with available data. This is justified by global scale results, but this procedure weights models that do a poor job simulating the region of interest equally with those that do a good job. Is there a better strategy?

An increasingly popular approach is to generate metrics of model skill, then pre-qualify models based on their ability to simulate climate in the region or variable of interest (e.g., Coquard et al., 2004; Gleckler et al., 2008; Brekke et al., 2008). Do models selected this way provide a better match to observed changes?

Models and Data

We use global model JFM minimum near-surface temperature (“tasmin”) over the western U.S. as a surrogate for the multi-variate analysis of B08. We also reuse the internal climate variability (noise) estimates from B08, obtained from 1600 years of simulation with two different models (cf. Santer et al., 2009).

Data from 21 global models forced by 20th century changes in anthropogenic and natural factors were obtained from the LLNL CMIP3 archive. The period analyzed is 1960-1999. To facilitate comparison, all model fields and the observations (below) were interpolated to a common 1°×1° grid over the western U.S.

We compare model temperatures and precipitation to a daily observed data set gridded at 1/8° longitude by latitude resolution across the western U.S. (Hamlet and Lettenmaier, 2005).

Metrics

We use 42 metrics to characterize each model. We use 4 seasonal (DJF, MAM, JJA, and SON) averages of 2 variables (surface air temperature [tas] and precipitation [pr]) in 4 aspects: the seasonal mean and the temporal standard deviation of the seasonal data averaged into 1, 5, and 10-yr blocks. This gives 32 metrics.

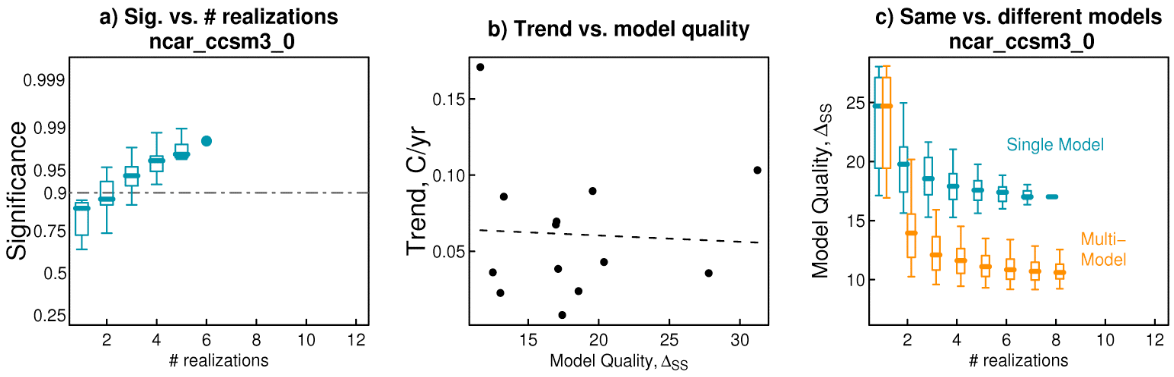
For ENSO and the PDO, we construct one metric describing the climate mode’s sea surface temperature pattern in the region where it is defined, and additional metrics describing the teleconnected effects of the climate mode in western tas and pr. This yields another 6 metrics. Finally we include the phase and amplitude of the tas and pr annual cycle, for a total of 42 metrics.

Modeled temperature trends

In evaluating the model temperature trends, we use most of the formal, fingerprint-based D&A methodology employed in B08. However, no downscaling is done due to the resources required for 21 models.

The models produce temperature trends in the western U.S. ranging from -0.05 to +0.21 °C/decade. Observations show +0.10 °C/decade. All 5 models with a negative trend have only 1 realization, while none of the 13 models with more than 1 realization has a negative ensemble-averaged trend. Natural variability means a single realization does not provide a reliable estimate of the warming signal.

The relationship between *N* (the number of realizations from a single model included in the ensemble average) and the significance of the ensemble averaged trend is



illustrated for a typical model in Figure 1a. Significance is assessed by comparing to trends found in the control runs, which lack anthropogenic forcing. All but one model show an upward trend in significance as the number of realizations increases, due to the averaging away of natural internal variability.

We calculate what the D&A results of B08 might have been if the 14 realizations used there had been chosen randomly from all the models available (63 realizations total). Using 10,000 random selections, we found 96% of the random trials resulted in a trend significant at the 90% level, and 90% of the trials gave a trend significant at the 95% level. Therefore, the finding that the JFM tasmin trend over the western U.S. is both detectable and attributable to combined anthropogenic and natural effects is robust to the range of trends found in the CMIP3 models.

The Role of Model Quality

We order the models in terms of quality by considering each model's skill scores to be a point in n_{metrics} (=42) dimensional space. In the results shown here, the ordering is given by Δ_{SS}, the Euclidian distance from the model's point to perfect skill at point (1,1,1,…,1). Lower values of Δ_{SS} indicate better simulations.

Fig. 1b shows how the magnitude of the JFM tasmin trend relates to Δ_{SS}. This has been calculated using only the 13 models that have more than 1 realization with tasmin, to reduce the effects of natural variability. There is no significant relationship between this measure of model quality and the regional tasmin trend.

These results are with individual models, but perhaps averaging across models is required for any relationships to be discerned. Accordingly, we separated the models into groups of the top 10 and bottom 11 based on Δ_{SS}, and computed the mean JFM tasmin trend for each group. The difference in trend between the groups was compared to Monte Carlo estimates of the difference using models partitioned randomly. We found no statistically significant difference in the distribution of trends obtained when partitioning by model quality compared to random partitioning.

The multi-model ensemble

The multi-model ensemble average, *MM*, is the best model in the overall skill score. Given the important role ensemble size plays, is *MM* better simply because it includes information from far more realizations than any individual model? Fig. 1c illustrates how Δ_{SS} changes as progressively more realizations from the *same* model (blue) or randomly selected *different* models (red) are added to the average. For most models, skill increases (Δ_{SS} decreases) more quickly when different models are added to the mix than when more realizations of the same model are included. This is true even for the same number of ensemble members.

MM's superiority can be understood by decomposing the model errors into the mean error, an error in the ratio of the model's variance to observed, and the pattern correlation between the model and observed. Mean errors are distributed around zero, and the variance ratio tends is distributed about 1. Averaging across models reduces the error in both these aspects, both in the mean climate and variability. For the pattern correlation, averaging across models tends to give better correlation with observations than any individual model, consistent with the argument that effectively random spatial errors are being reduced by averaging.

Summary

Our results show that the best way we currently have to use information from multiple global model runs in a regional D&A study is simply to form the multi-model ensemble average. Neither selecting the models based on the quality of their climate simulations in the region of interest nor forming an optimized ensemble average based on maximizing skill resulted in a superior result over the historical period.

References

Barnett T.P., D.W. Pierce, H.G. Hidalgo, C. Bonfils, B.D. Santer, T. Das, G. Bala, A.W. Wood, T. Nozawa, A.A. Mirin, D.R. Cayan, and M.D. Dettinger, 2008: Human-induced changes in the hydrology of the western United States. *Science*, **319**, 1080-1083.

Brekke L.D., M.D. Dettinger, E.P. Maurer, and M. Anderson, 2008: Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments. *Clim. Change.*, **89**, 371-394, doi:10.1007/s105 84-007-9388-3.

Coquard J., P.B. Duffy, K.E. Taylor, and J.P. Iorio, 2004: Present and future surface climate in the western USA as simulated by 15 global climate models. *Clim. Dyn.*, **23**, 455-472, doi:10.1007/s00382-00400437-6.

Hamlet A.F., and D.P. Lettenmaier, 2005: Production of temporally consistent gridded precipitation and temperature fields for the continental United States. *J. Hydromet.*, **6**, 330-336.

