Diss. ETH No. 19488

Statistical uncertainty analysis in an ensemble of global climate models

A dissertation submitted to the
ETH ZURICH

for the degree of
DOCTOR OF SCIENCES

presented by

David Masson

Dipl. Phys. ETH Zurich
born on 13 April 1979
citizen of Ecublens (VD)

accepted on the recommendation of

Prof. Dr. Reto Knutti, examiner
Dr. Benjamin Sanderson, co-examiner
Dr. Andreas Weigel, co-examiner

Zurich 2011

Contents

Abstract **i**

Résumé **iii**

1 Introduction **1**
- 1.1 Motivation 1
- 1.2 Background 2
  - 1.2.1 Simple energy-balance model 2
  - 1.2.2 Climate sensitivity 3
  - 1.2.3 Why is climate sensitivity so uncertain? 4
- 1.3 Outline of the thesis 6

2 Statistical methods **7**
- 2.1 Random forests 7
- 2.2 Empirical orthogonal functions 12
- 2.3 Cluster analysis 15
- 2.4 Kullback-Leibler divergence 18

3 Spatial scale analysis **21**
- 3.1 Introduction 23
- 3.2 Method 25
  - 3.2.1 Data 25
  - 3.2.2 Field smoothing 25
  - 3.2.3 Measures of uncertainty 26
  - 3.2.4 Optimal smoothing for present-day simulations 26
  - 3.2.5 Impact of resolution on model error 27
  - 3.2.6 Robustness of climate change signal 27
  - 3.2.7 Initial error preservation 27
- 3.3 Results 28
  - 3.3.1 Field smoothing and measures of uncertainty 28
  - 3.3.2 Optimal smoothing for present-day simulations 30
  - 3.3.3 Impact of the resolution on model error 33
  - 3.3.4 Robustness of climate change signal 33
  - 3.3.5 Initial error preservation 35
- 3.4 Discussion 35

3.4.1 Data 35
3.4.2 Measures of uncertainty 36
3.4.3 Optimal smoothing for present day simulations 37
3.4.4 Impact of resolution on model error 37
3.4.5 Robustness of climate change signal 38
3.4.6 Initial error preservation 38
3.5 Conclusion 39

4 Constraining climate sensitivity **41**
4.1 Introduction 43
4.2 Models and data 45
4.3 Constraining climate sensitivity from observations 46
4.3.1 CMIP3 ensemble 46
4.3.2 CPDN ensemble 51
4.4 Origin of the linear relation and tuning 52
4.5 Discussion and conclusion 55

5 Climate model genealogy **59**
5.1 Introduction 61
5.2 Method 62
5.3 Results and discussion 62
5.4 Conclusions 63
5.5 Supplementary material 68
5.6 Data 68
5.7 Methods 68
5.8 Robustness of results and additional discussion 70

6 Model weighting: a conceptual study **73**
6.1 Introduction 75
6.2 Role of the ensemble size using CMIP3 models 76
6.3 Role of model calibration using artificial models 77
6.4 Discussion and conclusion 80

7 Conclusions and outlook **83**
7.1 Conclusions 83
7.2 Outlook 85

Acknowledgments **87**

Curriculum Vitae **88**

References **91**

Abstract

This PhD thesis examines several ensembles of general circulation models from a statistical point-of-view. These models not only play a central role in detection and attribution studies on climate change, but also often serve as basis for projections and adaptation planning for policy makers. While scientific consensus points to human activities as the primary cause of the observed 20^{th} century warming, the details of future climate change remain uncertain. Apart from an agreement on the positive sign of future temperature change, all climate models disagree to a certain extent, depending on the variable, time horizon and spatial scale considered. While this disagreement is not a problem in itself, quantifying uncertainty from an ensemble of simulations whose performance and model dependence are not well identified remains challenging.

Model evaluation is often performed on a grid-point basis despite the fact that models are known to often be unreliable at such small spatial scales. A first study examines how different spatial and temporal scales affect the robustness of temperature and precipitation projections. One finding suggests that increasing the native resolution of a model does not substantially improve the simulation of present-day global mean annual temperature. Another finding is the preservation of small spatial scale model bias until the end of the century, when the deviations from the observations are considered independently from the present-day global bias. This justifies the common practice to focus on anomalies from a control simulation rather than absolute values.

The second study investigates why the predicted amount of global warming by different general circulation models is so difficult to constrain by the observations. The comparison between these state-of-the-art climate models with another ensemble made of a single model whose parameters have been perturbed suggests that parameter calibration plays an important role. As model calibration might have used all the information contained in the observation, then model climate sensitivity cannot be constrained a second time by the same observation. Moreover, if model parameters are tuned, this might give a false impression of validity for present-day while systematic biases could appear in the future.

The third study focuses on the relationships between general climate models from a statistical point of view. We find that models from the same institutions generally share similar behavior even after significant development. As a consequence, considering the climate models as independent is not justified and multi-model averages are probably biased towards certain model designs.

Finally, a fourth study explores the risks associated with weighting climate models according to their present-day performance. A conceptual experiment is constructed where two hypothetical IPCC reports are created at the same time. The consistency of both climate projections depends on the ensemble size and how strongly parameters are calibrated to match the observations.

Overall, this thesis has demonstrated the importance an appropriate spatial scale for model evaluation and the crucial role played by model inter-dependence and calibration for the uncertainty range sampled by future climate simulations.

Résumé

Cette thèse étudie divers ensembles de modèles climatiques d'un point de vue statistique. L'importance de tels modèles numériques n'est plus à démontrer : détection et attribution des changements climatiques ou études d'impact déstinées aux décideurs politiques sont des exemples fréquemment cités. Bien qu'un large consensus attribue aux activités humaines le réchauffement climatique observé durant le 20ème siècle, de larges incertitudes persistent sur l'ampleur et la répartitions géographique des futurs changements. Si les modèles s'accordent tous sur un réchauffement global, ils divergent en revanche plus ou moins dans les détails selon la variable considérée, l'éloignement dans le futur ou l'échelle spatiale à laquelle on les compare. Ce désaccord n'est pas problématique en soi ; le véritable enjeu consiste plutôt à correctement estimer les incertitudes reflétant l'état de nos connaissances. Le défi est de taille surtout lorsque des notions de performance ou d'inter-dépendence au sujet des modèles ne sont pas clairement définies.

L'évaluation des modèles est fréquemment effectuée à la plus petite échelle spatiale possible, c'est à dire celle de la cellule élémentaire du maillage global, ceci en dépit du fait que les simulations sont rarement fiables à une telle échelle. C'est ainsi que la première étude de cette thèse s'applique à définir une échelle spatiale et temporelle optimale à laquelle les modèles seraient consistants entre eux ou avec les observations. Il en ressort que l'accroissement de la résolution native d'un modèle n'apporte pas d'avantages pour la simulation de la température à l'échelle globale. Un autre résultat rend compte de la bonne préservation des erreurs à l'échelle locale jusqu'à la fin ce siècle, abstraction faite de l'erreur à l'échelle globale pour la période présente. Un tel résultat justifie une pratique répandue qui consiste à se focaliser sur les anomalies d'une simulation par rapport à sa version de contrôle plutôt que d'en analyser les valeurs absolues.

L'étude suivante consiste à comprendre pourquoi si peu de variables basées sur les observations sont à même de réduire l'incertitude sur la sensibilité climatique prédite par les modèles. Une comparaison est établie avec, d'un coté un ensemble de modèles à l'état de l'art, et de l'autre, un large ensemble de simulations générées par la perturbation des paramètres d'un seul modèle. Cette comparaison suggère que la calibration des modèles joue un rôle important : si toute les informations disponibles sont intégrées par les modèles au travers de la calibration, il n'est pas étonnant que les observations ne peuvent fournir davantage de contraintes quant à la réduction de l'incertitude touchant la sensibilité climatique. De surcroît, il se peut que les modèles compensent certains mécanismes afin de reproduire fidèlement les observations.

Un tel réglage pourrait produire une spécieuse apparence de validité et causer des déviations systématiques dans le futur.

Une troisième étude se focalise sur les similarités existantes entre les modèles d'un point de vue statistique. Nous démontrons que des modèles différents mais provenant des mêmes institutions produisent des simulations apparentées, ceci également au fil des développements successifs des modèles. Ainsi, le fait de considérer tous les modèles comme autant d'approches indépendantes ne reflète pas la réalité. En conséquence, l'incertitude des futures projections climatiques est probablement biaisée par certains concepts redondants.

Enfin, l'idée de donner à chaque modèle un poids reflétant son habileté à reproduire les observations est analysée dans une quatrième étude. Nous utilisons une expérience conceptuelle qui tend à démontrer que l'usage de tels coefficients est risqué s'il n'existe pas de relation entre la performance mesurée aujourd'hui et les projections climatiques futures. Le rôle du nombre de modèle dans un ensemble ainsi que celui de la calibration est également pris en compte.

Globalement, cette thèse démontre de quelle manière l'incertitude relative aux futurs changements climatiques est affectée par le choix de l'échelle spatiale, par la calibration des paramètres régissant les modèles et par certaines interdépendances conceptuelles.

Chapter 1

Introduction

Certainty is the anomalous condition for humanity, not uncertainty.
(Mike Hulme)

1.1 Motivation

In 1996, the second assessment report of the Intergovernmental Panel on Climate Change (IPCC) identified “a discernible human influence on global climate”. Eleven years later, the IPCC Fourth Assessment Report (AR4) reinforced this claim and stated that “most of the observed increase in global averaged temperatures (...) is very likely due to the observed increase in anthropogenic greenhouse gas concentrations”. The progress made in detecting and attributing climate change is a remarkable scientific achievement that required the coordination of very different fields.

Yet, even if a consensus exists that the Earth’s climate is changing as a response to anthropogenic emissions from burning fossil fuels, the amplitude and the pattern of climate change remain uncertain. Because societal, ecological and economical stakes would be very different if 1 or 6 °C of global warming were to be expected, a single value representing the best estimation of the projected climate change is almost useless if not accompanied by reliable uncertainty estimates. The future is uncertain because many unknowns of different nature are combined: how much more greenhouse gas is expected, how predictable is the natural climate variability, what do we really know about the climate system processes? (See Cox and Stephenson, 2007; Hawkins and Sutton, 2009, for more details).

The development of computer technology has allowed the exploration of several hypotheses by using elaborated General Circulation Models (GCMs). As these numerical models digest all relevant existing knowledge on the climate system, they play a central role in the projection of future climate change. To give an idea, the typical size of a GCM project implies about one million lines of source code across several hundreds of files. At the present day, only about 20 general circulation models from various institution worldwide are able to simulate the relevant climate processes over several decades, at the detailed spatial and time resolution required to provide reasonably reliable future projections (see Edwards, 2010, for a history of climate modeling). Although these models all disagree to a certain extent, a quantification of uncertainty based on their divergence remains subject of many debates. The complexity of this question is reflected by the amount of published studies on this topic (e.g. Stainforth et al., 2007; Tebaldi and Knutti, 2007; Knutti, 2008a; Knutti et al., 2008b, 2010b). Still, no consensus exists on the evaluation of climate models, neither on which statistical method to apply nor on the nature of the available sample of climate simulations. This thesis aims to get a better understanding of the ensemble of models used in the last IPCC AR4 report. More specifically, we are interested in answering the following questions:

- What are the spatial and temporal scales at which models can provide robust information?
- Why are there so few observational constraints on climate sensitivity?
- Is the history of climate models relevant and how much diversity is really contained is an ensemble of GCMs?
- Should GCMs be weighted according to their present-day performance?

1.2 Background

The representation of climate consists of several layers of complexity where mechanisms of different kinds interact with each others. This section provides a glimpse into the general background on which this thesis is built. We introduce basic understanding of the Earth system via a simple energy balance model and illustrate how radiative forcing and feedbacks inevitably affect uncertainty of future climate projections.

1.2.1 Simple energy-balance model

The Earth's global mean temperature is a key variable driving the climate. As temperature is directly related to energy, some fundamental notions of climate can be illustrated using a zero dimensional energy-balance model. The Earth receives short-wave energy from the sun and emits long-wave energy back to space. If the climate system is simplified to a thin atmospheric layer surrounding the Globe, the first law of thermodynamics applied at the top of the atmosphere is:

$$4\pi R^2 h\rho \cdot C \cdot \frac{dT}{dt} = \pi R^2(1-\alpha)S - 4\pi R^2 \epsilon\sigma T^4 \tag{1.1}$$

where the following terms are used:

T	average temperature of the atmosphere layer
$R = 6371$ km	Earth's radius
$h = 8.3$ km	height of the troposphere
$\rho = 1.2$ kg m^{-3}	air density
$C = 1000$ J kg^{-1} K^{-1}	air heat capacity
$\alpha = 0.3$	planetary albedo
$S = 1367$ W m^{-2}	Solar constant
$\epsilon = 0.6$	emissivity
$\sigma = 5.67 \cdot 10^{-8}$ W m^{-2} K^{-4}	Stefan-Boltzmann constant

The right-hand-side term ($\pi R^2 \cdot S$) represents the amount of short-wave energy received at the Earth surface. About 30% of the incoming solar radiation is directly reflected by the planetary albedo (α) formed mostly by reflecting components such as clouds and snow. The remaining portion is absorbed by the climate system and warms the oceans, the ground surface and the atmosphere. The energy is re-emitted back to space in the infra-red part of the spectrum according to the Stefan-Boltzmann law ($\sim \epsilon\sigma T^4$). As long as the infrared energy emitted by the top of the atmosphere does not exactly compensate the incoming solar energy, the atmosphere temperature varies to restore equilibrium.

An external *radiative forcing* is the change in net (downward minus upward) radiation (in W m^{-2}) at the tropopause and after stratospheric adjustment (IPCC, 2007). Radiative forcing can be caused by a change in solar activity, the natural cycle of the seasons, but also by the change in aerosol concentrations e.g. due to volcanic eruption or greenhouse-gas emission due to human activities. The simple equation (1.1) allows the estimation of the climate response after a radiative forcing has been applied and leads to a solution of the form $T(t) = \bar{T} + \mathrm{const} \cdot e^{-t/\tau}$, with $\bar{T}$ being the global equilibrium temperature before a forcing is applied, and τ is the typical time scale needed to restore equilibrium. Without the natural greenhouse effect, $\bar{T} = -18.3\,^{\circ}\mathrm{C}$ and $\tau \approx 34.6$ days after the external forcing has ceased. In reality, $\bar{T} = +14\,^{\circ}\mathrm{C}$ and τ depends on the type of forcing applied to the climate system and ranges from months to centuries due to different processes and heat capacities in the climate system (Stouffer, 2004; Knutti et al., 2008a). In addition, the vast majority of the heat is taken up by the mixed layer in the oceans, not by the air itself. In order to get a reliable picture of the climate system and estimate of climate change, the entire atmosphere, hydrosphere, cryosphere, lithosphere and biosphere need to be represented in comprehensive Atmospheric and Oceanic General Circulation Models (AOGCMs).