Santer B.D., K.E. Taylor, P.J. Gleckler, C. Bomfils, 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. Nat. Acad. Sci. USA*, **106**, p. 14778.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Using Paleo-Climate Data to Enhance Future Projections

Gavin Schmidt

NASA Goddard Institute for Space Studies, USA

CMIP5 will contain (for the first time) coordinated simulations for 3 periods of the past with substantial natural climate variability (the mid-Holocene, the Last Glacial Maximum and the last millennium). Model/data comparisons for these past climate changes will be a new tool in evaluating the projections of those same models in future scenarios. I outline a framework in which data synthesis combined with suitable modelling targets should be able to reduce uncertainty in both. The Intergovernmental Panel on Climate Change Fourth Assessment Report (IPCC AR4) highlighted a number of areas of uncertainty in future projections, due both to poor understanding of specific processes and to the wide range of sensitivities produced by current models. For many of these areas, there are relevant paleoclimate data that can be used to evaluate or weight model simulations of the future scenarions - provided that is some conformability between the models used for the past and future climate simulations. Specific targets include: the long-term behaviour of El Nino events and the potential response to volcanic and solar forcing; the variability of sub-tropical rainfall and the extent of the Hadley Circulation and their response to orbital and high-latitude forcing; ice sheet responses on sub-millennial timescales; multi-decadal changes in the North Atlantic ocean circulation and, certainly, overall climate sensitivity. In each case, I highlight data synthesis steps and modelling approaches necessary for reducing the uncertainty.

Probabilistic Projections of Climate Change at Global and Regional Scales

David M.H. Sexton, Ben Booth, Mat Collins, Glen Harris, and James Murphy

Met Office Hadley Centre, United Kingdom

In June 2009, the Met Office published the latest set of UK Climate Projections (http://ukcp09.defra.gov.uk/). The projections, which were probabilistic in nature, were provided at 25km resolution for three different SRES emission scenarios. They were produced in three stages. The first stage used a Bayesian framework (Goldstein and Rougier, 2004; Rougier, 2007) to produce probabilistic projections for the equilibrium response to doubled CO_2 levels. This stage allowed us to incorporate different sources of uncertainty: observational uncertainty, parametric uncertainty from an ensemble of perturbed physics runs (Murphy et al., 2004), and structural uncertainty from the CMIP3 multimodel ensemble. The second stage used data from coupled ocean-atmosphere runs to make the projections time-dependent for the 21st century. Several perturbed physics ensembles explored uncertainty in four different components of the Earth System (land/atmosphere, ocean, sulphur cycle, and terrestrial carbon cycle) and so forcing uncertainty was also included at this stage. In the final stage, data from regional climate model runs were used to downscale the projections to 25km resolution.

In this talk, we will concentrate on the first stage of production as this is most relevant to the problem of assessing and combining model projections.

The following issues will be discussed:

1. Motivation for using perturbed physics ensembles instead of just taking information from the multimodel archive.

2. Implementation of the Bayesian approach outlined in Rougier (2007) including multivariate weighting and the importance of taking into account the amount of structural error in the climate model, in particular its effect on one's ability to discern a relatively good climate model from a relatively poor climate model.

3. Estimation of the structural uncertainty explored by the multimodel ensemble that is additional to the parametric uncertainty explored by the perturbed physics ensemble.

4. Testing the sensitivity of probabilistic pro jections to key assumptions in the method and comparison with alternative techniques.

Finally we discuss lessons learnt from the Bayesian framework and what implications these lessons have for assessing and combining multimodel projections.

References

Goldstein, M., and J. Rougier, 2004: Probabilistic formulations for transferring inferences from mathematical models to physical systems. SIAM *J. Sci. Comput.*, **26**, 467-487.

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*, **430**, 768-772.

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

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Quantifying Uncertainty in Future Warming Using Observational Constraints Derived from Past Temperature Changes

Peter Stott

Met Office Hadley Centre, United Kingdom

Increased understanding of the past provides greater confidence in predictions of likely changes in future. Optimal detection analyses (Allen and Tett, 1999) have been used to quantify the contributions of greenhouse gases, other anthropogenic forcings and natural factors to past temperature changes on global and continental scales (Tett et al., 1999; Stott, 2003). It has been shown that there is a close relationship between past and future greenhouse warming (Allen et al, 2000; Frame et al, 2006) and that uncertainties in future warming can be derived based on estimates of past warming attributable to anthropogenic and natural factors (Allen et al., 2000; Stott et al., 2006a; Frame et al., 2006). These observationally constrained analyses indicate that it is very likely that aerosol cooling is suppressing a major portion of current greenhouse warming (Stott et al., 2008) and as a result additional warming is implied if aerosol pollution is removed from the atmosphere in future.

The technique described above, which has been denoted "ASK", obtains observationally constrained uncertainties in future warming conditional on particular emissions scenarios by deriving factors by which climate models can be scaled to be consistent with observed responses to greenhouse gas forcing and other forcings and applying these scaling factors and their uncertainties to climate model simulations of future changes. While global temperature changes appear to be relatively well constrained using this approach there is greater uncertainty in regional patterns of temperature change. Probabilistic predictions of continental scale temperatures were produced using two versions of the attribution scaling technique by Stott et al. (2006b). The first version projected future continental changes according to past changes in the same region (thus obtaining relatively conservative estimates of uncertainty by neglecting possible constraints from aspects of past change remote to the region of interest); the second version scaled future continental changes according to errors in past spatial and temporal patterns of change over the whole globe (thus obtaining narrower estimates of uncertainty, although this does not take account of possible errors in the regional pattern of response, since it scales the model's pattern of response over the whole globe by the same factor, with uncertainties, for each region).

Recent work has developed such techniques further to produce probabilistic predictions on sub-continental scales by taking fuller account of model uncertainty. The model set we analyse here consists of two ensembles of simulations; an ensemble of 17 simulations of HadCM3 with the same specification of anthropogenic and natural forcings but different perturbed parameter combinations (Harris et al., 2006) and a parallel ensemble of 17 simulations in which the same parameter combinations are used but greenhouse gas forcing is omitted from the forcings. Global scaling factors on the modeled response to greenhouse gas forcing and to other forcings are produced for each of the 17 models and these predictions are then combined in a weighted average to produce probabilistic predictions on sub-continental scale "Giorgi" regions. The results are shown in Figure 1 where the ASK predictions for the northern Europe region are compared with predictions based on the methodology used for producing probabilistic predictions for the UK in the UKCP09 project (ukclimateprojections.defra.gov.uk). The UKCP09 methodology weights model predictions based on measures of a model's ability to simulate mean climate as well as past climate, whereas the ASK methodology uses only past climate changes to weight predictions. The degree of agreement between the two approaches appears to indicate that on sub-continental scales future temperatures are observationally constrained largely by past changes over the globe rather than by mean climate. The observed temporal evolution of the land-ocean temperature contrast, of hemispheric temperature differences and of the ratio of high latitude to low latitude warming appear to provide important constraints on future warming and its regional patterns (Stott et al., 2006a). Further work will seek to improve our understanding of the processes that control the climate's response to past greenhouse gas forcing and other forcings in order to refine probabilistic estimates of future regional climate changes.

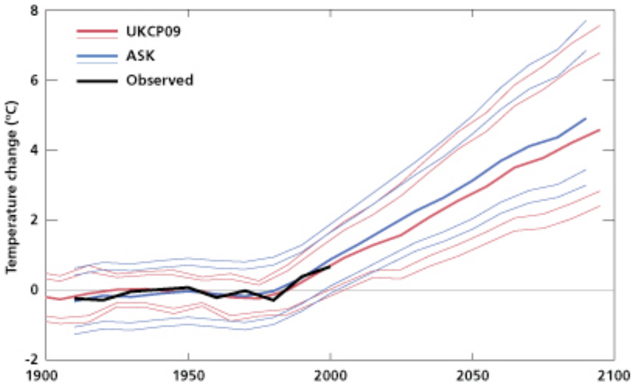


Figure 1. Comparison of probabilistic climate projections for changes in 10-year annual mean 1.5 m temperature (ºC) in response to SRES A1B emissions. Changes shown are for Northern Europe, relative to 1906-2005, from two methods: UKCP09 (red) and using ASK (blue). The probability levels are 2.5%, 10%, 50% (thick), 90%, and 97.5% as used in Stott et al. (2006a). The observations are also shown as the black line.

References

Allen, M.R., and S.F.B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Clim. Dyn.*, **15**, 419-434.

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

Allen, M.R., and P.A. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting, part I: theory. *Climate Dynamics*, **21**, 477-491, doi: 10.1007/s00382-003-0313-9.

Frame, D.J., D.A. Stone, P.A. Stott, and M.R. Allen, 2006: Alternative to stabilization scenarios. *Geophys. Res. Letters.*, **33**(5), L14707.

Harris, G.R., 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. *Climate Dynamics*, **27**, 357-375.

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., 2003: Attribution of regional scale temperature changes to natural and anthropogenic causes. *Geophys. Res. Lett.*, **21**, 493-500, doi:10.1029/2003GL017324.

Stott, P.A., J.F.B. Mitchell, M.R. Allen, T.L. Delworth, J.M. Gregory, G.A. Meehl, and B.D. Santer, 2006a: Observational constraints on past attributable warming and predictions of future global warming. *J. Climate*, **19**(13), 3055-3069.

Stott, P.A., J.A. Kettleborough, and M.R. Allen, 2006b: Uncertainty in predictions of continental scale temperature rise. *GRL*, **33**, doi:10.1029/GL0 24423.

Stott, P.A., C. Huntingford, C.D. Jones, and J.A. Kettleborough, 2008: Observed climate change constrains the likelihood of extreme global warming. *Tellus*, **60B**, 76-81.

Tett, S.F.B, P.A. Stott, M.R. Allen, W. Ingram, and J.F.B. Mitchell, 1999: Causes of twentieth century temperature change. *Nature*, **399**, 569-572.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Evaluations of CMIP3 Model Performances for Various Phenomena in the Atmosphere and Oceans, in the Present-Day Climate and in Future Projections

Yukari N. Takayabu

Center for the Climate System Research, University of Tokyo, Japan

Considering the global warming impacts on our society, changes in associated shorter-term or regional phenomena in the atmosphere and oceans, such as tropical cyclones, storm tracks, monsoon, ENSO, etc., can influence our lives in near future more directly than the gradual trends of temperature and/or precipitation. For the IPCC's Fourth Assessment Report (AR4), climate model experiments for the present day climate simulations and for the future projections are performed in various organizations worldwide, and collected for the Coupled Model Intercomparison Phase 3 (WCRP CMIP3). In order to extract more reliable information on future climate variability, skills of climate models are evaluated utilizing CMIP3 data.

We have carried out a study 'Evaluations of CMIP3 Model Performances for Various Phenomena in the Atmosphere and Oceans, in the Present-Day Climate and in Future Projections' under a project entitled 'Integrated Research on Climate Change Scenarios to Increase Public Awareness and Contribute to the Policy Process (FY2007-FY2011)' funded by the Ministry of Environment, Japan. In this study, multi climate model simulation data of CMIP3 are analyzed in eight subgroups allocated to 16 phenomena which are influential for our lives in Japan. Sixteen selected phenomena are listed as follows: 1) tropical cyclones, 2) intertropical convergence zones (ITCZ), 3) Pacific-Japan teleconnections, 4) storm tracks, 5) surface temperature variations, 6) rainfall distribution in relation to the El Nino and Southern Oscillation (ENSO), 7) rainfall characteristics in Baiu, 8) the ocean heat content and ENSO, 9) the Arctic Oscillation, 10) Pacific decadal oscillation, 11) Asian summer monsoon, 12) equatorial westerly bursts and ENSO, 13) the Madden-Julian oscillation, 14) cloud radiative forcing, 15) Baiu and Meiyu front and the subtropical high, 16) cloud amount and the large-scale circulation fields.

In the first year, metrics for reproducibilities of individual phenomena are suggested and the CMIP3 climate models are evaluated for 20th Century Climate in Coupled Model (20C3M) runs in comparison with the observations. In the second year, reproducibility metrics of various phenomena for 25 climate models are gathered from each subgruoup and synthesized. Based on these reproducibilities in the current climate, we evaluated future projections of some phenomena with reduced uncertainties compared to a single model or a simple multi-model projection.

Lastly, interrelationships among the reproducibilities of individual phenomena are examined. Tight relationships of their performances with the reproducibility of large-scale environmental fields are suggested for several phenomena. Following this result, we are now working on to produce the "Asian Metrics" to represent the performance of large-scale environments, which are essential for determining characteristics of short-term atmospheric and oceanic phenomena in the Asian region.

References

Ichikawa, H., H. Masunaga, and H. Kanzawa, 2009: Evaluation of precipitation and high-level cloud areas associated with large-scale circulation over the tropical Pacific in the CMIP3 models. *J. Meteor. Soc. Japan*, **87**, 771-789.

Inoue, T., and H. Ueda, 2009: Evaluation for the seasonal evolution of the summer monsoon over the Asian and western North Pacific sector in the WCRP CMIP3 multi-model experiments. *J. Meteor. Soc. Japan*, **87**, 539-560.

Kitoh, A., and T. Mukano, 2009: Changes in daily and monthly surface air temperature variability by multi-model global warming experiments. *J. Meteor. Soc. Japan*, **87**, 513-524.

Kosaka, Y., and H. Nakamura, 2008: A comparative study of the Pacific-Japan (PJ) teleconnection pattern based on reanalysis datasets. *SOLA*, **4**, 9-12, doi:10.2151/sola.2008-003.

Kosaka, Y., H. Nakamura, M. Watanabe, and M. Kimoto, 2009: Analysis on the dynamics of a wave-like teleconnection pattern along the summertime Asian jet based on a reanalysis dataset and climate model simulations. *J. Meteor. Soc. Japan*, **87**, 561-580.

Ninomiya, K., 2009: Characteristics of precipitation in the Meiyu-Baiu season in the CMIP3 20th century climate simulations. *J. Meteor. Soc. Japan*, 829-843.

Nishii, K., T. Miyasaka, Y. Kosaka, and H. Nakamura, 2009: Reproducibility and future projection of the midwinter storm-track activity over the Far East in the CMIP3 climate models in relation to "Haru-Ichiban" over Japan. *J. Meteor. Soc. Japan*, **87**, 581-588.

Ose, T., and O. Arakawa, 2009: Characteristics of the CMIP3 models simulating realistic response of tropical western Pacific precipitation to Niño3 SST variability. *J. Meteor. Soc. Japan*, **87**, 807-819.

Oshima, K., and Y. Tanimoto, 2009: An evaluation of reproducibility of the Pacific Decadal Oscillation in the CMIP3 simulations. *J. Meteor. Soc. Japan*, **87**, 755-770.

Sato, N., C. Takahashi, A. Seiki, K. Yoneyama, R. Shirooka, and Y.N. Takayabu, 2009: An evaluation of the reproducibility of the Madden-Julian oscillation in the CMIP3 multi-models. *J. Meteor. Soc. Japan*, **87**, 791-805.

Sueyoshi, M., and T. Yasuda, 2009: Reproducibility and future projection of the ocean first baroclinic Rossby radius based on the CMIP3 multi-model dataset. *J. Meteor. Soc. Japan*, **87**, 821-828.

Yokoi, S., and Y.N. Takayabu, 2009: Multi-model projection of global warming impact on tropical cyclone genesis frequency over the western North Pacific. *J. Meteor. Soc. Japan*, **87**, 526-538.

Yokoi, S., Y.N. Takayabu, and C.L. Chan, 2009: Tropical cyclone genesis frequency over the western North Pacific simulated in medium-resolution coupled general circulation models, *Climate Dyn.*, **33**, 665-683.

A Partial Explanation of the Apparent Superior Performance of the Multi-Model Mean Simulations

Karl E. Taylor

Lawrence Livermore National Laboratory, USA

Statistical measures that gauge differences between model-simulated and observed climatology often indicate that the mean fields computed from a multi-model ensemble are apparently in better accord with observations than any of the individual model fields comprising those means. We examine whether this result can be explained simply in terms of the smoother character of the mean field. We define a skill score that penalizes models that have unrealistically smooth fields, but even so the mean model appears to excel. We then show that the formation of a mean result tends to filter the shorter spatial scales preferentially. Although this filtering has little effect on the overall variance (and thus is penalized little by the skill score), it can reduce the RMS error in the field and improve its correlation with observations. If we apply a comparable effective filter to individual model results, we see their apparent skill improve. The degree to which the preferential smoothing of smaller spatial scales improves the apparent model skill depends on the field considered.

An Overview of Approaches to Future Projections Based on Multi-Model Ensembles

Claudia Tebaldi

Department of Statistics, University of British Columbia, Canada

In this talk, I will try to describe the assumptions and methods (and when possible the results) that have been proposed over the last ten years for the characterization of future projections and their uncertainty based on Multi-Model Ensembles (MMEs).