1.2.2 Climate sensitivity

In the context of anthropogenic climate change, equilibrium climate sensitivity (S) is an important number and is defined as the equilibrium global average temperature change for a doubling of the atmospheric $\mathrm{CO_2}$ concentration. Reformulating equation (1.1) gives

$$\Delta Q = \Delta F - \lambda \Delta T \tag{1.2}$$

where ΔF is a radiative forcing (in W m^{-2}) leading to a surface warming ΔT and an increased heat flux ΔQ, mainly taken by the oceans (see Hansen et al., 1981; Levitus et al., 2000). The constant λ is known as the *climate feedback parameter*. At thermal equilibrium, ΔQ is zero and $\lambda = \Delta F / \Delta T$. The inverse value $1/\lambda = \Delta T / \Delta F$ is known as *climate sensitivity parameter*. Because many aspects of climate change scale linearly with S, equilibrium climate sensitivity is a central quantity in climate modeling. Unfortunately, some physical processes are poorly understood and studies that have tried to estimate S have also found large uncertainties (see Fig. 1.1). Although all results are consistent on the positive sign of temperature change, the distribution is generally positively skewed and do not exclude large value. The state-of-the-art GCMs run for the IPCC AR4 report predict a likely climate sensitivity range of $2 - 4.5$ K (IPCC, 2007, Box 10.2).

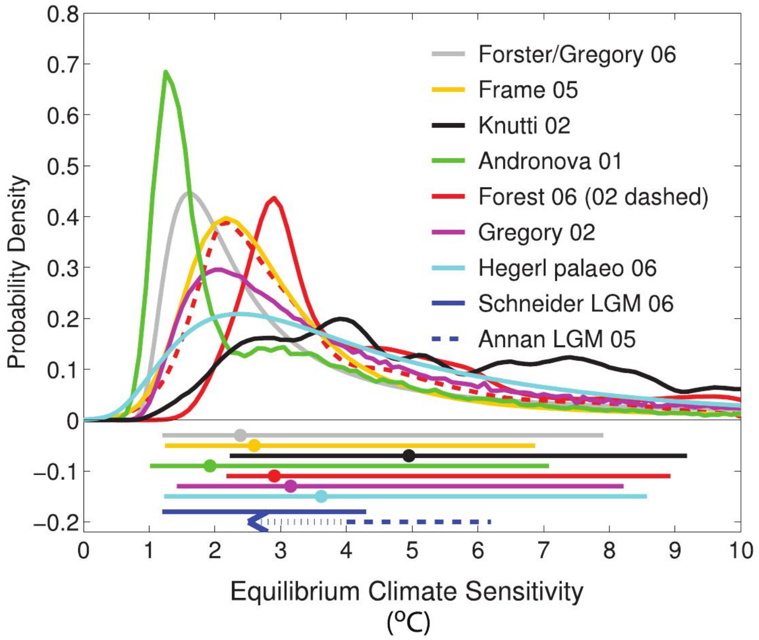


Figure 1.1: *Probability density functions for the equilibrium climate sensitivity from different studies based on observational constraints. Also shown are the 5 to 95% approximate ranges with the points denoting the median values. Reproduced from IPCC (2007), Fig. 9.20, where more details are given.*

1.2.3 Why is climate sensitivity so uncertain?

The primary source of model disagreement concerning climate sensitivity is due to limited knowledge about feedback processes, and particularly to the parameterization of cloudiness and cloud-radiation interactions (IPCC, 2007). The concept of *feedback* can be understood as processes amplifying or damping the initial blackbody response of $\Delta T_0 = 1.2$ K after doubling CO_2 concentration. The positive skewness of the climate sensitivity distribution is inherent to the climate system if we measure the total feedback parameter f (Roe and Baker, 2007) and can be phrased by

$$S = \Delta T_0/(1-f) \tag{1.3}$$

Because f and S are non-linearly related, a symmetric distribution for f results in a skewed distribution in S as shown in Fig. 1.2.

The climate sensitivity is therefore conditioned by the representation of feedback processes in climate models. A more accurate classification makes a distinction between structural and parametrical uncertainty. *Strutural* divergences among GCMs are the result of different choices made for the numerical schemes, resolution, representation of the atmosphere, vegetation, but also of different computer hard- and softwares (Stainforth et al., 2007; Knight et al., 2007; Knutti, 2008a). A systematic sampling of the structural uncertainty is conceptually difficult to achieve and the complexity of the GCMs makes them expensive to run several times even for large institutions. Therefore, the structural uncertainty of the climate sensitivity is sampled only by about twenty structurally different GCMs from various institutions that are gathered in the CMIP3 project (Meehl et al., 2007a). The second kind of uncertainty is related to the parametrization of physical and chemical processes and exists because of lack of knowledge or because some direct dynamical calculations are computationally too expensive. As parameter

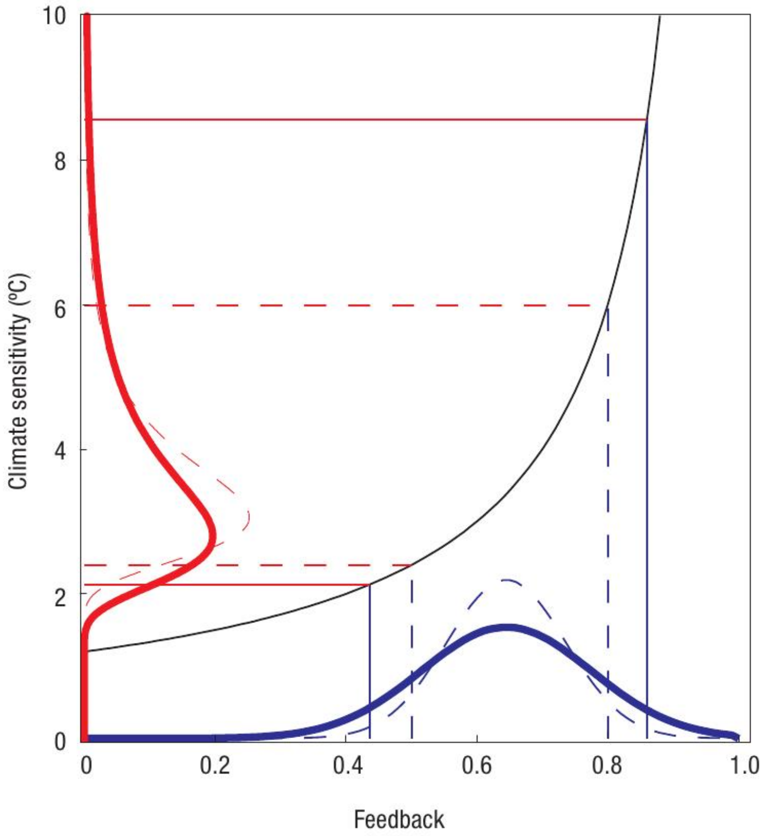


Figure 1.2: *Relation between amplifying feedbacks f and climate sensitivity S (black line). A symmetric distribution of the total feedback parameter f (blue line) results in a skewed climate sensitivity distribution (red line) with larger uncertainty for high climate sensitivity value. The dashed lines shows that reducing the feedback uncertainty by* 30% *does not affect the skewness of climate sensitivity. Figure from Knutti and Hegerl (2008).*

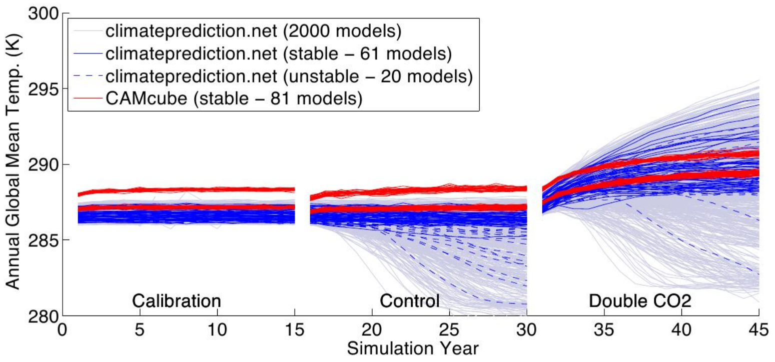


Figure 1.3: *Global annual mean temperature simulated by different perturbed physics ensemble. The three different sequences correspond to the calibration, the control and the doubling of* CO_2 *phases. Figure from Sanderson (2010).*

values are not alway well constrained by observations, the *parametric uncertainty* remains a large source of model disagreement. In contrast to the structural uncertainty, the parametrical uncertainty can be systematically explored by *perturbed physics ensemble* (PPE), i.e. by running several parameter combinations of the same GCM (Kennedy and O'Hagan, 2001; Murphy et al., 2004). Fig. 1.3 shows some runs from the climateprediction.net and CAMcube perturbed ensembles (Stainforth et al., 2005; Sanderson, 2010) where different perturbations result in different temperature projections.

1.3 Outline of the thesis

The next chapter is an introduction to specific statistical methods used in this thesis. Chapters 3 to 5 form the main body of this dissertation and consist of studies submitted or in preparation for international peer-reviewed journals:

- Chapter 3, **"Spatial scale dependence of climate model performance in the CMIP3 ensemble"** *(published in Journal of Climate)*. Which are the spatial and temporal scales where GCMs are able to provide reliable results? While small spatial scales are important to determine specific climate impacts, inter-model disagreements at this scale suggest that some form of spatial aggregation might improve model consistency.
- Chapter 4, **"Constraining climate sensitivity from interannual variability: an illustration of tuning in climate model ensembles"** *(in revision in Journal of Climate)*. Why is climate sensitivity so difficult to constrain from the observations? The role of model calibration is highlighted.
- Chapter 5, **"Climate model genealogy"** *(published in Geophysical Research Letters)*. While many studies consider the 24 GCMs gathered for the IPCC AR4 as independent, some evidence suggests that the number of truly independent models is smaller. This study focuses on climate model relationships from the statistical point-of-view.
- Chapter 6, **"Model weighting, ensemble size and tuning: a conceptual study"***(not published)*. This conceptual experiment explores the benefit and risk of weighting models by present-day performance.

Chapter 2

Statistical methods

This thesis deals with large datasets, sometimes in range of several gigabytes. When a vast amount of data needs interpretations, machine learning algorithms are often used to reveal patterns and structures. The learning problems can be categorized into two classes: *supervised* and *unsupervised*. In supervised learning, the goal is to predict an output from a configuration of input parameters. Decision trees, random forests and neural networks belong to this category and are presented in this chapter. Unsupervised learning includes empirical orthogonal functions and clustering techniques. These algorithms seek for organized patterns in the input ensemble itself and no output is analyzed. The last section of this chapter describes the Kullback-Leibler divergence for multivariate normal distributions and is also a short glimpse on information theory. With increased computational power and shared database worldwide, these techniques have become popular in various fields such as biostatistics, econometrics and of course climate science.

2.1 Random forests

Random forests belong to pattern recognition algorithms whose task is the following: given a set of N input variables $\mathbf{x} = (x_1, x_2, \cdots, x_N)^t$, predict a real-valued output y via the model $y = f(\mathbf{x}) + \epsilon$, where f is an underlying but unknown function describing the model and ϵ some noise contaminating the data. A fit is particularly interesting when the relation f is complex and non-linear. A pattern recognition algorithm constructs an estimate $\hat{f}$ that minimizes the errors $\left(y_i - \hat{f}(\mathbf{x}_i)\right)$ between the original and the generated outputs. In the learning phase, only part of the original data is used and is known as the *training set*. The performance of the fit $\hat{f}$ is estimated on the set of remaining data by predicting output values $\hat{y}$ that are compared with the true values y not used during the training phase. Typical supervised learning techniques are neural networks, regression trees and random forests. We present random forest and regression tree algorithms and compare their performance with those of a neural network on a specific example in climate modeling.

Regression tree analysis is conceptually simple yet powerful. The idea behind is to fit a multi-dimensional step function to a continuous function f. Let us first consider the result of a regression tree analysis $\hat{f}(x_1, x_2) = \hat{y}$ which is represented in Fig. 2.1a. The input space is divided into the five response partitions shown in Fig. 2.1b and $\hat{y}$ takes fives possible response

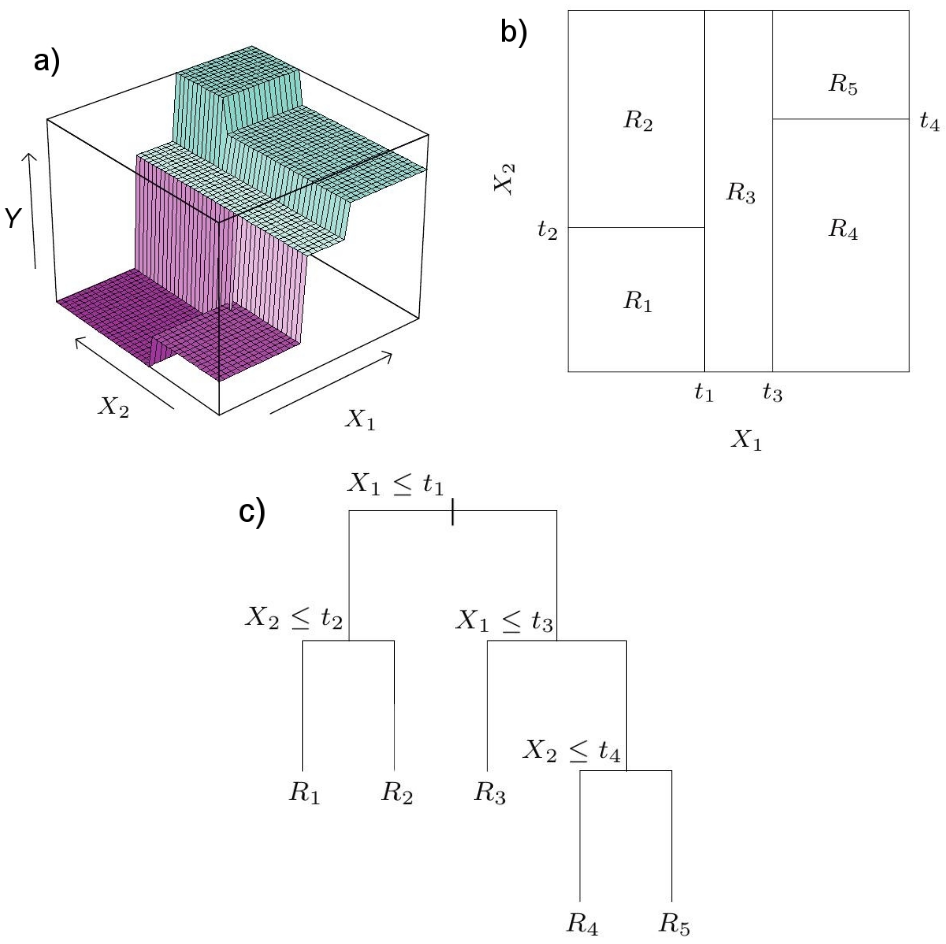


Figure 2.1: *Illustration of a regression tree. (a) Perspective plot of the prediction surface as a function of the input parameters X_1 and X_2. (b) The input space spanned by X_1 and X_2 is parted into five response regions $R_1, \ldots, R_5$. (c) Decision tree based on the partition $R_1, \ldots, R_5$. Figures from Hastie et al. (2001).*

values $R_1, \ldots, R_5$ and the corresponding decision-tree shown in Fig. 2.1c. If a continuous function f is fitted, the question is how small should we divide the partition of the input space? A too detailed tree will start to fit the noise in the training set leading to inappropriate estimate $\hat{f}$. This phenomenon is known as *overfitting*. Many techniques exist to avoid overfitting and are referred to *pruning*. More details can be found in Hastie et al. (2001).

We illustrate this technique with a perturbed physics ensemble. The climateprediction.net project perturbs the parameter values of the HadCM3L climate model and generates thousands of simulations, each representing a possible evolution of the Earth's climate (Stainforth et al., 2005). The combination of the input parameters $(x_1, \ldots, x_N)$ affects climate sensitivity (y) of the simulated climate system considered as the output response. We are interested whether a group of leading parameters exists. Another possible application is to save time by avoiding the explicit simulation of the parameters by a climate model $f(\mathbf{x})$ and to rely on the fit $\hat{f}(\mathbf{x})$ as model surrogate instead. In this case, $\hat{f}(\mathbf{x})$ is often called an *emulator*.

The result of a decision tree analysis using 60% of the data for the training phase is shown in Fig. 2.2. The input parameters affect the extent of cloud cover, precipitation, the convection scheme, etc. For a more complete description of the parameters, see Sanderson et al. (2008). Focusing on the left branch of the tree as an illustration, when the adjusted cloud fraction

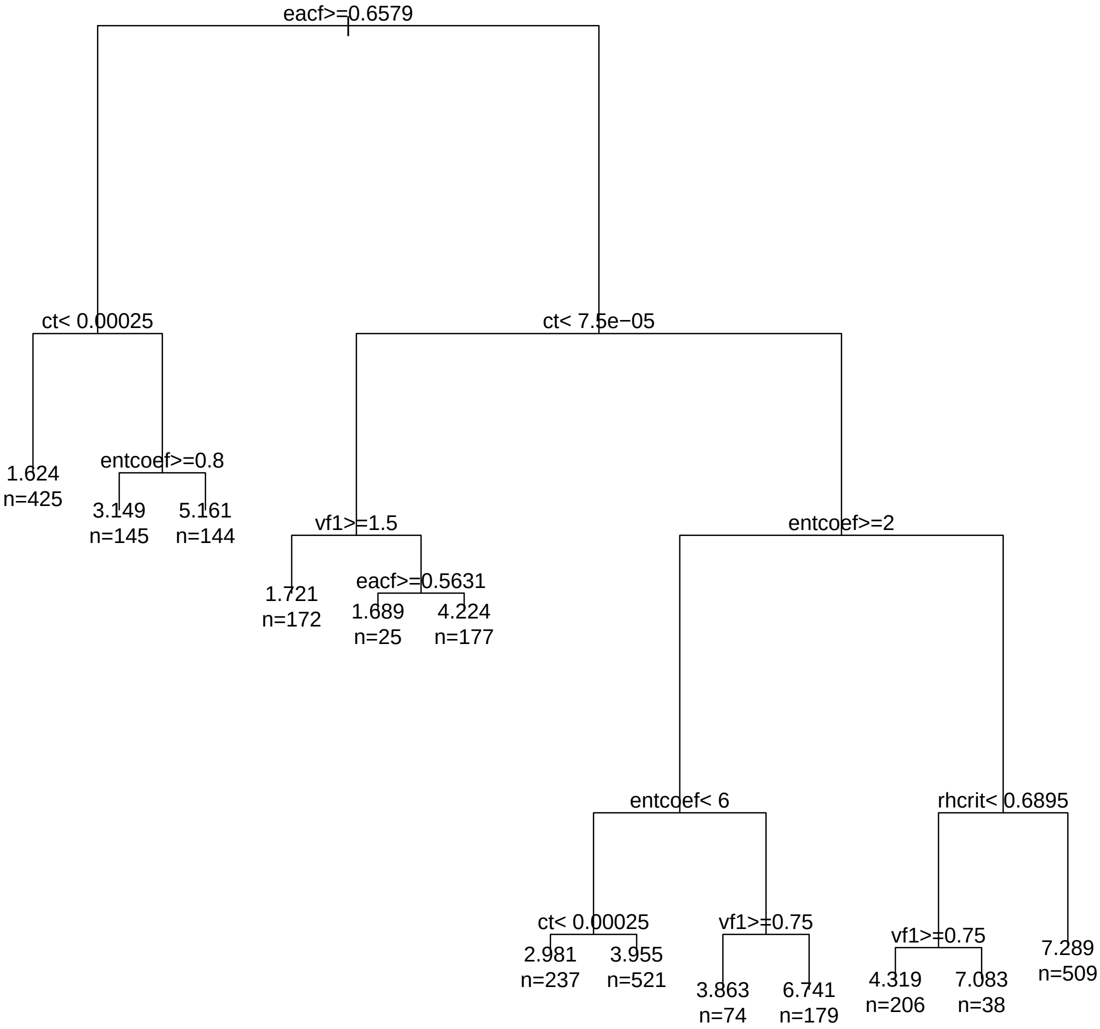


Figure 2.2: *Regression tree applied to the climateprediction.net simulations. The input is the parameters combination and the output is the simulated climate sensitivity. See text for explanations. More details about the parameters can be found in Sanderson et al. (2008)*

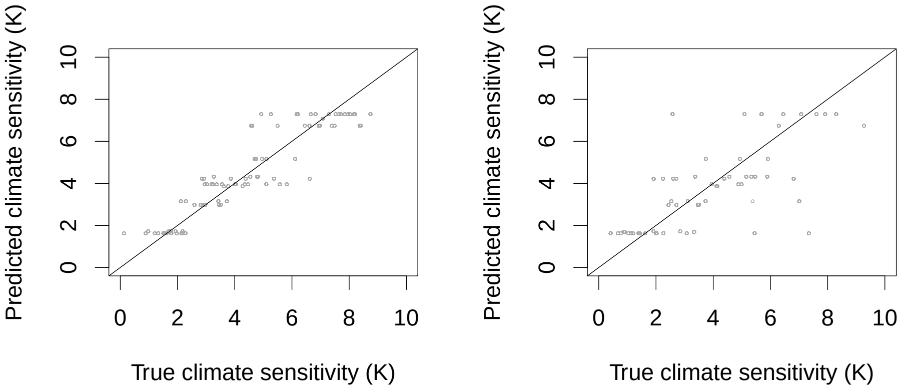


Figure 2.3: *Predicting climate sensitivity using a regression tree algorithm. The performance is measured in terms of linear correlation between the true climate sensitivity value versus the estimated value. Left panel: training set representing* 60% *of the data. The correlation is equal to* 0.95*. Right panel: validation set with the remaining* 40% *of the data. The correlation is equal to* 0.75.

(eacf) is larger than 0.66 and the accretion constant (ct) smaller than $2.5 \cdot 10^{-4}$, the tree-based prediction of climate sensitivity is 1.62 K. The number $n = 425$ refers to the amount of models in the training set that belong to this category. The efficiency of the fit can be measured by the linear correlation between the predicted versus the true values in the validation set and is equal to 0.75 in this case. Fig. 2.3 shows the correlation in the training and in the validation sets. The correlation in the training set is higher and reaches 0.95 but a closer look shows that there is still large uncertainties even during the training. (Note that other skill measures such as root mean square errors can be used instead of the linear correlation. While systematic biases are ignored in a linear correlation, the potential skill can be quantified.)

We continue with the random forest algorithm which is built on regression trees. The general idea behind is to combine a large amount of different decision trees. A bootstrapping of the predictor variables is done at every node to evaluate the performance of different tree configurations. All trees get weighted and are finally recombined into the fit $\hat{f}$. The same experiment done before on the climateprediction.net ensemble is repeated using the random forest algorithm. The performance in the training set and in the validation set in shown in Fig. 2.4.

The learning is very efficient and is able to predict climate sensitivity almost perfectly. This performance is reflected in validation set, though with more uncertainties. Nevertheless, the correlation is 0.85 and clearly outperforms the fit done with a single regression tree. The bootstrapping necessary to generate the forest is also useful to assess the variable importance in relation to the output variable. In this example, Fig. 2.5.a reveals that the oceanic parameters (four last parameters to the right) are almost irrelevant to climate sensitivity. This result is expected because the ocean heat uptake in equilibrium is zero.

As a final comparison, a feed-forward single-hidden-layer neural network with 6 neurons is used to fit climate sensitivity (more explanations on neural networks can be found in Hastie et al., 2001). The performance is shown in Fig. 2.6. The correlation is 0.97 in the training set and 0.78 in the validation set. Depending on the nature of the data analyzed, some algorithms

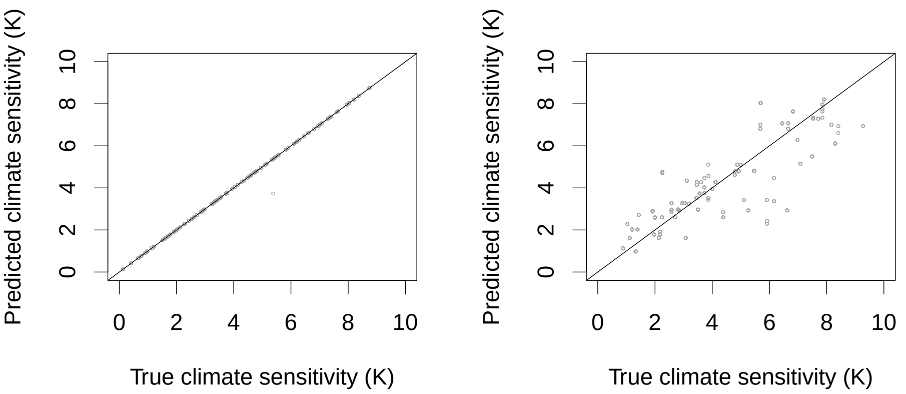


Figure 2.4: *Predicting climate sensitivity using a random forest algorithm. Left panel: training set representing* 60% *of the data. The correlation is equal to* 0.99. *Right panel: validation set with the remaining* 40% *of the data. The correlation is equal to* 0.85.

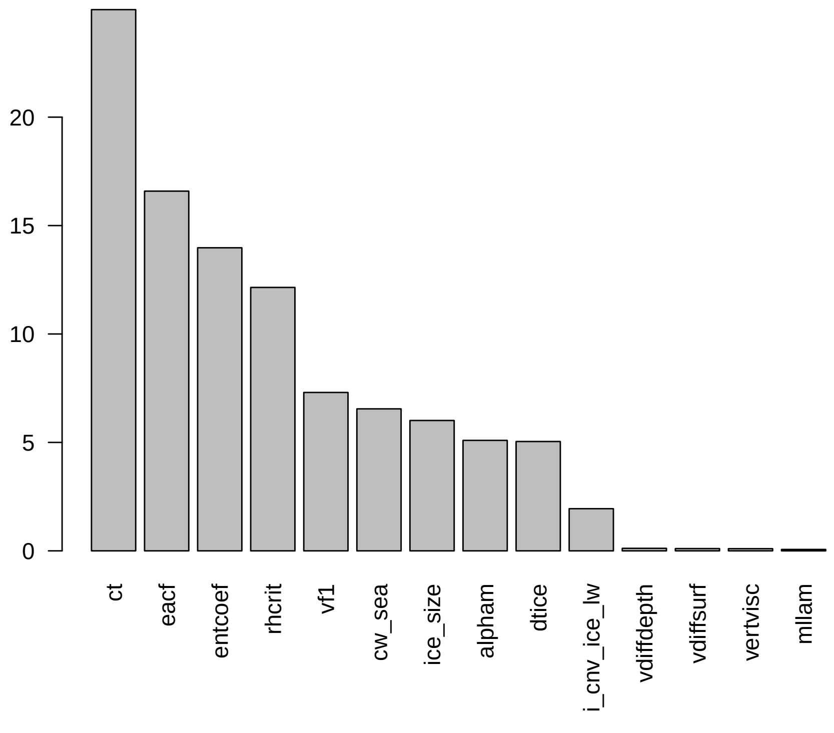


Figure 2.5: *Variable importance (in* $\%$*) of the perturbed parameters in relation to climate sensitivity. This estimation is done using random forest.*

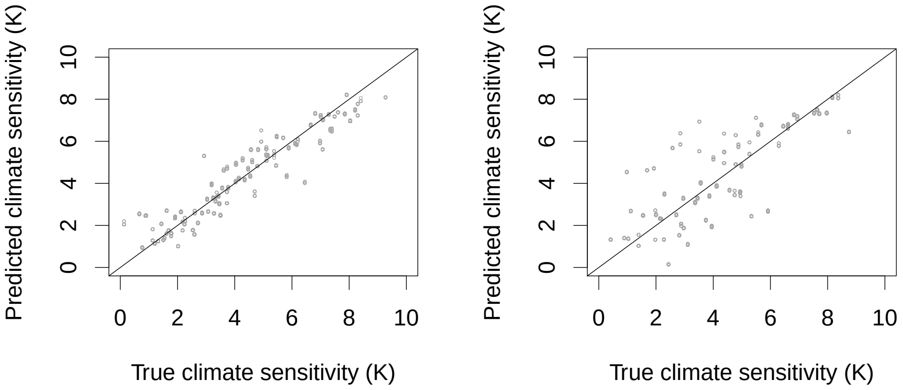


Figure 2.6: *Predicting climate sensitivity using a feed-forward single-hidden-layer neural network with 6 neurons. Left panel: training set representing* 60% *of the data. The correlation is equal to* 0.94*. Right panel: validation set with the remaining* 40% *of the data. The correlation is equal to* 0.78*.*

can be more or less efficient. The climateprediction.net project typically samples only two or three possible values per parameter. The example done above illustrates that while a step-function-based algorithm as random-forest is more adapted to fit a discrete input space, some other algorithms based on smoother functions as neural network are more likely to generate errors. An application of random forests is given in chapter 4.

2.2 Empirical orthogonal functions

A typical problem in climate science is the large amount of data. Large phase spaces often lead to computationally expensive calculations or singular matrices when the information is redundant. A solution is to decompose the data according to typical patterns that explain most of the variance, yet substantially reduce the dimensionality. In contrast to the Fourier decomposition, the basis vectors are patterns constructed directly from the empirical data and not from harmonical functions. This section starts with a general mathematical formulation of Empirical Orthogonal Functions (EOFs). We then compare the EOFs technique with the Spherical Harmonic Functions (SHFs) decomposition and show why EOFs are often more appropriate to decompose climatological fields.