Starting from the work by Raisanen and Palmer (2001) that utilized runs from CMIP2 and coming to the latest papers on the subject (e.g., Annan and Hargreaves, 2010), how the data from MMEs has been interpreted in relation to the true climate change signal that we seek to characterize has varied: some methods have implicitly or explicitly considered each model's simulation as a possible future trajectory, and have built empirical histograms or reweighted versions of them as an approximation to the range and likely distribution of possible futures; other methods have focused on the idea of a common signal underlying all models' simulations, and have characterized this unobserved consensus' estimate and its uncertainty as if each of the members of the MME was a "truth+error" version of it.

There have also been approaches that have combined MMEs and other sources of information (observations, simplified model results, pattern scaling, perturbed physics experiments) in order to consider as comprehensive a representation of model results -- and their performance -- as our computational resources allow.

Among these latter approaches, many have attempted to utilize observational constraints to reweight individual MME members into a synthesis that reflects model performance. These attempts have been challenged by the complexity of model performance evaluation and the non-obvious relation between performance over past and current climate and reliability of future simulations. We will survey several proposals that have been put forward in order to quantify and /or rank MME members and utilize the result to construct future projections.

As I go along and describe these approaches I would like to focus on the underlying assumptions (some of which I just sketched) as the most relevant input to this workshop, as opposed to specific results. I will present however some comparison of actual results in order to exemplify how the shape and range of the final projections that ensue from each of these methods (embodying our best guesses of the future and their uncertainties) may be sensitive to some of these assumptions – or not.

Following is a list of papers relevant to my discussion. If I missed some relevant work at this point it is only because of oversight, and I hope to have a more complete list by the time of the workshop.

Bibliography (in progress)

Annan J., and J. Hargreaves, 2010: Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, doi:10. 1029/2009GL041994,

Benestad, R., 2004: Tentative probabilistic temperature scenarios for northern Europe. *Tellus A* **56** 89-101.

Berliner, L.M., and Y. Kim, 2008: Bayesian design and analysis for superensemble-based climate forecasting. *J. Clim.*, **21**(9), 1891-1910.

Brekke, L.D. , M.D. Dettinger, E.P. Maurer, and M. Anderson, 2008: Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments *Clim. Ch,.* **89**(3-4), 371-394.

Buser, C.M., H.R. Kunsch, D. Luthi, M. Wild, and C. Schar, 2009: Bayesian multi-model projection of climate: bias assumptions and interannual variability. *Clim. Dyn.,* **33**(6), 849-868.

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

Furrer, R., S. Sain, D. Nychka,and G. Meehl, 2007: Multivariate Bayesian analysis of atmosphere-ocean general circulation models. *Environ. Ecol. Stat.*, **14**(3), 249-266.

Giorgi, F., 2008: A simple equation for regional climate change and associated uncertainty. *J. Clim.*, **21**(7), 1589-1604.

Giorgi, F., and L. 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.

Giorgi, F., and L. Mearns, 2003: Probability of regional climate change calculated using the reliability ensemble average (REA) method. *Geophys. Res. Lett.*, **30**, 1629-1632.

Gleckler, P.J., K.E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *JGR-Atmosphere* **113**, D06104.

Greene, A., L. Goddard, and U. Lall, 2006: Probabilistic multimodel regional temperature change projections. *J. Clim.,* **19**, 4326-4343.

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**, L08708.

Murphy, J.M., B.B.B. Booth, M. Collins, et al., 2007: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Phil. Trans. of the Royal Soc. A.,* **365**(1857), 1993-2028.

Palmer, T.N., F.J. Doblas-Reyes, R. Hagedorn, and A. Weisheimer, 2005b: Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Phil. Trans. R. Soc. B.*, **360**, 1991-1998.

Perkins, S.E., and A.J. Pitman, 2009: Do weak AR4 models bias projections of future climate changes over Australia? *Clim. Ch.,* **93**(3-4), 527-558.

Pierce, D.W., T.P. Barnett. B.D. Santer, et al., 2009: Selecting global climate models for regional climate change studies. *PNAS*, **106**(21), 8441-8446.

Raisanen, J. and T.N. Palmer, 2001: A probability and decision-model analysis of a multimodel ensemble of climate change simulations. *J. Clim.*, **14**, 3212–3226.

Rougier, J., 2007: Probabilistic inference for future climate using an ensemble of climate model evaluations. *Clim. Change*, **81**(3) 247-264.

Smith, R., C. Tebaldi, D. Nychka, and L. Mearns, 2009: Bayesian modeling of uncertainty in ensembles of climate models. *JASA*, **104**(485), 97-116.

Suppiah, R., K. Hennessy, P.H. Whetton, et al., 2007: Australian climate change projections derived from simulations performed for the IPCC 4th Assessment Report. *Australian Meteorological Magazine,* **56**(3), 131-152.

Tebaldi, C., and B. Sanso, 2008: Joint projections of temperature and precipitation change by synthesizing multi-model ensembles in a Bayesian hierarchical framework. J. *Royal Stat. Soc., Ser. A.*, **172**, 83-106.

Tebaldi, C., R. Smith, D. Nychka, and L. 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.

Watterson, I., 2008: Calculation of probability density functions for temperature and precipitation change under global warming. *JGR-Atmosphere*, **113**, D12106.

Wigley, T.M.L. and S.C.B. Raper, 2001: Interpretation of high projections for global-mean warming. *Science*, **293**, 451-454.

Assessing Future Projections from Models with Differing Climate Sensitivities

Michael Wehner

Lawrence Berkeley National Laboratory, USA

Upcoming climate change assessment reports will be expected to discriminate among models in their estimates of future climate change based on their abilities to reproduce observed climate statistics. Different climate models exhibit different sensitivities to external forcing for a multitude of reasons. This presents difficulties in making unbiased ensemble projections if the climate sensitivities of the acceptable models are not representative of the accepted range of climate sensitivity. Using the ability of the CMIP3 models to reproduce the Palmer Drought Sensitivity Index (PDSI) as an example, we illustrate a method to more properly frame a limited set of future projections.

In a forthcoming paper, we analyzed the ability of nineteen climate models to simulate PDSI over the North America. We found that models more skillful in their ability to reproduce observed drought statistics of the second half of the twentieth century tend to exhibit less severe projections of future drought under the SRES A1B emissions scenario. The reason for this is an apparent correlation between drought skill and climate sensitivity, with the models projecting the least amount of warming being the best at reproducing recent drought statistics. Because there is little reason to believe that this is a fundamental result, projections of future drought in a specified time period are biased low if one accepts the full range of climate sensitivity estimates. A more unbiased way of approaching the problem is to express the change in PDSI at a specific warming amount rather than at a specific time. This entails identifying when each model projects such a specific warming amount and combining the drought estimates among models from these presumably different time intervals.

Risks of Model Weighting in Multi-Model Climate Projections

Andreas P. Weigel

Federal Office of Meteorology and Climatology, MeteoSwiss, Switzerland

In seasonal forecasting, performance-based weighting schemes have been successfully implemented and have been demonstrated to improve the average prediction skill significantly (e.g., Weigel et al., 2008). Such weighting schemes are typically based on 20 to 40 years of independent hindcast data, which mimic real forecasting situations and can thus serve as a data basis to derive optimum weights. However, if only an insufficient number of hindcast data are available (e.g., 10 years or less), the weight estimates get less robust and skill may drop below the level obtained with unweighted multi-models (Weigel et al., 2010). In other words, model weighting in seasonal forecasting is only successful, if the weights are robust enough to represent the true underlying model skill.

What does this imply for the combination of climate change projections? What would be the consequences in terms of projection accuracy, if "wrong" weights are applied, i.e., weights which do not represent the true model performance? This is an important question to be discussed, given that at present there is no consensus on how model weights should be obtained in a climate change context (e.g., Knutti et al., 2010), nor is it clear that appropriate weights can be obtained at all with the data and methods at hand.

To address these questions, we have recently introduced a simple conceptual model of climate change projections (Weigel et al., 2010). This "toy model" allows us to analyze the effects of equal, optimum and inappropriate weighting in generic terms by controlled combination experiments. The model is designed such that the impacts of (i) model error magnitude, (ii) error correlation, and (iii) internal variability can be considerd. The key results, many of which are consistent with experience in seasonal forecasting, can be summarized as follows:

- Equally weighted multi-models yield, on average, more accurate projections than do the participating single models alone. The projection errors can be further reduced by model weighting, assuming the optimum weights are known.