Let $\mathbf{X}(t)$ be an N-dimensional random vector, for example the global monthly temperatures field. The main patterns of variability are computed from the $N \times N$ covariance matrix S,

$$S_{ij} = \frac{1}{T} \sum_{t=1}^{T} (X_i(t) - \bar{X}_i) \cdot (X_j(t) - \bar{X}_j)$$

using the singular value decomposition:

$$\mathbf{S} = \mathbf{E} \cdot \mathbf{\Lambda} \cdot \mathbf{E}^T$$

where Λ is a diagonal matrix. The $N \times N$ matrix $\mathbf{E}$ contains the orthonormal basis $\mathbf{E}_k$ ($k = 1, \ldots, N$), column-wise arranged, and represents the typical patterns of variability. The original

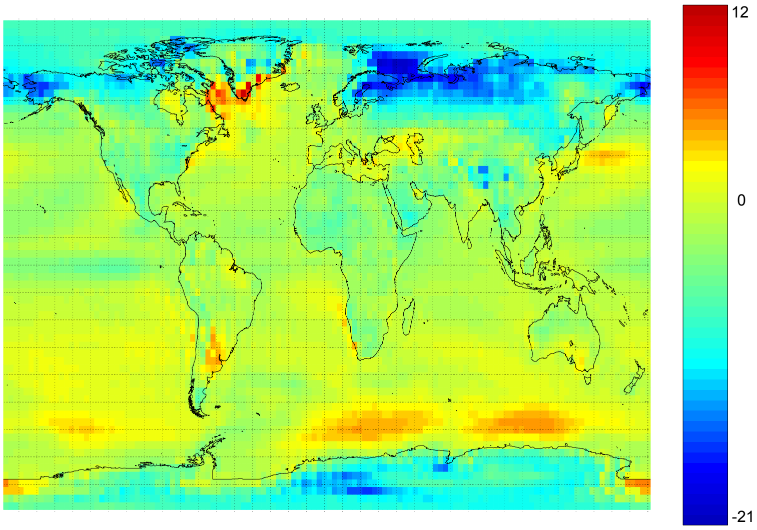


Figure 2.7: *Average temperature bias field (in K) between an observational dataset and a certain GCM for boreal winter during the period 1980-1999.*

field can be projected on the basis vectors, called Empirical Orthogonal Functions (EOFs), and are reconstructed according to:

$$\mathbf{X}(t) = \sum_{k=1}^{L \le N} U_k(t) \cdot \mathbf{E}_k \tag{2.1}$$

$\mathbf{U}(t)$ is the representation of the data in the new coordinate system and the components are by construction mutually uncorrelated. The reconstruction in equation (2.1) is perfect if $L = N$ and approximate otherwise. The quality of the compression is measured by R^2:

$$R^2(L) = \sum_{k=1}^{L} \operatorname{cov}(U_k(t)) \Bigg/ \sum_{i=1}^{N} \operatorname{cov}(X_i(t))$$

which is the cumulated fraction of variability explained by the typical patterns $\mathbf{E}_{k=1,\ldots,L}$.

EOFs techniques are often used in data mining to find typical patterns of variability (e.g. the Northern Atlantic Oscillation), to filter redundancies and uninteresting noise from a dataset, for data reconstruction, and to reduce the dimensionality. The last point is illustrated with an example where EOFs are particularly efficient. Let us consider the average temperature bias field between an observational dataset and a certain GCM for boreal winter during the period 1980-1999 as shown in Fig. 2.7. The size of this field is equal to the number of grid-points N (8192 here). We would like to reduce this size while preserving the maximum quantity of information. The average boreal winter temperature over the same period is simulated by an ensemble of $M = 24$ different GCMs. The typical modes of variability from these M temperature patterns are calculated using singular value decomposition and form a new basis made of N EOFs. More precisely, the singular value decomposition is performed on the $N \times N$ covariance matrix $\mathbf{S}$ representing the covariance between all possible pairs of grid-points using M temperature patterns $\mathbf{X}(m)$, $m = 1, \ldots, 24$. The three leading EOFs are shown in Fig. 2.8 and explain 67% of the total variance. The first EOF (upper figure) reflects the typical model

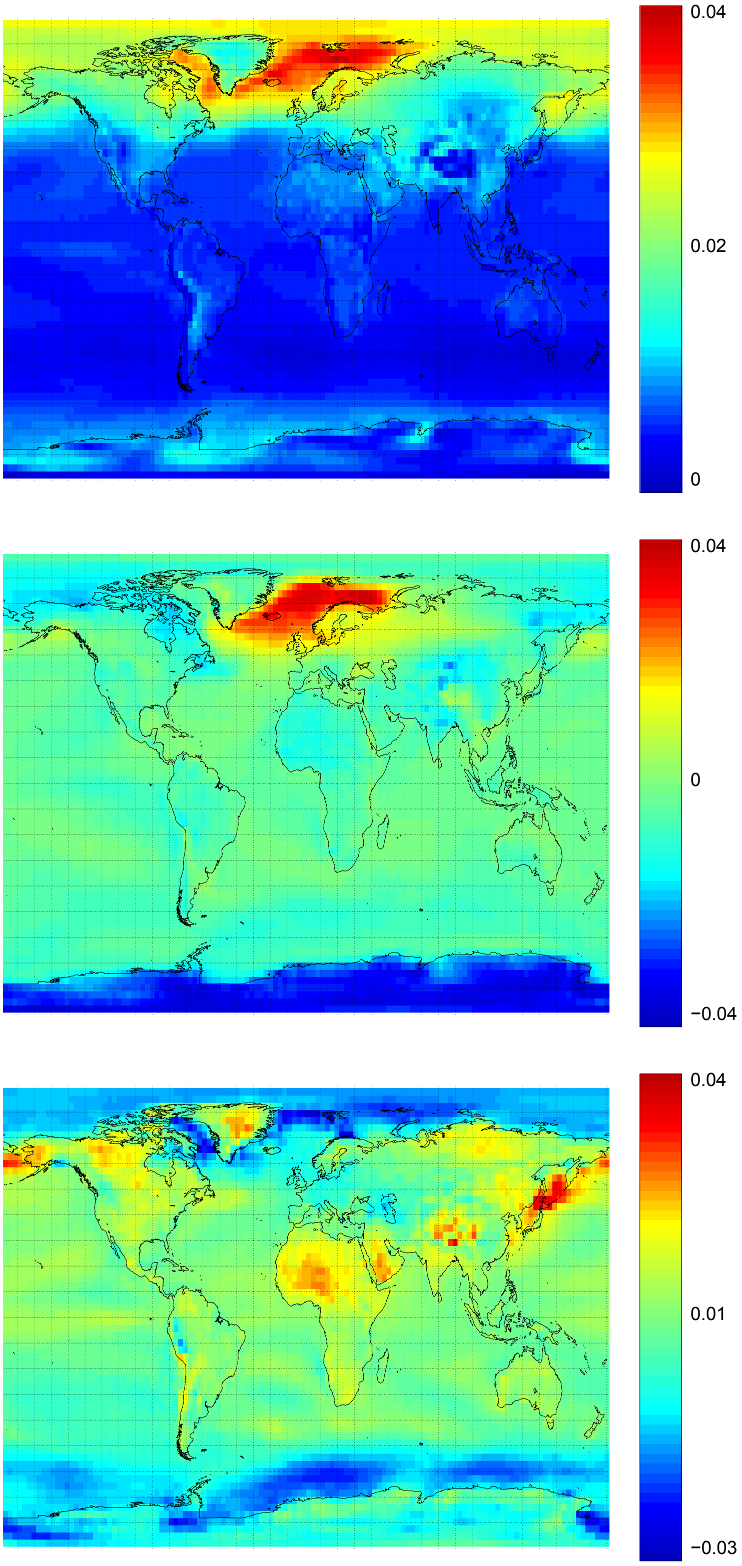


Figure 2.8: *The three first EOFs describing the main modes of variability across an ensemble of 24 different GCM simulations of boreal winter temperature (see text for more details). The EOFs have no unit. These three patterns explain* 67% *of the total variance.*

errors in the Arctic region. Note that the EOFs have no unit. The next EOF (figure in the middle) is a dipole between the Norwegian Sea and the Antarctic continent. Finally, the third EOF (lower figure) is related to sea-ice variations. While a total of $N = 8192$ EOFs exists, only a small number carries a physical interpretation.

For most of the applications, EOFs are used to find the main modes of variability and to relate them to physical processes (e.g. El Niño-Southern Oscillation or the North Atlantic Oscillation). In addition to that, this example seeks for an appropriate basis in order to reduce the number of coordinates needed to express the bias pattern shown in Fig. 2.7. Another possible basis is provided by Spherical Harmonic Functions (see e.g., Jackson, 1998) that can decompose any field on the sphere. Fig. 2.9 shows three possible basis functions from an infinite set of SHFs. These functions can explain 11% of the total variance in the same ensemble of 24 simulated boreal winter patterns. The final step is to compare the ability of the EOFs versus the SHFs to reconstruct the bias patterns shown in Fig. 2.7. The performance depends on two aspects: first, the quality of the reconstructed pattern in the new basis, measured in terms of linear correlation between the original and the reconstructed pattern. The second aspect is the reduction of the spatial dimension from N to a lower number of coordinates. If 8192 EOFs are used, the reconstruction is perfect but the size of the data has not been reduced. If 23 EOFs are used, the correlation between the original and the reconstructed field is 0.96, indicating that a small number of EOFs are enough to reconstruct most of the original information (see middle figure in Fig 2.10). If SHFs are used instead, even a large amount of 256 basis functions are not enough to accurately reconstruct the original field and show a correlation of only 0.63 with the original field (see bottom figure in Fig. 2.10). This example illustrates that a bias pattern of size $N = 8192$ can be reduced to 23 components without significant loss of information. Such techniques can be useful for computational purposes. As an example, certain applications require the inversion of the covariance matrix S. However, the inversion is impossible when $N \geq M$. The solution is to express the covariance matrix S in a new basis in order to reduce the spatial dimensionality. This technique is used in an application given in chapter 5.

2.3 Cluster analysis

Cluster analysis belongs to the class of unsupervised learning technique. The goal is to structure a set of data according to the degree of similarity between individual objects. The algorithm makes a partition of the data in K subgroups called *clusters*. The clustering algorithm groups the data such that the pairwise dissimilarities between those data assigned to the same cluster tend to be smaller than those in different clusters. Several clustering methods exist and a delicate part is generally to determine the optimal number K of clusters. Popular techniques are those algorithms that partitions the data with a specified number K of clusters and avoid (in a first step) the difficulty of finding the optimal number K. More important than the algorithm, the choice of the dissimilarity measure is decisive for the clustering outcome. It does not need to be a distance in the sense of fulfilling the triangle inequality (i.e. $D_{ij} \leq D_{ik} + D_{kj}$) but can be any dissimilarity symmetric measure. In the absence of pre-existing dissimilarity matrix, the squared Euclidian distance is by far the most common choice, but other options such as the Manhattan distance or the linear correlation are possible. In those case, the K-

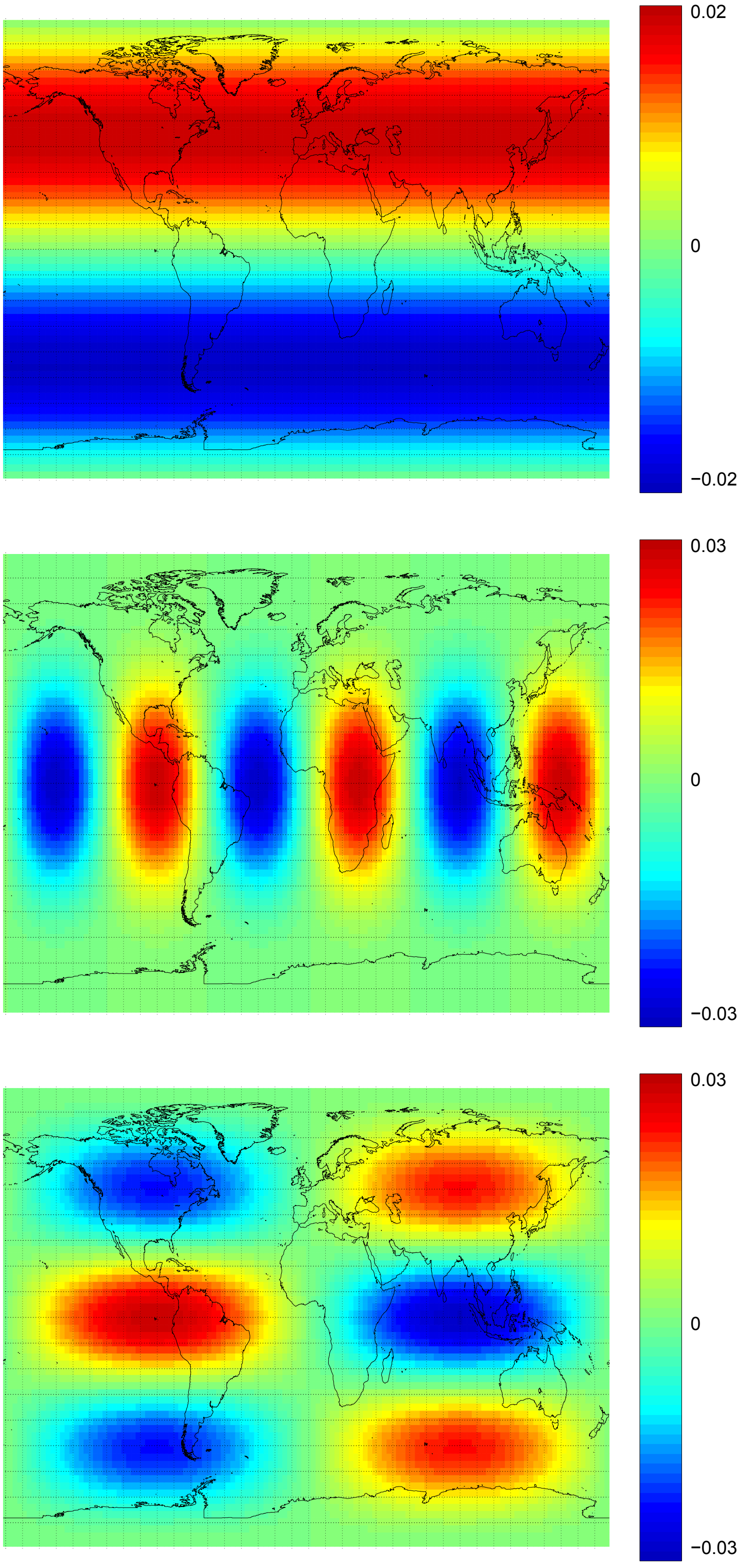


Figure 2.9: *Example of Spherical Harmonic Functions. These functions have no unit. These three patterns are able to explain 11% of the total variance in an ensemble of 24 different GCM simulations of boreal winter temperature (see text for more details).*