- The optimum weights are not only a function of the single model error uncertainties, but also depend on the degree of model error correlation and the amount of internal variability. Neglecting internal variability and model error correlation can lead to severely biased estimates of optimum weights.

- If model weights are applied which do not reflect the true model uncertainties, then the weighted multi-model may have much lower skill than the unweighted one. In many cases more information may actually be lost by inappropriate weighting than can potentially be gained by optimum weighting (Figure 1).

- This asymmetry between potential loss due to inappropriate weights and potential gain due to optimum weights grows under the presence of internal variability. In fact, if the internal variability is of comparable or even larger magnitude than the model errors, then equal weighting essentially becomes the optimum way to construct a multi-model.

These results do not imply that the derivation of performance based weights is impossible by principle. However, our results do imply that a decision to weight climate models should be taken with great care. Unless there is a clear relation between what we observe and what we predict, the risk of reducing the projection accuracy by inappropriate weights appears to be higher than the prospect of improving it by optimum weights. Given the current difficulties in determining reliable weights, equal weighing may for many applications well be the safer and more transparent way to.

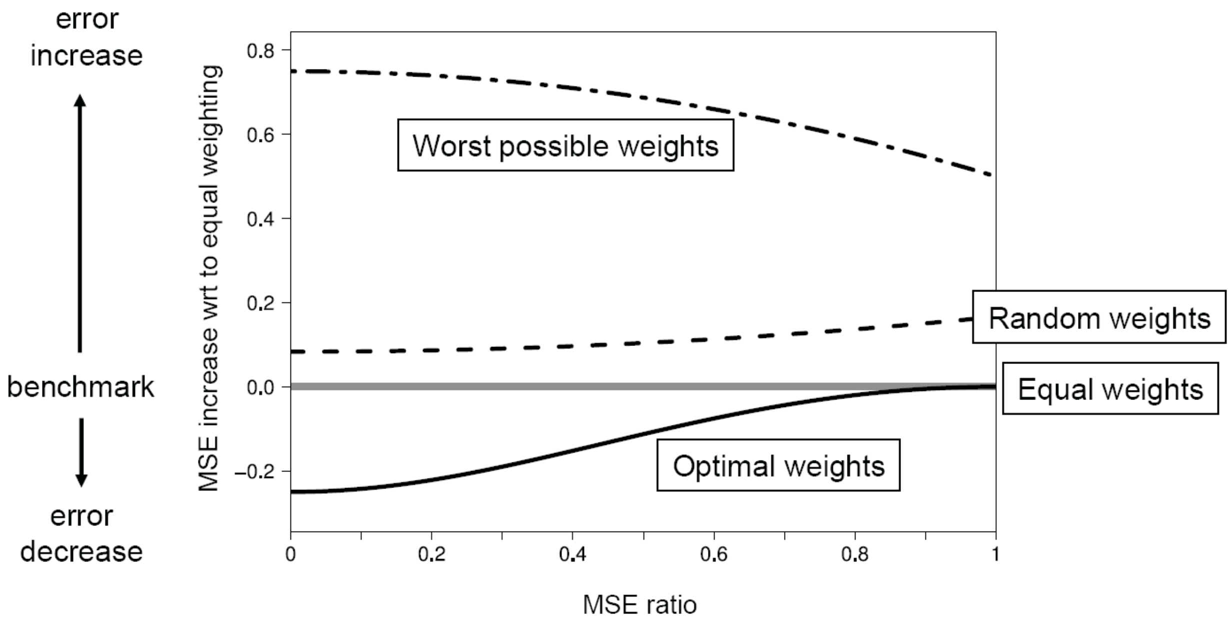


Figure 1. Increase/decrease of the expected mean squared error (MSE) of weighted averages of two single models (solid black: optimum weights; dot-dashed: worst possible weights; dashed: random weights) with respect to the benchmark of equal weighting. The results are plotted as a function of the MSE ratio of the two single models to be combined. The combination experiments are based on the conceptual model of Weigel et al. (2010).

References

Knutti R., R. Furrer, C. Tebaldi, and J. Cermak, 2010: Challenges in combining projections from multiple climate models. *J. Clim*, submitted

Weigel A.P., R. Knutti, M.A. Liniger, and C. Appenzeller, 2010: On the risk of applying unrobust model weights in multimodel climate change projections. *J. Clim.*, submitted

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.

IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections
Boulder, Colorado, USA, 25-27 January 2010
Conference Volume

Regional Projections and Model Evaluation: Potential Benefits of 'Representative Future Regional Climates'

Penny Whetton, Kevin Hennessy, Bryson Bates, David Kent

CAWCR and CSIRO Marine and Atmospheric Research, Australia

A key challenge for the science of climate projection and impact assessment is to serve the rapidly growing climate information needs of adaptation planning in an environment where the following apply:

- A perception that impact assessments based on older GCMs and emission scenarios become 'out-of-date' and need to be re-done as new modelling results become available;
- The view that substantial differences between GCMs in simulated future regional climate is a major challenge for adaptation planning (see analysis by Dessai et al., 2009)
- A common desire amongst many users to use a few as possible future climate scenarios in impact assessments (e.g., AGO, 2006), often a 'best guess' assessment, just when there is rapid growth in the number of potentially relevant GCM and downscaled results available for application (e.g., CMIP5);
- Standard methods for producing probabilistic projections (which can synthesis results from multiple GCMs into a PDF) and applying them in impact studies are yet to be established.

If effective methods of filtering out unreliable models were to be found, this could reduce uncertainty, new model runs could be routinely assessed as they come along, and a small set of simulations could be used in impact applications, simplifying this work. This view has stimulated much recent interest in model evaluation in the Australian context (e.g., see Smith and Chandler 2009), as equivalent pressures have done so internationally (e.g., this workshop). There are two problems with this response: first, there is no consensus that we have yet found, or are about to find, a robust process for filtering available models down to a few 'best' models (e.g. see Knutti et al., in press). Second, even if we could identify the best models, this need not serve the requirements of adaptation planning, such as the 'robust decision making' perspective (see Dessai et al., 2009) in which the full range of plausible future climates, and not just the more likely future climates, need to be considered in the impact assessment. In other words, low probability scenarios with high impact should not be ignored in risk assessment. A further complexity is that model evaluation has also been taken up by researchers primarily with application interests (e.g., Chiew et al.,, 2009). Although the impact perspective is a vital consideration in the development of relevant regional projection products, user involvement in model evaluation increases the likelihood that models will be selected on local current climate realism criteria, which may or may not relate to reliability of the anthropogenically forced change.

Our assessment is that model evaluation by itself is unlikely to adequately address the issues described above. If the simplicity that is craved in this area is going to be achieved, it will have to be reached some other way.

'Representative concentration pathways' have recently been adopted as a basis for running GCM simulations for the IPCC AR5 (Moss et al., 2008). This approach recognizes that the climate system responds to the evolving greenhouse gas and aerosol concentrations of the atmosphere, and does not need to 'know' about the varying socio-economic scenarios that may underlie any particular concentration pathway. This means that as long as a suitably wide set of pathways is developed, climate modelers can apply these in new simulations for the AR5, while in parallel, the integrated assessment modelers can assess potential socio-economic scenarios and estimate which of the concentration pathways any scenario is most likely to follow. This movement from a linear approach to a parallel approach shortens the

2030 A1B				
	Little change up to 0.5 °C warmer	Warmer 0.5 to 1.5 °C warmer	Hotter 1.5–3.0 °C warmer	Much hotter more than 3.0 °C warmer
Much wetter (more than +15%)	*No evidence*	*No evidence*	*No evidence*	*No evidence*
Wetter (0 to 15% wetter)	*No evidence*	Unlikely *5 models*	*No evidence*	*No evidence*
Drier (0 to 15% drier)	Very unlikely *GISS AOM, PCM*	Likely *16 models*	*No evidence*	*No evidence*
Much drier (More than 15% drier)	*No evidence*	*No evidence*	*No evidence*	*No evidence*
2070 A1FI				
Much wetter (more than +15%)	*No evidence*	*No evidence*	*No evidence*	*No evidence*
Wetter (0 to 15% wetter)	*No evidence*	*No evidence*	Unlikely *4 models*	Very unlikely *(CGM3.1 T47)*
Drier (0 to 15% drier)	*No evidence*	Very unlikely *(GISS AOM)*	As likely as not *10 models*	Unlikely *3 models*
Much drier (More than 15% drier)	*No evidence*	*No evidence*	Very unlikely *CNRM-CM3,CSIROmk3*	Very unlikely *CSIRO Mk3.5, IPSL*

Figure 1. Example of RFRCs based on annual temperature and precipitation change for Victoria, Australia. Calculated from 23 CMIP3 simulations for two time slices and selected emission scenarios (using pattern scaling, see Mitchell, 2003), and with ranges of change arbitrarily chosen. The total model count is shown, or, for counts of two models or less, the models named. Likelihood terminology is consistent with the IPCC, based on the percentage of models in a given square (Risbey and Kandlikar, 2007) (e.g., 'likely' is >66%). Where no square is 'likely' the heavy red border links squares that together would reach the 'likely' threshold.

development time for the production of relevant assessment work for the IPCC and also provides an ongoing framework for relating socio-economic scenarios to available GCM runs.

Here we extend the above concept to consider the idea of sets of 'representative future regional climates (RFRCs)' for any region of concern. If these could be developed, they would greatly help address the challenge and associated issues identified in the first two paragraphs above. This perspective would not sideline model evaluation, although it would redirect this work into the dual goals of 1) identifying plausible (broadly defined) regional climates and their likelihood (which will depend on assessing model processes driving particular simulated regional responses), and 2) assessing the applicability of particular models for specific applications (which is likely to depend on local realism).

To construct RFRCs, we would need to consider changes to a range of variables commonly needed in impact assessments, such as changes to annual and seasonal mean temperature, rainfall, potential evaporation, solar radiation and windspeed, and daily extremes such as heavy rainfall, hot days and strong winds. A set of such climates could be perhaps be found using cluster analysis, or related data reduction techniques, but for illustrative purposes here we simply classify simulated future regional climates by the changes to annual rainfall and temperature. This choice is guided by fact that these variables are the two most commonly required in impact assessment as well as being the two most commonly reported in the literature. We also expect that temperature and rainfall changes effectively classify changes in a range of other variables (e.g., extreme rainfall). Figure 1 gives an example of such a classification for a selected region based on results from the CMIP3 archive. A set of plausible future climates is indicated, and we see how the hotter and drier climates are more likely to occur at a later date and under the higher emission scenario. Relative likelihoods, based density of model results, are also indicated (see caption).

The RFRCs would be defined by their high level descriptions (in this case, 'warmer and much drier', etc.) but under these there can be a richer level of detail based on wider results of the models that fall in a classification, such as the seasonal distribution of the changes or changes to other variables (which are also likely to further discriminate the climates from each other, although this needs investigation). The expansion of an RFRC may include reference to various future climate data sets (such as downscaled data) that fall under that classification and possibly impact results that have employed those data sets. Likelihood assessment could be modified to account for assessed model reliability (through weighting or model exclusion), or a range of other potentially relevant considerations such as physical arguments (e.g., expansion of Hadley Cell) or recent observed climate trends (extrapolation of observed trends could be entered as a plausible scenario). Application scientists engaging with a set of RFRCs may choose to use a subset, such as the 'likely' RFRC plus, say, two RFRCs that might lead to 'extreme risk' for the impact system under consideration, or a 'least change' RFRC. Future climate data sets suitable for use in the application need only be developed (through selecting GCMs, downscaling etc.) to populate those cases. One or two more cases may need to be added to cover the needs of a number of impact sectors if a general purpose set of RFRCs is to be prepared.

Systematic application of this approach would require various challenges to be addressed, such as robustly classifying future regional climates into a small high level set, estimating likelihoods, and deciding on suitable regionalisations. However, if this approach could be established, the needs of the robust decision making perspective could be efficiently addressed, while still benefiting from assessment of the likelihood of various future climates. Furthermore, RFRCs, and not GCMs, could be the 'boundary objects' (Hulme and Dessai, 2008) which anchor discussion between climate science and impact communities. Since the high level RFRC descriptions need not change as new GCM results emerge (although the RFRC likelihoods, and the datasets the RFRCs are populated with, will evolve), they can provide a framework for assimilating impact assessments undertaken at different times with different sets of GCMs.

References

AGO, 2006: *Climate Change Impacts & Risk Management: A Guide for Business and Government*. Published by the Australian Greenhouse Office, in the Department of the Environment and Heritage.

Chiew, F.H.S., J. Teng, J. Vaze, and D.G.C. Kirono, 2009: Influence of global climate model selection on runoff impact assessment. *Journal of Hydrology*, **379**, 172–180

Dessai, S., M. Hulme, R. Lempert, and R. Pielke, Jr., 2009: *Climate prediction: a limit to adaptation? Adapting to Climate Change: Thresholds, Values, Governance*. Published by Cambridge University Press. pp. 64-78.

Hulme, M., and S. Dessai, 2008: Negotiating future climates for public policy: a critical assessment of the development of climate scenarios for the UK. Science Direct, *Environmental Science & Policy*, **11**, 54-70.

Knutti, R. et al., Challenges in combining projections from multiple models, *J. Clim.*, in press.

Mitchell, T.D., 2003: Pattern scaling. *Climatic Change*, **60**, 217-242.

Moss, R.H., et al., 2008: *Towards New Scenarios for Analysis of Emissions, Climate Change, Impacts, and Response Strategies*. IPCC, Geneva, Switzerland, 132 pp.

Risbey, J., and M. Kandlikar, 2007: Expressions of likelihood and confidence in the IPCC uncertainty assessment process. *Climatic Change*, **85**(1-2), 19-31.

Smith, I., and E. Chandler, 2009: Refining rainfall projections for the Murray Darling Basin of southeast Australia: the effect of sampling model results based on performance. *Climatic Change*, doi :10.1007/s10584-009-9757-1.

Reconstruction, Projection and Detection

Francis Zwiers

Environment Canada, Canada

There are important parallels between the millennial climate reconstruction problem on the one hand, and the climate projection problem on the other.

In the millennial reconstruction problem one has available (a) a network of instrumental observations that extend over the past 100-150 years, (b) a network of annually or seasonally resolved climate proxies extending back in time over one to two millennia from recent decades, and (c) perhaps also the output of one or more climate models that have been run over the millennial period using reconstructed forcings. The reconstruction problem is one where information from one source (proxies) is related to information from another source (instruments) over the instrumental period, and that relationship is then applied to the pre-instrumental period values of the proxies to estimate what the instruments would have said in the past had they been present. The quality of the reconstruction and the ability to assess uncertainties depends at least, in part, on the statistical method that is used to relate proxies to instruments. The reconstruction may benefit from additional constraints from climate models if output from one or more climate models is also built into the reconstruction process (e.g., via a Kalman Filter, as was recently demonstrated by Lee et al., 2008). When undertaken in this way, the reconstruction process may simultaneously allow detection and attribution, and may provide the possibility of making projections forward in time that are constrained by historical relationships and perhaps by pre-instrumental proxy behaviour.

In the climate projection problem one has available (a) a network of instrumental observations that extend over the past 100-150 years, (b) ensembles of historical climate simulations of the instrumental period with estimated historical forcing, and (c) additional ensembles of climate simulations of the next one-to-two centuries that have been driven with a forcing scenario. Detection and attribution studies, which analyze the relationship between the instrumental observations and the historical climate simulations, attempt to separate the observed climate fields into components that can be attributed as responses to forcing and components that are consistent with internal climate variability. The statistical relationship that is established in this way between the estimated responses to forcing and observed climate changes can subsequently be exploited to constrain projections of future change (e.g., Allan and Ingram, 2002), in much the same way that the link between proxies and instruments is used to interpret the pre-instrumental variations of the proxies.