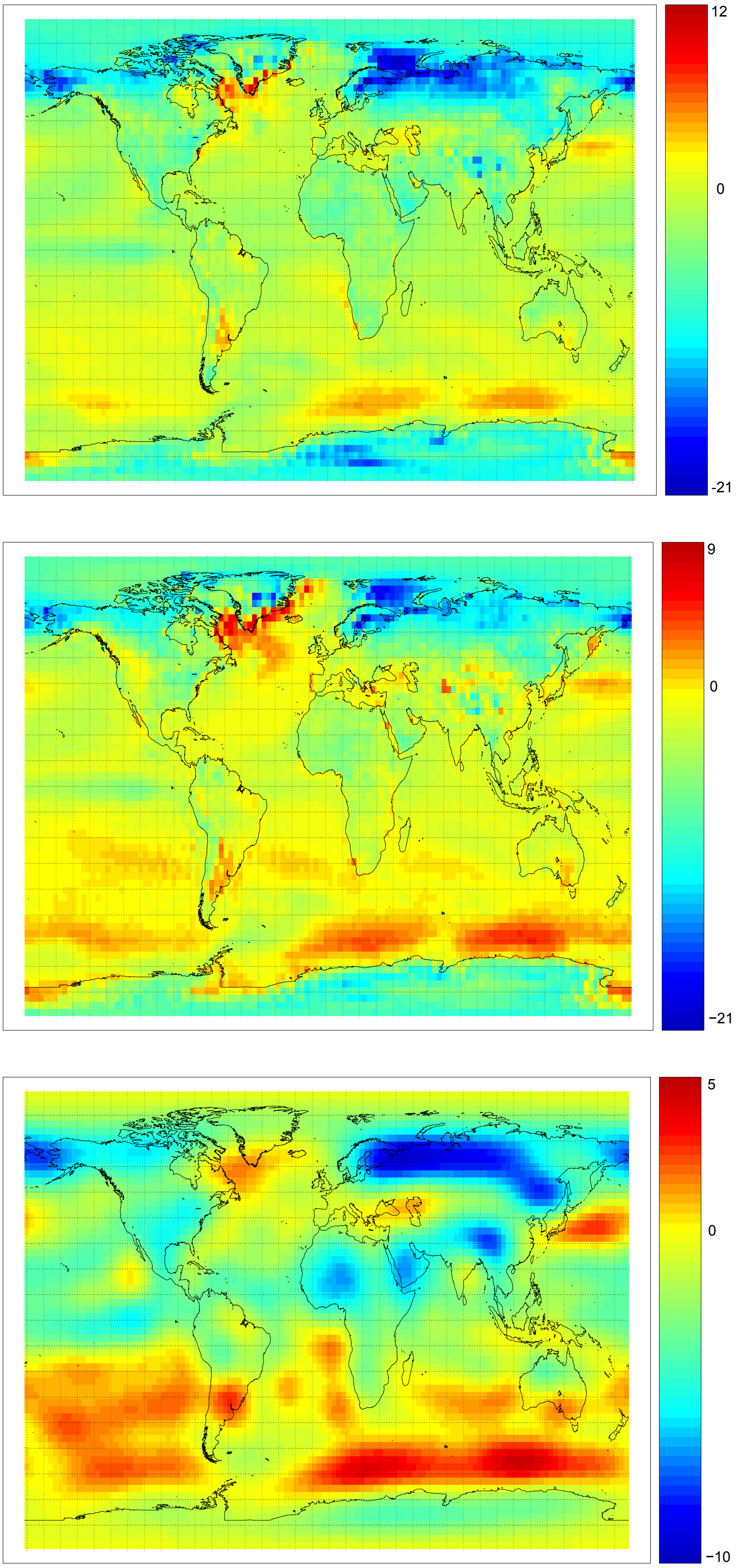


Figure 2.10: *Top: original bias field (in K) for boreal winter temperature during the period 1980-1999. Middle: Reconstructed field using 23 EOFs. The correlation between the original bias field and the EOFs reconstruction is* 0.96*. Bottom: Reconstructed field using 256 spheric harmonical functions. The correlation between the original bias field and the SHFs reconstruction is* 0.63*.*

means clustering algorithm is generally used. Fig. 2.11a illustrates the clustering of a set of data $A, \ldots, J$ characterized by the bi-dimensional Euclidian distance of their position in the plane. The number of specified clusters is $K = 3$ and is indicated by the blue, green and orange ensembles.

The algorithm first starts by randomly choosing K objects promoted to be the cluster centroids. The data is then divided into K groups whose criterion is the distance to the centroids. After this, the average positions of the nearest neighbors define K new centroids. Finally, the last two steps are repeated until convergence has been reached. In the case of a pre-existing symmetric dissimilarity matrix $D_{ij} = d(X_i, X_j)$ between the individual objects $\mathbf{X} = (X_1, X_2, \cdots, X_N)$, the *K-medoids* algorithm is used instead of the K-means. In contrast to the K-means algorithm, the K-medoids algorithm chooses data objects as cluster centroids.

An extension of cluster analysis is *hierarchical clustering* with dendrograms as graphical representation. It produces a hierarchical representation of the data where the clusters at each level are created by merging those of the next lower level. The lowest level contains as many clusters as individual data and the highest level is a single cluster grouping all the data. The dendrogram of the example cited above is shown in Fig. 2.11b. The height of each node measures the intergroup dissimilarity between its two daughters. An application of hierarchical clustering is given in chapter 5.

2.4 Kullback-Leibler divergence

Before the expression of Kullback-Leibler divergence is given, a glimpse on information theory introduces some insightful concepts. The idea of *information* in statistics can be illustrated with the log-likelihood function. As a reminder, the expression of the log-likelihood for a probability density function $f(x|\sigma)$ parameterized by σ is:

$$l(\sigma) = \int_{-\infty}^{\infty} \log(f(x|\sigma)) \, dx$$

The maximum of the log-likelihood function defines the most probable parameter $\hat{\sigma}_{ML}$ given the observed data X. The Fisher-information is defined as the second derivative of the log-likelihood:

$$I(\sigma) = -\frac{d^2 l(\sigma)}{d\sigma^2}$$

Evaluated at $\hat{\sigma}_{ML}$, the Fisher-information is known as the *observed Fisher-information* $I(\hat{\sigma}_{ML})$. Intuitively, the amount of information contained in a probability density function (PDF) is related to its concavity, i.e. how sharp the density is. As an example, the smaller the standard deviation σ of a normal distribution $\mathcal{N}(\mu, \sigma)$ is, the sharper and the more informative is the PDF.

In the case of a discrete probability distribution $P(x_i)$, the expression $\log\left(\frac{1}{P(x_i)}\right)$ is called the *surprise*, as observing a highly improbable event is very surprising. Reciprocally, observing a highly probable event with probability close to one provides almost no surprise. The use of the logarithm makes the surprise additive for a product of several events.

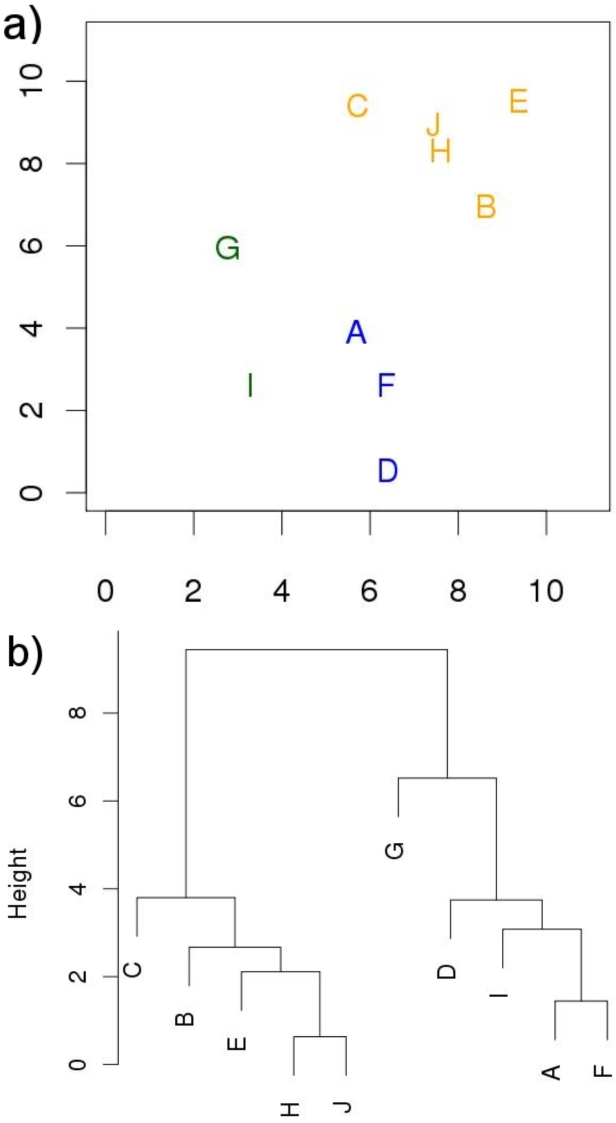


Figure 2.11: Illustration of clustering using randomly generated data. The points $A, \dots, J$ are characterized by their coordinates in the plane. (a) The data is clustered into three groups (green, blue and yellow). (b) Corresponding dendrogram showing the complete hierarchical structure. The height of each nodes measures the intergroup dissimilarity between its two daughters

For a continuous PDF, the *entropy* is defined as the expected value of the surprise:

$$H = -\int_{-\infty}^{\infty} f(x) \log\left(f(x)\right) dx$$

and is a measure of uncertainty associated with a random variable. If the density function carries no information, as in an uniform distribution $u(x) = \frac{1}{c}$ for a bounded range of x and a normalizing constant c, then the entropy H is maximal and the underlying random variable X is unpredictable within the range of x.

Suppose we have two distribution functions $f(x)$ and $g(x)$. The hypothesis that the random variable X is distributed among f (or g) is denoted by H_f (respectively H_g). The *relative entropy* is defined as

$$I(g:f) = \int_{-\infty}^{\infty} g(x) \log\left(\frac{g(x)}{f(x)}\right) dx \tag{2.2}$$

and is the mean information per observation for discrimination in favor of the hypothesis H_g against H_f. A symmetric version of the relative entropy exists and is defined as the *Kullback-Leibler divergence* $J(f,g) = I(f:g) + I(g:f)$. It measures the divergence between the hypotheses H_f and H_g, or the difficulty to discriminate between them. Even if $J(f,g)$ is symmetric, it is not a distance measure since the triangle inequality is not fulfilled. From the statistical test point of view, the equation (2.2) expresses the divergence between two hypotheses and can be used as a dissimilarity measure.

We finish this section by examining the case of multivariate normal population. For any random vector $\mathbf{X} = (x_1, \ldots, x_N)^t$ with multivariate normal distribution

$$f(\mathbf{X}) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{X}-\mu)^t \Sigma^{-1} (\mathbf{X}-\mu)\right)$$

with μ the N-dimensional mean vector and Σ the $N \times N$ covariance matrix, the expression of Kullback-Leibler divergence becomes:

$$\begin{aligned} J(A,B) &= \frac{1}{2}\mathrm{Tr}\left\{(\Sigma_A - \Sigma_B)(\Sigma_B^{-1} - \Sigma_A^{-1})\right\} \\ &+ \frac{1}{2}\mathrm{Tr}\left\{(\Sigma_A^{-1} + \Sigma_B^{-1})(\mu_A - \mu_B)(\mu_A - \mu_B)^t\right\} \end{aligned}$$

More details on this topic can be found in Kullback (1968). An application of the Kullback-Leibler divergence is given in chapters 5 and 6.

Chapter 3

Spatial scale dependence of climate model performance in the CMIP3 ensemble

Spatial scale dependence
of
climate model performance in the CMIP3 ensemble

David Masson and Reto Knutti

Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland

(published in Journal of Climate)

Abstract

About twenty global climate models have been run for the IPCC Fourth Assessment Report (AR4) in order to predict climate change due to anthropogenic activities. Evaluating these models is an important step to establish confidence in climate projections. Model evaluation however is often performed on a grid-point basis despite the fact that models are known to often be unreliable at such small spatial scales. In this study, the annual mean values of surface air temperature and precipitation are analyzed. Using a spatial smoothing technique with a variable scale parameter it is shown that the intermodel spread as well as model errors from observations are reduced as the characteristic smoothing scale increases. At the same time the ability to reproduce small scale features is reduced and the simulated patterns become fuzzy. Depending on the variable of interest, the location and the way data is aggregated, different optimal smoothing scales from the grid-point size to about 2000 km are found to give good agreement with present-day observation, yet retain most regional features of the climate signal. Higher model resolution surprisingly does not imply much better agreement with temperature observations, in particular with stronger smoothing, and resolving smaller scales therefore does not necessarily seem to improve the simulation of large scale climate features. Similarities in mean temperature and precipitation fields for a pair of models in the ensemble persist locally for about a century into the future, providing some justification for subtracting control errors in the models. Large scale to global errors however are not well preserved over time, consistent with a poor constraint of the present day climate on the simulated global temperature and precipitation response.

3.1 Introduction

In order to assess future climate changes and the anthropogenic contribution to global warming, about twenty global climate models ran scenarios in a coordinated model intercomparison targeted for the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4). These models came from various institutions and differ in their structure. While there is agreement between these models concerning the anthropogenic contribution to global warming and some agreement on the projected future changes, the uncertainty quantification is still problematic and a consensus on performance metrics for models is lacking (Tebaldi and Knutti, 2007; Knutti et al., 2010b). Some questions related to model evaluation have rarely been asked. For example, what is the typical spatial scale at which models can provide reliable results? Climate scientists have an intuitive feeling for it and use it when interpreting results. However, models are also often compared to observations or other models at the grid-point scale, sometimes leading to the conclusion that models do not even agree on the sign of predicted future changes over large areas on the globe. This is particularly true for variables other than temperature, e.g., for precipitation (see Fig. 10.12 of Meehl et al. 2007b). Evaluating the models directly on the smallest spatial scale can be misleading because resolving a feature requires several grid-points at least. Therefore, errors on small spatial scales can be large even if the models agree better on larger scales. Apart from this numerical aspect, natural variability is also an important source of model disagreement. Aggregating changes over larger regions reduces internal variability (Räisänen, 2001) and leads to more consistent projections across models even for variables that are difficult to simulate, e.g., precipitation where zonal averaging leads to a more consistent pattern across the models (Zhang et al., 2007).

While much of the community's effort goes into improving the models (shown for example by Reichler and Kim, 2008), it is unclear at what scale the models can provide useful information and agreement with data, how that scale depends on the variable and projection lead time, and whether higher model resolution leads to more useful information on a smaller spatial scale (Stainforth et al., 2007). Between the grid-point scale where models are less reliable and the global scale which is of limited use for local projections, a whole range of spatial scales exists and can be explored. Climate projections were often aggregated regionally in order to reduce model uncertainty (e.g. Tebaldi et al. 2005) but the regions were chosen in a rather ad hoc way. In the present paper we attempt to estimate optimal spatial smoothing scales for temperature and precipitation in a more formal way by minimizing a penalty function, which is a combination of the model's error to a observation-based dataset and a measure of spatial information that is lost through averaging. In simple words, the full local information is provided at every grid-point without smoothing, but model errors and model spread may be large and confidence in local projections is therefore low. On very large scales, errors are smaller and models are more likely to agree, but the information for local impacts is lost and the projection is again rather useless. Somewhere in between is a regional to continental aggregation or smoothing (implicit for example in Christensen et al. 2007) where information is most likely to be useful and robust against model assumptions.

We first test the agreement of present-day simulated climate with observations at different spatial scales. Then we focus on how model resolution affects model errors for different areas and scales. Next, we consider the ratio of the climate change signal to the model disagreement and how smoothing the data affects when and where models agree on a predicted change. Finally, we study the persistence of model errors of the initial period 1960-1979 (or equivalently the similarity of two models in the ensemble) through time and for different spatial scales in a perfect model approach. This is an important point, since past agreement with observations is often used to support projections into the future (Stott and Kettleborough, 2002; Giorgi and Mearns, 2002, 2003; Tebaldi et al., 2005; Knutti, 2008b,a), i.e. it is assumed that a model that is close to observations in the past will be close to the real world in its simulated future response.

3.2 Method

3.2.1 Data

This study uses a subset of the data produced for the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project Phase 3 (CMIP3) (Meehl et al., 2007a), a coordinated model intercomparison for the AR4 report of the IPCC. One ensemble member for each of the 24 Atmosphere Ocean General Circulation Models (AOGCMs) simulations under the A1B emission scenario (Nakicenovic et al., 2000) is used.

The observation-based data sets are ERA-40 (Uppala et al., 2005) for surface temperature and CMAP (Xie and Arkin, 1997) and GPCP (Adler et al., 2003) for precipitation. Each of these observation-based datsets obviously has its limitations and errors, and they partly rely on models as well, but still these are probably the best observation-based datasets available for our purpose. Because individual models and observations come at different spatial resolutions, the data is interpolated using bilinear interpolation to a common T42 grid (Gaussian grid associated with spectral truncation, 128 latitudinal by 64 longitudinal grid-points). The original data contains monthly averaged fields from which climatological averages over periods of 20 years (annual means) have been extracted. For future periods where no observations exist, the perfect model approach is used in some cases, i.e. each model is treated as reference once and the results are averaged afterwards.

3.2.2 Field smoothing

In order to study uncertainty of models on different spatial scales, a field smoothing technique is used. Instead of evaluating models at the grid-point scale, the fields are smoothed by weighted spatial averaging, whereby the weight $w_{i,j}$ of each point (i, j) decreases exponentially with the squared distance $d_{i,j}(k, l)$ from the original location (k, l):

$$w_{i,j}(k, l, \lambda) = e^{-d^2_{i,j}(k,l)/2\lambda^2}$$

with λ being the parameter representing the characteristic smoothing length scale of the Gaussian weighting. Therefore, the original data $V_{k,l}$ is replaced by its smoothed value $\bar{V}_{k,l}$ according to:

$$\bar{V}_{k,l}(\lambda) = \frac{\sum_{i,j}^{I,J} V_{i,j} \cdot w_{i,j}(k, l, \lambda)}{\sum_{i,j}^{I,J} w_{i,j}(k, l, \lambda)}$$

with I, J being the total number of longitude and latitude grid-points. To characterize how smooth or homogeneous the climate signal is at a certain location, we introduce the spatial variation $\sigma^s(\lambda)$, which is essentially a standard deviation of all grid-points weighted by $w_{i,j}$ defined above:

$$\sigma^s_{k,l}(\lambda)^2 = \frac{\sum_{i,j}^{I,J} \left(V_{i,j} - \bar{V}_{k,l}(\lambda)\right)^2 \cdot w_{i,j}(k, l, \lambda)}{\sum_{i,j}^{I,J} w_{i,j}(k, l, \lambda)}$$

This characterizes the spatial heterogeneity in a region determined by an area proportional to λ^2. A large spatial variation σ^s indicates that the value at that location is not very informative

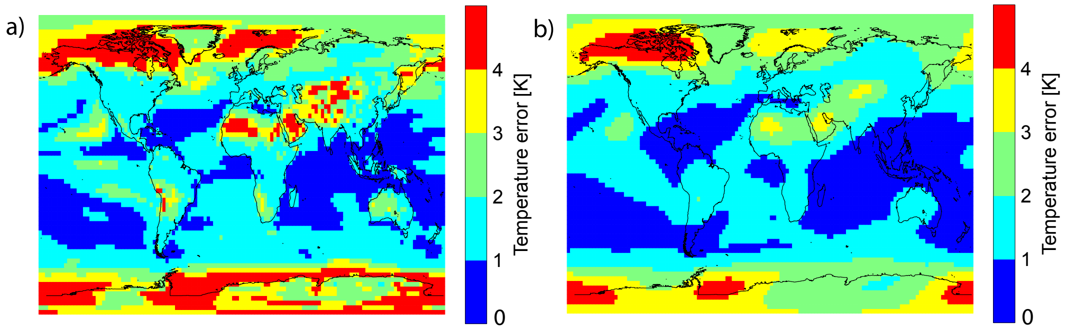


Figure 3.1: *Effect of (a) weak ($\lambda = 100$ km) and (b) medium ($\lambda = 700$ km) field smoothing on the average magnitude of the temperature error of the CMIP3 ensemble for the period 1980-1999. With smoothing some details are lost but the average error is smaller.*

about the original spatial details, whereas a small spatial variation σ^s indicates that most points within a distance of about λ are similar, such that not much of the spatial pattern is lost with smoothing. The characteristic length λ was sampled logarithmically between 100 km and 10,000 km, since most variations in the field occur at small scales and tend to decrease towards global scales. For illustration, the typical magnitude of the CMIP3 present-day temperature errors are shown in Fig. 3.1 after a small ($\lambda = 100$ km) and a medium ($\lambda = 700$ km) smoothing. As an alternative to the Gaussian smoothing, a simple average over grid-points is performed using a step-function with weight 1 inside a circular region of radius λ and 0 elsewhere. The difference between the Gaussian and the step-function smoothing is examined in the result section.

3.2.3 Measures of uncertainty

Three sources of uncertainty have been considered: 1) the model error $E_{i,j}(\lambda)$ at location (i, j) after smoothing with a fixed scale parameter λ, defined as the absolute value of the difference between the simulation and a observation-based set, 2) the spatial variation σ^s, defined as the spatial standard deviation of the variable in a given area (see definition above) and 3) the intermodel spread σ^m, defined as the CMIP3 ensemble standard deviation for a grid-point or a region. In a first part, the error and spatial variation σ^s are analyzed for the period 1980-1999 (termed present day here). For the error and this part only, we have subtracted the present day global error from all simulated data to account for the fact that some models have a global warm or cold error that is not obviously related to their ability to simulate patterns. In a second part, the evolution of these quantities over all successive periods of 20 years between 1960 and 2100 is studied.

3.2.4 Optimal smoothing for present-day simulations

One possible way to define an optimal spatial scale for present day is to use a penalty function that accounts for both the error and the spatial variation. The optimal spatial scale should consider the error magnitude E and the spatial variation σ^s as two independent components of uncertainty. To give similar relative importance to both quantities, both components are normalized with $E_{i,j}^{max}$ and $\sigma_{i,j}^{s,max}$, the local maximum values of E_{ij} and σ_{ij}^{s} across all tested spatial scales. We define the global penalty function $U(\lambda)$ as

$$U(\lambda) = \sqrt{\text{Global Mean}\left[\left(\frac{E_{i,j}(\lambda)}{E_{i,j}^{max}}\right)^2 + \left(\frac{\sigma_{i,j}^{s}(\lambda)}{\sigma_{i,j}^{s,max}}\right)^2\right]}$$

The optimal spatial scale $\lambda_{opt,m}$ for a given model m minimizes $U(\lambda)$. The optimal spatial length for the entire CMIP3 ensemble is obtained by computing the median value λ_{opt} over all $\lambda_{opt,m}$. We interpret this optimal spatial scale as a typical scale on which the model errors E are reasonably small yet a large portion of the spatial signal is preserved (small σ^s indicating that the unsmoothed values of points nearby are similar to the smoothed value in the center of the area). The resulting optimal scale depends on the choice of the normalization and the definition of the penalty function which are both partly subjective, and as a consequence the results should be interpreted as illustrative (see sections 3.3.2, 3.4.3).

3.2.5 Impact of resolution on model error

Larger computational capacities are often justified with the need for higher resolution. It is assumed that a model run at a higher spatial resolution will provide more reliable information than a model run at a lower resolution. We test this assumption by doing a regression of the error against the resolution for different smoothing values. This is particularly interesting for strong smoothing, as it answers whether higher resolution (in addition to resolving smaller features) also improves the simulation of the large scale pattern. Resolution is defined here as the typical edge length of a grid-cell, calculated as the square root of the Earth surface after dividing by the number of grid-cells.

3.2.6 Robustness of climate change signal

In order to assess the strength of the predicted climate change signal compared to the inter-model spread σ^m, the ensemble robustness ratio R is defined:

$$R_{ij}(\lambda) = \left|\frac{\Delta\bar{V}_{i,j}(\lambda)}{\sigma_{i,j}^{m}(\lambda)}\right|$$

This ratio is defined as a function of the location (i, j) and the smoothing scale λ with $\Delta\bar{V}$ being either the smoothed field of temperature or precipitation change (based on the multi-model mean value). As proposed by Murphy et al. (2004), the climate change signal is considered robust if the absolute value of the ratio is larger than 2, i.e. the predicted climate change signal is at least twice as large than the uncertainty across models.

3.2.7 Initial error preservation

Models are often evaluated and calibrated towards a observation-based dataset with the hope that this would ensure skill for a prediction. But do the initial model errors in 1960-1979 also explain the errors in the future? Is a good model for present-day still good in the future, and on what spatial scale is this relationship strongest? We examine these questions with the help of a perfect model approach. As the global error of the initial period 1960-1979 dominates the future error signal at all spatial scales, we subtract it from the data. This is justified since the focus lies rather on the spatial error pattern generated after 1960-1979. The relation between initial and future errors and simulated change and the role of the spatial scale is studied by a squared correlation index $I(t, \lambda)$ as a function of smoothing scale and projection lead time. For a given time period t and λ the following squares ρ^2 of the correlation value are calculated at each grid-point and then globally averaged:

$$I^{(1)}_{i,j}(t,\lambda) = \rho^2 \left(\vec{E_{i,j}}(1960 - 1979, \lambda), \vec{E}_{i,j}(t,\lambda) \right)$$
$$I^{(2)}_{i,j}(t,\lambda) = \rho^2 \left(\vec{E_{i,j}}(1960 - 1979, \lambda), \vec{C_{i,j}}(t,\lambda) \right)$$

with $\vec{E}_{i,j}$ being a vector of the 23 error values for a certain time, smoothing and grid-point (i, j). The length of the vector is 23 and is equal to the number of CMIP3 models 24 minus one model that serves as reference to calculate the error. This perfect model approach is repeated 24 times so that each model is used as reference once. The 24 possible squared correlation indices $I(t, \lambda)$ are then averaged and return a single representative explained variance for a certain time period and smoothing. $\vec{C_{i,j}}$ is the difference between the simulated variable change (temperature or precipitation) of two models for a certain time, smoothing and grid-point. We assume the relations to be linear. The squared correlation coefficient thus equals the explained variance under the hypothesis that the predicted variable is normally distributed at any predicting value (von Storch and Zwiers, 2004, section 8.2.4). This hypothesis is met by the distributions of the projected error magnitudes among the 23 simulations. As the explained variance is an additive value, $I^{(1)}_{i,j}$ and $I^{(2)}_{i,j}$ can be globally averaged for a given time t and smoothing scale λ. The quantity $I^{(1)}$ therefore measures the fraction of future error in the variable that is explained by the initial error in 1960-1979 (control error), while $I^{(2)}$ measures the fraction of error in the variable change (i.e., the simulated difference rather than the variable itself) that is explained by the initial error in the reference period 1960-1979. In other words, $I^{(1)}$ describes the persistence of the initial errors over time, whereas $I^{(2)}$ describes the relation between initial errors and trend errors. High explained variance values in $I^{(2)}$ indicate that the mean state climate for the present day period is a good indicator for the model consistency in the future, i.e. two models with a similar present day state will simulate similar changes, and therefore a model close to the present day climate of the real world would hopefully produce an accurate prediction of the changes of the real world.

3.3 Results

3.3.1 Field smoothing and measures of uncertainty

In a first step, the error and spatial variation σ^s (absolute values averaged over space) are quantified for present-day simulations at various smoothing scales and for each model. The largest errors compared to the observation-based dataset are found at the smallest tested smoothing ($\lambda = 100$ km, essentially equivalent to no smoothing) for all models. As shown in Fig. 3.2 the errors decrease monotonically and all curves converge to zero as λ increases. Note that the global error was initially subtracted. For precipitation the error reduces faster than for temperature and even a weak smoothing significantly improves the agreement with observation.

The ranking of the models (from the error point of view) depends on the spatial scale considered. In general, a model performing well on small scales tends to also perform well after smoothing while the opposite is not necessarily true. Because the models get more and more similar with increasing spatial scales, performing well at large scales does not guarantee good agreement with the observation-based dataset at smaller scales. A more detailed study of how such model rankings evolve over time is given later in section 3.3.5. Two observational datasets (CMAP and GPCP) are available in the case of precipitation. If one reference is treated as the true data and the other serves as an additional model, the best performing model at local scales is the alternative observation-based dataset. However, that is not true at large scales. In the case of CMAP being the reference, GPCP is even among the five worst models. Not surprisingly, both datasets differ from all models on small scales, because the models are unable to resolve some small scale patterns, while this does not seem to hold for large scales. Without judging which observation-based dataset is more realistic, this analysis highlights that observational uncertainty in variables other than temperature may be large and should be considered when developing metrics for model evaluation.

In contrast to the error, the spatial variation σ^s is monotonically increasing, as shown in Fig. 3.3. For the smallest smoothing, the spatial variation σ^s is near zero, because the standard deviation over all points gives virtually no weight to neighboring points. At large scales, it converges to a constant value representing the standard deviation across all grid-points. For temperature data, the curves are approximately constant above 6000 km while for precipitation they stabilize earlier at 4000 km.

Because different smoothing techniques are likely to produce different results, the Gaussian smoothing and a simple average over neighboring grid-points using a step-function (see section 3.2.2) are compared. Fig. 3.4 shows the typical CMIP3 global average error magnitude E and spatial variation σ^s as function of the spatial length. Booth techniques return the same values at the grid-point and the global spatial scales. The largest difference is the rate of change which is faster in case of the Gaussian smoothing. The step-function smoothing shows irregularities between 0 and 1000 km as a hard threshold is more likely to create artifacts when moving across mountain ranges or coastlines. The qualitative behavior however is similar.