Similarities and differences between these two "projection" problems are contrasted and discussed. While the signal of interest is different in the two cases, there are generally considerable similarities. Detection and attribution is involved, either implicitly or explicitly, in both cases. Also, in both cases, the statistical models that are developed can be used to contribute to the elaboration of uncertainties in the "projected" temperatures. Further, the effectiveness of the constraint that is imposed on the "projected" temperatures by the statistical model outside the instrumental period presumably depends upon whether the instrumental period provides enough variation to effectively train the statistical models used for projection. Finally, in both cases, methods are often used that seek to optimize signal-to-noise ratio, and in doing so, avoid drawing upon aspects of the response to forcing, or aspects of the proxies, where information is limited.

References

Allen, M.R., and W.J. Ingram, 2002: Constraints on future changes in climate and the hydrologic cycle. *Nature*, **419**, 224–232.

Lee, T.C.K., F.W. Zwiers, and M. Tsao, 2008: Evaluation of proxy-based millennial reconstruction methods. *Climate Dynamics*, **31**, 263-281, doi:10.1007/s00382-007-0351-9.

Annex 4: Participant List

Gabriel ABRAMOVITZ
Climate Change Research Centre
University of New South Wales
AUSTRALIA

Myles ALLEN
Atmospheric, Oceanic and Planetary Physics
University of Oxford
UNITED KINGDOM

Bryson BATES
Commonwealth Scientific and Industrial Research Organization
Marine and Atmospheric Research
AUSTRALIA

Rasmus BENESTAD
Climate Department
Norwegian Meteorological Institute
NORWAY

Sandrine BONY-LENA
Institut Pierre Simone Laplace
Université Pierre et Marie Curie
FRANCE

Timothy CARTER
Finnish Environment Institute
FINLAND

Ole Bøssing CHRISTENSEN
Danish Climate Centre
Danish Meteorological Institute
DENMARK

Matthew COLLINS
Hadley Centre for Climate Prediction and Research
Met Office
UNITED KINGDOM

Wolfgang CRAMER
Department of Global Change and Natural Systems
Potsdam Institute for Climate Impact Research
GERMANY

Francisco DOBLAS-REYES
Catalan Institute for Climate Sciences
SPAIN

Kristie EBI
IPCC WGII TSU
USA

Seita EMORI
Center for Global Environmental Research
National Institute for Environmental Studies
JAPAN

Veronika EYRING
Deutsches Zentrum für Luft- und Raumfahrt
Institut für Physik der Atmosphäre
GERMANY

Christopher FIELD
IPCC WGII Co-Chair
USA

Pierre FRIEDLINGSTEIN
Institut Pierre Simon Laplace
Laboratoire des Sciences du Climat et de l'Environnement
FRANCE

Marjorie FRIEDRICHS
Department of Biological Sciences
Virginia Institute of Marine Science
USA

Xuejie GAO
National Climate Centre
China Meteorological Administration
CHINA

Nathan GILLETT
CCCma
Environment Canada
University of Victoria
CANADA

Peter GLECKLER
Program for Climate Model Diagnosis and Intercomparison
Lawrence Livermore National Laboratory
USA

Jonathan GREGORY
University of Reading and
Hadley Centre for Climate Prediction and Research
Met Office
UNITED KINGDOM

Eric GUILYARDI
Institut Pierre Simon Laplace
Laboratoire d'Océanographie et du Climat
FRANCE

Alex HALL
Department of Atmospheric Sciences
University of California Los Angeles
USA

Gabriele HEGERL
School of GeoSciences
University of Edinburgh
UNITED KINGDOM

Isaac M. HELD
Geophysical Fluid Dynamics Laboratory
National Oceanic and Atmospheric Administration
USA

Bruce HEWITSON
Department of Environmental & Geographical Sciences
University of Cape Town
SOUTH AFRICA

Charles JACKSON
Institute for Geophysics
The University of Texas at Austin
USA

Vladimir KATTSOV
Voeikov Main Geophysical Observatory
RUSSIA

Jeffrey T. KIEHL
Climate Change Research Section
National Center for Atmospheric Research
USA

Reto KNUTTI
Institute for Atmospheric and Climate Science
ETH Zürich
SWITZERLAND

Won-Tae KWON
Climate Research Laboratory
National Institute of Meteorological Research
Korea Meteorological Administration
KOREA

Hong LIAO
State Key Laboratory of Atmospheric Boundary Layer
Physics and Atmospheric Chemistry
Institute of Atmospheric Physics
Chinese Academy of Sciences
CHINA

Victor MAGAÑA RUEDA
Centro de Ciencias de la Atmósfera
Ciudad Universitaria
Universidad Nacional Autónomia de México
MEXICO

Josè MARENGO
Centro de Previsão de Tempo e Estudos Climáticos
Instituto Nacional de Pesquisas Espaciais
BRAZIL

Linda MEARNS
National Center for Atmospheric Research
USA

Gerald MEEHL
Climate and Global Dynamics Division
National Center for Atmospheric Resarch
USA

Pauline MIDGLEY
IPCC WGI TSU
SWITZERLAND

Wendy PARKER
Department of Philosophy
Ohio Center for Ecology and Evolutionary Studies
Ohio University
USA

Sarah PERKINS
Commonwealth Scientific and Industrial Research
Organization
Marine and Atmospheric Research
AUSTRALIA

David PIERCE
Division of Climate, Atmospheric Science and Physical Oceanography
Scripps Institution of Oceanography
USA

Gian-Kasper PLATTNER
IPCC WGI TSU
SWITZERLAND

Dahe QIN
IPCC WGI Co-Chair
CHINA

Benjamin SANTER
Program for Climate Model Diagnosis and Intercomparison
Lawrence Livermore National Laboratory
USA

Gavin SCHMIDT
NASA Goddard Institute for Space Studies and
Center for Climate Systems Research
Columbia University
USA

David SEXTON
Hadley Centre for Climate Prediction and Research
Met Office
UNITED KINGDOM

Thomas STOCKER
IPCC WGI Co-Chair
SWITZERLAND

Daíthí STONE
Climate Systems Analysis Group
University of Cape Town
SOUTH AFRICA

Peter STOTT
Hadley Centre for Climate Monitoring and Attribution
Met Office
UNITED KINGDOM

Yukari TAKAYABU
Center for Climate System Research
University of Tokyo
JAPAN

Karl TAYLOR
Program for Climate Model Diagnosis and Intercomparison
Lawrence Livermore National Laboratory
USA

Claudia TEBALDI
Department of Statistics
University of British Columbia
CANADA

Melinda TIGNOR
IPCC WGI TSU
SWITZERLAND

Thuc TRAN
Ministry of Natural Resources and Environment
Institute of Meteorology Hydrology and Environment
VIETNAM

Yawei WANG
China Meteorological Administration
CHINA

Michael WEHNER
Scientific Computing Group
Lawrence Berkley National Laboratory
USA

Andreas WEIGEL
NCCR Climate
MeteoSwiss
SWITZERLAND

Penny WHETTON
Commonwealth Scientific and Industrial Research Organization
Marine and Atmospheric Research
AUSTRALIA

Francis ZWIERS
IPCC WGI Vice-Chair
Climate Research Division
Environment Canada
CANADA

Annex 5: Bibliography

This non-exhaustive list of references is provided by the participants of the Expert Meeting on Multi Model Evaluation as a resource for the reader.

Abe, M., H. Shiogama, J.C. Hargreaves, J.D. Annan, T. Nozawa, and S. Emori, 2009: Correlation between inter-model similarities in spatial pattern for present and projected future mean climate. *SOLA,* **5**, 133-136.

Allen, M.R., and W.J. Ingram, 2002: Constraints on future climate change and the hydrologic cycle. *Nature*, **419**, 224-233.

Allen, M.R., P.A. Stott, J.F.B. Mitchell, R. Schnur, and T. 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**, L02703.

Barnett, T.P., et al., 1999: Detection and Attribution of Recent Climate Change: A Status Report. *Bull. Am. Meteorol. Soc.,* **80**, 2631-2660.

Benestad, R.E., 2004: Tentative probabilistic temperature scenarios for northern Europe. *Tellus,* **56A(2)**, 89-101.

Benestad, R.E., 2005: Climate change scenarios for northern Europe from multi-model IPCC AR4 climate simulations. *Geophys. Res. Lett.,* **32**, L17704.

Benestad, R.E., 2006: Can We Expect More Extreme Precipitation on the Monthly Time Scale? *J. Clim.*, **19(4)**, 630-637.

Berliner, L.M., and Y. Kim, 2008: Bayesian Design and Analysis for Superensemble-Based Climate Forecasting. *J. Clim.*, **21(9)**, 1891-1910.