The intermodel spread σ^m (measured as the standard deviation across all models after smoothing and representing model dissimilarities) are larger at local scales and can be reduced with a stronger smoothing, as in the case of the error (see Fig. 3.5a). At local scales, the intermodel spread σ^m over time remains relatively constant. At large scales however, σ^m is smaller but clearly increases with time. The reason is that global scale dissimilarities are related to the transient temperature change and evolve with the same magnitude as the global error (Knutti et al., 2008b). The local dissimilarities however are less related to global warming and dominated by model errors, and thus almost time independent. The reduction of the intermodel spread σ^m for precipitation occurs more rapidly than for temperature, similar to the case of the error (see Fig. 3.5b). In contrast to temperature, the precipitation intermodel spread σ^m does

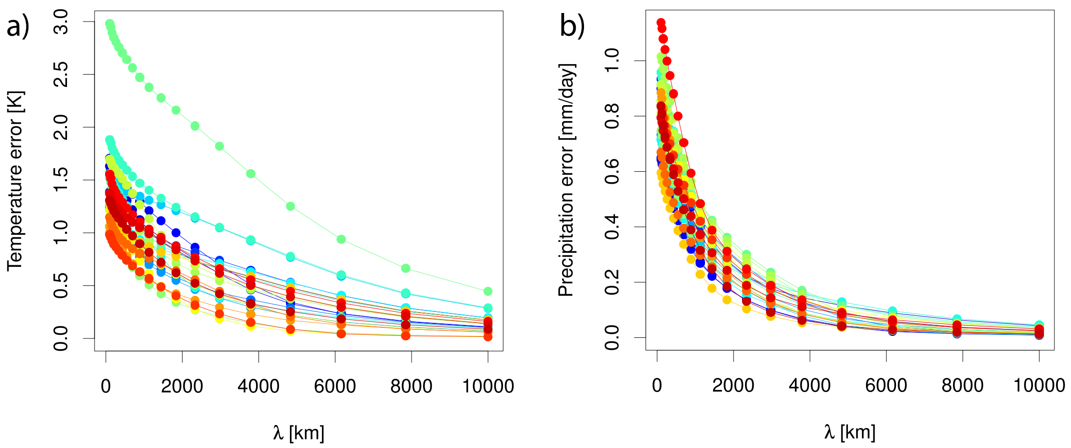


Figure 3.2: *Global average of the absolute value of the error as function of the spatial scale* λ*, for each CMIP3 model. (a) Temperature (ERA-40 as reference) and (b) precipitation (CMAP as reference) for the period 1980-1999. Smoothing the data reduces the error between simulation and observation.*

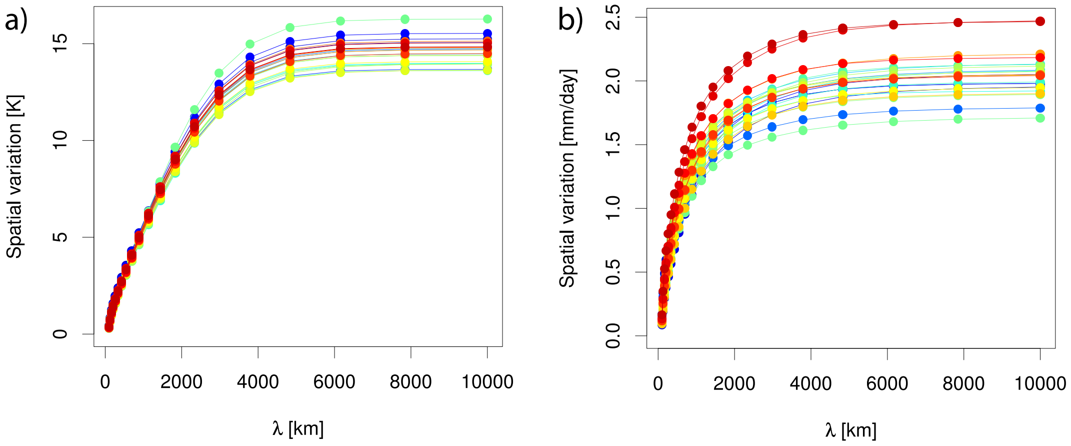


Figure 3.3: *Global average of the spatial variation* σ^s *as function of the spatial scale* λ*, for each CMIP3 model and for a) temperature and b) precipitation for the period 1980-1999. The spatial variation is defined as the spatial standard deviation of the variable within a given spatial area. The larger the spatial scale, the larger the standard deviation of the variable encompassed by the smoothing.*

not change much over time at any of the tested scales because the precipitation trends are rather small compared to the model dissimilarities.

3.3.2 Optimal smoothing for present-day simulations

The optimal smoothing that minimizes the root-mean-square value of the penalty function $U_{i,j}(\lambda)$ indicates an approximate scale where the error is reasonably small through spatial averaging, yet most of the local climate pattern is preserved. The results for different regions and smoothing techniques range between 185 km and 2080 km and are shown in Table 3.1. The step-function smoothing produces optimal scales about two times larger than the Gaussian smoothing but the results are qualitatively similar. The reason is that the step-function eliminates all influences from grid points further away that the distance λ, whereas the Gaussian gives non-zero weight even to very remote points. In general, the temperature field allows a larger smoothing than the precipitation field because the spatial variation for temperature is

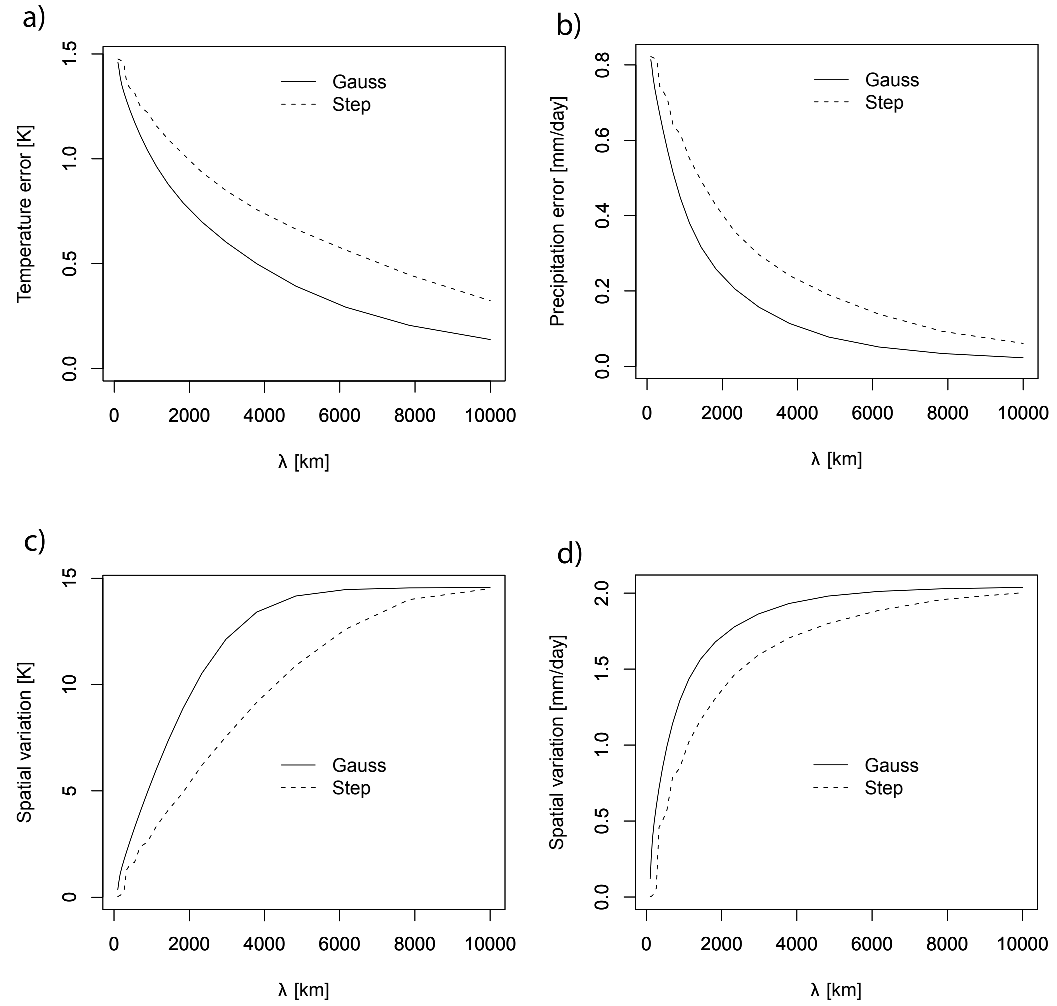


Figure 3.4: *Gaussian (solid line) versus step-function (dashed line) smoothing for surface temperature (left column) and precipitation (right column) applied to the typical CMIP3 global average error magnitude (first row) and spatial variation (second row).*

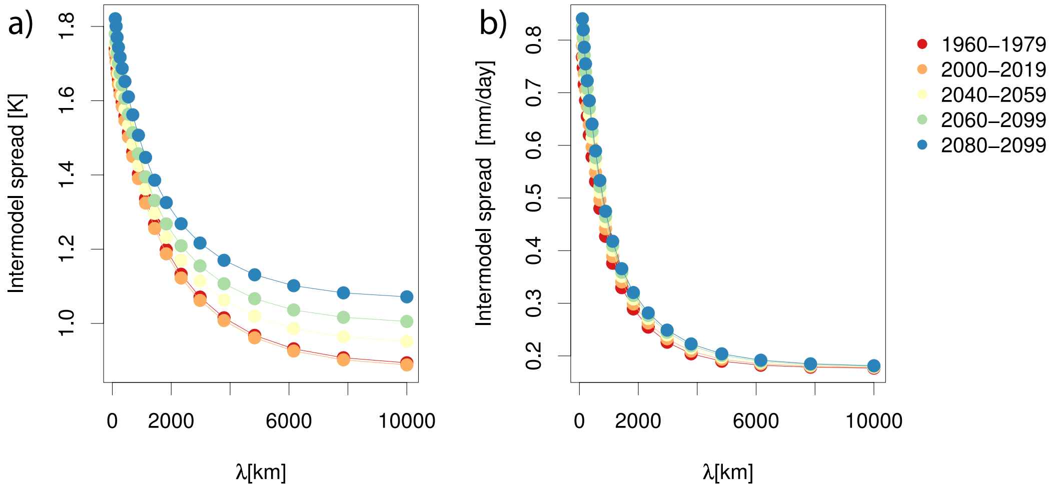


Figure 3.5: *Intermodel spread σ^m representing the model dissimilarities, globally averaged and as a function of the spatial scale λ for a) temperature and b) precipitation for different time periods from 1960 to 2080. Models tend to show smaller dissimilarities at larger spatial scales.*

smaller. This is even more pronounced in the tropics for convective precipitation, where the spatial variation is much larger than the error magnitude, leading to an optimal smoothing close to the grid-point scale. As the choice of the penalty function is partly subjective, we have investigated two other definitions. For example, if the error and the spatial variation are simply added without normalization, σ^s quickly dominates E and the grid-point scale is the optimal choice. Different applications may require different weighting in the penalty function. Rather than defending any particular choice of a penalty function, the idea here is to demonstrate the two opposing trends of model error and spatial variation. Trying to minimize both of these components implies a typical length scale over which the model results should be aggregated. That length scale is larger for temperature than for precipitation globally, larger in the tropics for temperature, larger in the extratropics for precipitation, and larger over ocean than over land for temperature. These general conclusions should be robust against different definitions of the penalty function, provided that an optimal scale exists.

Table 3.1: *Optimal smoothing length (in km) for surface air temperature (TAS) and precipitation (PR) for various regions using the Gaussian or the step-function smoothing.*

	Gauss		Step-function	
Region	TAS	PR	TAS	PR
global	695	207	1280	428
tropics	1130	162	2340	428
extra-tropics	428	1130	886	2080
land only	428	207	886	428
ocean only	886	185	1440	428

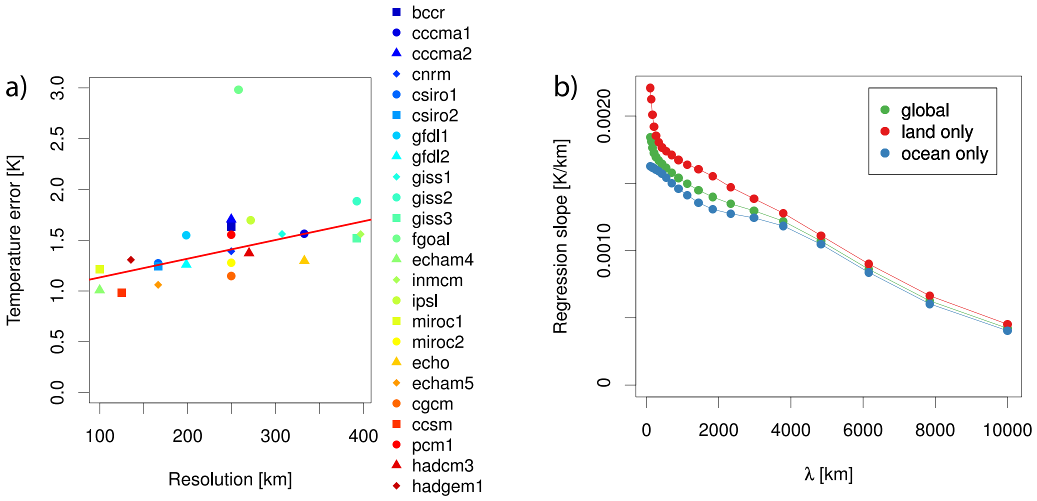


Figure 3.6: *(a) Relation between absolute value of the temperature error (globally averaged) at the grid-point scale ($\lambda = 100$ km) and native model resolution. The linear regression is indicated by a red line (with the fgoals model excluded). (b) Regression slope versus λ for land or ocean only and the entire globe. Whereas a higher resolution improves performance at local scales, its impact at large scales is limited. In the case of precipitation no relation between resolution and error could be established.*

3.3.3 Impact of the resolution on model error

The correlation between error and model resolution (i.e., the original resolution at which the model is run, not the smoothing scale) is an indication of the benefit of higher resolution in representing current climate. In the case of precipitation, correlations between error and resolution were lower than 0.5 and the regression slopes were never statistically significant using the F-test with a 0.05 significance level, thus no clear relation seems to exist at least within the relatively narrow range of resolutions covered by CMIP3. In Fig. 3.6a, a scatter plot of the relation is shown for temperature at the grid-point scale ($\lambda = 100$ km), considering the whole globe (land and oceans). At this scale, the error is reduced from about 1.5K to 1K from the coarsest to the highest resolution. The slope of the regression line characterizes the strength of the relation between the resolution and the error and is calculated for all spatial scales, global, land and oceans. The fgoals model is a clear outlier and is excluded for this part of the analysis. The linear regression slopes are displayed in Fig. 3.6b and are always statistically significant using the same test as before. Not surprisingly, the benefit of high resolution is largest at the smallest scales and over land, where the topography is more complex and higher resolution can probably resolve more local processes. However, the benefit of the resolution quickly decreases with smoothing approaching 500 km. At scales above about 500 km, the slope dependence on λ is similar in all cases.