Boulanger, J.P., F. Martinez, and E.C. Segura, 2007: Projection of future climate change conditions using IPCC simulations, neural networks and Bayesian statistics. Part 2: Precipitation mean state and seasonal cycle in South America. *Clim. Dyn.*, **28(2-3)**, 255-271.

Brekke, L.D., M.D. Dettinger, E.P. Maurer, and M. Anderson, 2008: Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments. *Clim. Change*, **89(3-4)**, 371-394.

Buser, C.M., H.R. Kunsch, D. Luthi, M. Wild, and C. Schar, 2009: Bayesian multi-model projection of climate: bias assumptions and interannual variability. *Clim. Dyn.*, **33(6)**, 849-868.

Coelho, C.A.S., S. Pezulli, M. Balmaseda, F.J. Doblas-Reyes, and D.B. Stephenson, 2004: Forecast Calibration and Combination: A Simple Bayesian Approach for ENSO. *J. Clim.,* **17**, 1504-1516.

Collins, M., 2007: Ensembles and probabilities: a new era in the prediction of climate change. *Phil. Trans. R. Soc. A,* **365**, 1957-1970.

Dessai, S., X.F. Lu, and M. Hulme, 2005: Limited sensitivity analysis of regional climate change probabilities for the 21st century. *J. Geophys. Res.,* **110(D19)**, D19108.

Dettinger, M.D., 2006: A component-resampling approach for estimating probability distributions from small forecast ensembles. *Clim. Change,* **76(1-2)**, 149-168.

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**, L14707.

Frame, D.J., N.E. Faull, M.M. Joshi, and M.R. Allen, 2007: Probabilistic climate forecasts and inductive problems. *Phil. Trans. R. Soc. A,* **365(1857)**, 1971-1992.

Furrer, R., S.R. Sain, D. Nychka, and G.A. Meehl, 2007: Multivariate Bayesian analysis of atmosphere - Ocean general circulation models. *Environ. Ecol. Stat.,* **14(3)**, 249-266.

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(6)**, L06711.

Giorgi, F., 2008: A Simple Equation for Regional Climate Change and Associated Uncertainty. *J. Clim.,* **21(7)**, 1589-1604.

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(10)**, 1141-1158.

Greene, A.M., L. Goddard, and U. Lall, 2006: Probabilistic Multimodel Regional Temperature Change Projections. *J. Clim.*, **19(17)**, 4326-4343.

Hegerl, G.C., P.A. Stott, M.R. Allen, J.F.B. Mitchell, S.F.B. Tett, and U. Cubasch, 2000: Detection and attribution of climate change: Sensitivity of results to climate model differences. *Clim. Dyn.*, **16**, 737-754.

Hulme, M., and S. Dessai, 2008: Predicting, deciding, learning: can one evaluate the 'success' of national climate scenarios? *Environ. Res. Lett.*, **3**, 045013.

Jackson, C.S., 2009: Use of Bayesian Inference and Data to Improve Simulations of Multi-physics Climate Phenomena. *J. Phys. Conf. Ser.*, **180**.

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(24)**, 6698-6709.

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

Knutti, R., 2010: The end of model democracy? *Clim. Change*, doi:10.1007/s10584-010-9800-2.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G.A. Meehl, 2010: Challenges in Combining Projections from Multiple Climate Models. *J. Clim.*, **23**, 2739-2758.

Laurent, R., and X.M. Cai, 2007: A maximum entropy method for combining AOGCMs for regional intra-year climate change assessment. *Clim. Change*, **82(3-4)**, 411-435.

Lempert, R., N. Nakicenovic, D. Sarewitz, and M. Schlesinger, 2004: Characterizing climate-change uncertainties for decision-makers. *Clim. Change*, **65**, 1-9.

Lopez, A., C. Tebaldi, M. New, D. Stainforth, M.R. Allen, and J.A. Kettleborough, 2006: Two Approaches to Quantifying Uncertainty in Global Temperature Changes. *J. Clim.*, **19(19)**, 4785-4796.

Min, S.K., and A. Hense, 2007: Hierarchical evaluation of IPCC AR4 coupled climate models with systematic consideration of model uncertainties. *Clim. Dyn.*, **29(7-8)**, 853-868.

Min, S.K., D. Simonis, and A. Hense, 2007: Probabilistic climate change predictions applying Bayesian model averaging. *Phil. Trans. R. Soc. A*, **365(1857)**, 2103-2116.

Moise, A.F., and D.A. Hudson, 2008: Probabilistic predictions of climate change for Australia and southern Africa using the reliability ensemble average of IPCCCMIP3 model simulations. *J. Geophys. Res.*, **113(D15)**, D15113.

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. R. Soc. A*, **365(1857)**, 1993-2028.

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. *Phil. Trans. R. Soc. B*, **360**, 1991-1998.

Palmer, T.N., F.J. Doblas-Reyes, A. Weisheimer, and M.J. Rodwell, 2008: Toward Seamless Prediction: Calibration of Climate Change Projections Using Seasonal Forecasts. *Bull. Am. Meteorol. Soc.*, **89(4)**, 459-470.

Perkins, S.E., and A.J. Pitman, 2009: Do weak AR4 models bias projections of future climate changes over Australia? *Clim. Change*, **93(3-4)**, 527-558.

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. U.S.A.*, **106(21)**, 8441-8446.

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(12)**, 1-50.

Räisänen, J., 2007: How reliable are climate models? *Tellus*, **59A**, 2-29.

Räisänen, J., and T.N. Palmer, 2001: A Probability and Decision-Model Analysis of a Multimodel Ensemble of Climate Change Simulations. *J. Clim.*, **14(15)**, 3212-3226.

Räisänen, J., and L. Ruokolainen, 2006: Probabilistic forecasts of near-term climate change based on a resampling ensemble technique. *Tellus*, **58(4)**, 461-472.

Reichler, T., and J. Kim, 2008: How Well Do Coupled Models Simulate Today's Climate? *Bull. Am. Meteorol. Soc.*, **89**, 303-311.

Reifen, C., and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill? *Geophys. Res. Lett.*, **36**, L13704.

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

Ruosteenoja, K., H. Tuomenvirta, and K. Jylha, 2007: GCM-based regional temperature and precipitation change estimates for Europe under four SRES scenarios applying a super-ensemble pattern-scaling method. *Clim. Change*, **81**, 193-208.

Santer, B.D., et al., 2009: Incorporating model quality information in climate change detection and attribution studies. *PNAS,* **106**, 14778-14783.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter, and D. Paolino, 2006: Climate model fidelity and projections of climate change. *Geophys. Res. Lett.,* **33(7)**, L07702.

Smith, R.L., C. Tebaldi, D. Nychka, and L.O. Mearns, 2009: Bayesian Modeling of Uncertainty in Ensembles of Climate Models. *J. Am. Stat. Assoc.,* **104(485)**, 97-116.

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., and C.E. Forest, 2007: Ensemble climate predictions using climate models and observational constraints. *Phil. Trans. R. Soc. A,* **365**, 2029-2052.

Stott, P.A., J.A. Kettleborough, and M.R. Allen, 2006: Uncertainty in continental-scale temperature predictions. *Geophys. Res. Lett.,* **33(2)**, L02708.

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(13)**, 3055-3069.

Suppiah, R., K.J. Hennessy, P.H. Whetton, K. McInnes, I. Macadam, J. Bathols, J. Ricketts, and C.M. Page, 2007: Australian climate change projections derived from simulations performed for the IPCC 4th Assessment Report. *Aust. Meteorol. Mag.,* **56(3)**, 131-152.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Phil. Trans. R. 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**, 1524-540.

Watterson, I.G., 2008: Calculation of probability density functions for temperature and precipitation change under global warming. *J. Geophys. Res.,* **113(D12)**, D12106.

Weigel, A.P., R. Knutti, M.A. Liniger, and C. Appenzeller, 2010: Risks of Model-Weighting in Multimodel Climate Projections. *J. Clim.*, in press.

Whetton, P., I. Macadam, J. Bathols, and J. O'Grady, 2007: Assessment of the use of current climate patterns to evalate regional enhanced greenhouse response patterns of climate models. *Geophys. Res. Lett.,* **34**, L14701.