3.3.4 Robustness of climate change signal

The ensemble robustness R is the absolute value of the ratio of the climate change to the intermodel spread σ^m. The geographic distribution of the robustness at the grid-point scale

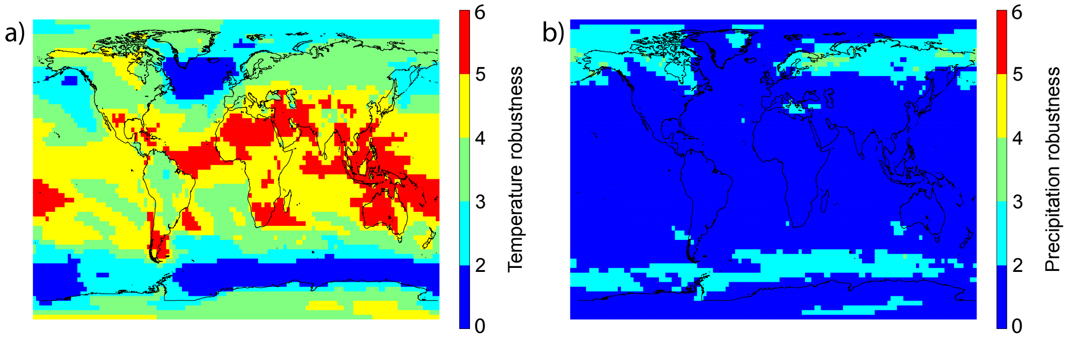


Figure 3.7: *Ensemble robustness* R *for the 2080-2099 climate change signal since 1960-1979 at the grid-point scale (*$\lambda = 100$ *km), for temperature (a) and precipitation (b). The climate change signal is said to be robust if* R *is larger than 2.*

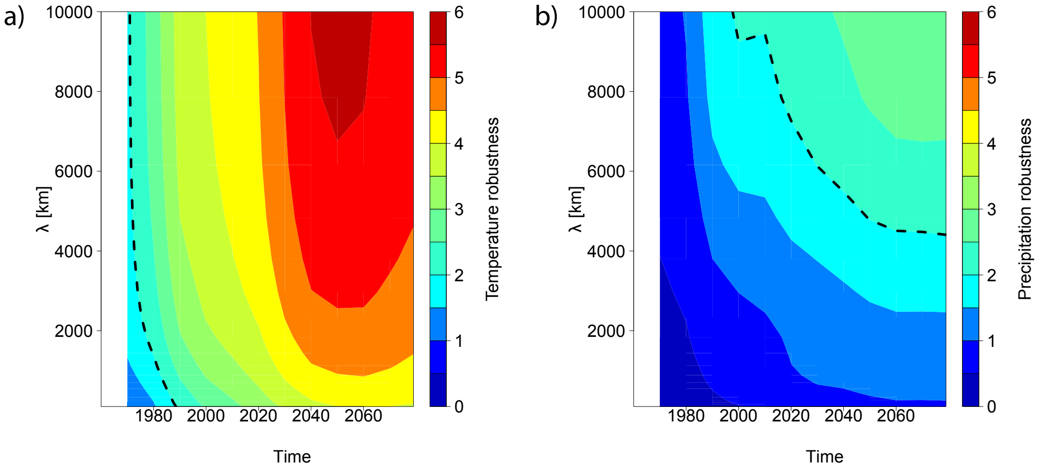


Figure 3.8: *Ensemble robustness* R *for the climate change signal since 1960-1979 (globally averaged), for temperature (a) and precipitation (b).* R *is defined to be robust above the dashed line where the signal is twice as large as the standard deviation characterizing the intermodel spread. While temperature simulations are mostly robust everywhere, precipitation only gets robust later and for continental or large scales.*

($\lambda = 100$ km) is first shown in the maps in Fig. 3.7 for the end of the century as an example. A striking but well known feature is that temperature changes are clearly more robust than precipitation (Räisänen, 2001). While temperature robustness is especially weak in the North Atlantic and over the Southern Ocean where some models indicate cooling while most show warming, the models agree for the rest of the globe. Precipitation simulations show good agreement over high latitudes and the Mediterranean sea. The temporal and spatial behavior of R is depicted in Fig. 3.8. Not unexpected and in agreement with detection/attribution and future projection studies (Barnett et al., 2005; Meehl et al., 2007b), the temperature signal is robust at all scales after a few decades and clearly exceeds the intermodel spread σ^m. In contrast, simulated precipitation changes agree about 50 years later and on continental scales only, both due to larger model differences and large interannual variability.

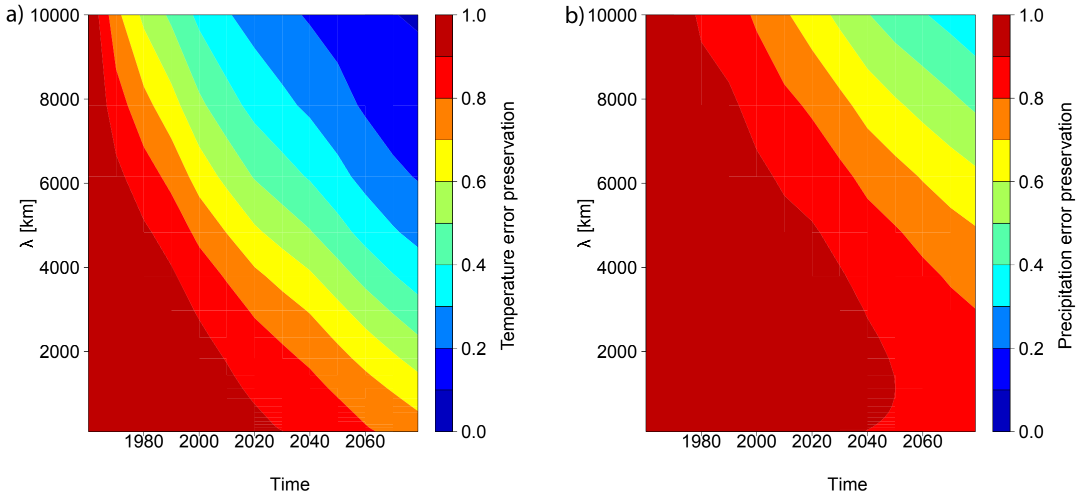


Figure 3.9: *Preservation of the initial errors among the models through time and spatial scales, defined as the fraction of variance of the future errors that is explained by the initial model errors in 1960-1979 (globally averaged). The time axis is divided into 12 intervals between 1960 and 2099. The perfect model approach was used here for (a) temperature and (b) precipitation. Values near one indicate that differences between models strongly persist over time, while values near zero indicate no relation between differences at the beginning and the end of the simulation.*

3.3.5 Initial error preservation

The global average of the explained variance values between initial and future model errors, $I^{(1)}(t, \lambda)$, is shown in a contour plot in Fig. 3.9 as function of time and spatial scale. The time axis is divided into 12 intervals between 1960 and 2099. For both temperature and precipitation the explained variance is high at local scales until the end of the 21st century. At that time and at larger scales however the explained variance vanishes. The decay of the explained variance with time is slower at local scales than at large scales. In agreement with Giorgi and Mearns (2002) and Räisänen (2007), a linear regression between initial and future model errors for scales and periods exhibiting a high correlation yield regression slopes very close to one, supporting the conclusions that initial differences between models are well preserved over time, in particular on small scales. The relation $I^{(2)}(t, \lambda)$ between initial model differences and trend errors was also analyzed. In contrast to the case before, only very small correlations are seen for both temperature or precipitation (not shown). Therefore, the initial error seems to relate only weakly to simulated future trends. The interpretation of these results is given in the Discussion section.

3.4 Discussion

3.4.1 Data

Regarding the data, we have used the A1B scenario for the analysis of future projections, but the choice of a specific scenario should not impact the presented results, because the ratio of temperature change to forcing is approximately constant across scenarios (Knutti et al., 2008b; Gregory and Forster, 2008), and simulated patterns tend to be similar for all scenarios (e.g.

Meehl et al., 2005; Washington et al., 2009; Meehl et al., 2007b). The number of runs available for each model varies from one to nine but very few models have more than five runs. We selected arbitrarily the first run in the list. This choice is not critical since internal variability after averaging the data on 20 years is relatively small in comparison to intermodel differences. Choosing periods of 20 years is a compromise between having enough data inside a period to avoid too much internal variability and having enough periods to cover the observation period between 1960 and 2000. Averaging initial condition ensemble members only for some of the models would inappropriately reduce variability for some models. The model evaluation was only done for temperature and precipitation, but of course other important fields such as sea-level pressure would be important to consider for a more complete study.

3.4.2 Measures of uncertainty

The magnitude of the error rather than error itself has been chosen for the analysis because part of the multi-model error get canceled in a multi-model mean (Räisänen, 2007) and the tendency of the error magnitude to decrease with increasing smoothing is masked. Our results suggest that smoothing the data before the model evaluation reduces the disagreement between the models and an observational dataset, in particular in the case of a variable that is difficult to simulate and locally heterogeneous, such as precipitation. This does not ignore the weaknesses of the models but may help to focus on features that are relevant at a certain spatial scale. Two datasets were used to evaluate precipitation and it is interesting that on larger scales the similarity between some models and one dataset is larger than the similarity of the two datasets. For small scales, the two datasets are similar and all models are different because the models all share similar limitations in parameterizing or discretizing physical processes (Tebaldi and Knutti, 2007; Knutti et al., 2010b) and are unable to correctly capture certain small scale patterns in precipitation.

As model errors decrease when results are aggregated on larger scales, grid-points from more climatic regimes are averaged together and the uncertainty related to the spatial variation σ^s increases. The information provided by the models gets blurred and the simulation signal is less precise than at the grid-point scale. Similar to the error, the intermodel spread σ^m for precipitation is reduced faster by smoothing than for temperature. This is because precipitation differences are dominated by small-scale features whereas temperature differences vary at larger spatial scales. Although the intermodel spread σ^m of temperature is lowest at large scales, it increases faster with time at large scales due to the different transient warming rates of the models. The case of precipitation is interesting because the intermodel spread σ^m is almost constant in time at each tested spatial scale. The reason is that precipitation changes are only on the order of a few percent and can be of opposite sign in nearby areas and therefore get partly averaged out at larger spatial scales. On the other hand, simulated precipitation in the baseline climate can easily vary by a factor of two locally, so the intermodel spread σ^m is dominated by the climatological errors at all times.

3.4.3 Optimal smoothing for present day simulations

Combining the error and spatial variation σ^s into an overall penalty function where the error is substantially reduced yet the main spatial patterns are still preserved is possible but the results depend on the definition adopted. Our results suggest that there are different optimal scales, depending on which variable is analyzed as well as which region is considered. From a numerical perspective several grid-points are needed to discretize the partial differential equations governing the climate models. Therefore, interpreting scales smaller than at least several grid-points is dangerous because of numerical instabilities and errors generated. It is interesting to compare the optimal scales obtained here with the choices made in publications that aggregate regional climate change (e.g. Giorgi and Francisco 2000; Tebaldi et al. 2005; Christensen et al. 2007; Mahlstein and Knutti 2009). Many of these studies provided regional averages over 26 land regions. If we divide the total land area by 26 and take the square root of that (corresponding to the length and width of a region if the land was divided into 26 regions of equal area), we find a characteristic length scales of about 2400 km, larger than those obtained here with the Gaussian smoothing but in closer agreement with the step-function scales. We argue that climate scientists may in fact often choose regions and scales based on a similar informal optimization procedure, trying to maximize the regional detail (e.g. for impact studies). But knowing that errors and intermodel spread σ^m are largest at local scales, they aggregate results into regions encompassing multiple grid points. Of course, other aspects such as the communication of the results play a important role, and the optimal spatial scales found in this analysis should be seen as an approximate estimate that depends on the location, the variable, the temporal and spatial variability as well as the uncertainty in the observation. Different definitions of a penalty function and an optimal spatial scale are possible, and the one chosen here should be interpreted as an illustration that provides insight into how different quantities depend on the spatial scale, rather than as a definitive answer.

3.4.4 Impact of resolution on model error

The resolution of models is correlated with their performance in simulating temperature, but apparently not precipitation (at least in the CMIP3 ensemble). This might be because resolving some processes related to clouds and precipitation would require much higher resolution than any of the global models currently have.

Alternatively, it might also be that higher-resolution models produce precipitation with higher geographical variability. Since the precipitation field is more variable than temperature, a simple shift in precipitation pattern could penalize the model performance. High resolution models have the clearest advantage in reproducing current temperature over land, on scales between local and 500 km, partly because of a better representation of the topography. The globally averaged error is reduced from about 1.5 K to 1 K from the coarsest to the highest resolved models at the smallest scale, where the relation is strongest. However, given the cost of higher resolution the benefit may be seen as rather small. As already noted by Santer et al. (2009), the Canadian Climate Centre's CGCM3.1 and the Japanese MIROC3.2 were both run at higher and lower resolution configurations, but despite the use of higher resolution the error was not much reduced. For some variables, a higher resolution may eliminate a parameterization and allow direct dynamical computations instead. In the ocean for example, a lot of energy is contained in small-scale eddies. Therefore the relation between resolution and sea surface temperature might be stronger than for surface air temperature. In general, it is not easy to separate the effects of higher resolution and a more comprehensive representation of processes. The groups running models at highest resolution are often also those with the longest experience in building models and with the largest number of people developing the model. So resolution, rather than just a numerical property, should probably be seen more as an indicator of overall sophistication, effort, computing power and financial resources going in a model.

3.4.5 Robustness of climate change signal

Maximizing the ensemble robustness R is desirable and is likely to improve with newer climate models generations. A closer look at Fig. 3.8b for precipitation compared to Fig. 10.12 of Meehl et al. (2007b) summarizes the benefit of this study: if models are compared at the grid-point scale they do not agree in their trends over large areas on the globe, whereas our study shows that they do, but only for regions with 4000 km as typical scale and trends beyond the period 2050-2069. Trends are relatively weak and local precipitation is difficult to simulate, therefore larger regions are needed in order to detect the precipitation change signal and simulate robust trends, consistent with the results of the precipitation attribution study by Zhang et al. (2007). Note that even if the global average of the robustness is below a given threshold, there are of course regions where the robustness is high. For example, the simulated increase in precipitation in high northern latitudes is significant and robust even at small scales and in the near future.

3.4.6 Initial error preservation

The climatological errors in mean temperature and precipitation in temperature and precipitation in the CMIP3 models are surprisingly constant over time, in particular on small scales. This error preservation vanishes towards the end of the century at large scales, indicating that differences in climate change have a larger scale than the differences in error magnitudes. This is consistent with the fact that climatological errors and simulated changes are weakly correlated (Murphy et al., 2004; Knutti et al., 2006, 2010b). Local errors are more likely to be the result of deficiencies in simulating particular processes that are important locally (e.g. not resolving a mountain range), and these are usually more persistent over time. The explained variance between initial and future errors is larger for precipitation than for temperature, which may seem surprising at first since precipitation is more difficult to simulate. The reason for this persistence is that changes in precipitation are quite small in many regions compared to the control errors. Therefore, future errors are essentially a sum of initial errors plus some trend, with the former dominating the latter. The result is that two models having similar errors in their initial state will tend to have similar errors in the future at the same location.

The lack of correlation between the initial errors and future trends errors is rather disturbing. It is commonly assumed that models with small initial error are more accurate in predicting future trends (e.g., IPCC AR4 FAQ 8.1, Randall et al. 2007). But reality suggests otherwise, as shown by the lack of correlation between past or future predicted warming with present-day simulated temperatures (Tebaldi and Knutti, 2007; Jun et al., 2008b; Knutti et al., 2010b; Weigel et al., 2010). As a consequence, knowing the discrepancy between present-day simulations and observations of the mean climate state does not immediately help to constrain the estimation of future trend error. On the other hand, there is clear evidence and physical reasons for a relation of past greenhouse gas attributable warming and future warming (e.g. Stott and Kettleborough 2002). Clear relations also exist for local processes, e.g. a correlation between past and future sea-ice loss in the Arctic (Boé et al., 2009b). These points are important to keep in mind when weighting the models for the future projections, with weights based on performance in the past (Knutti et al., 2010a; Knutti, 2010).

3.5 Conclusion

Although small spatial scales are most important to determine specific climate impacts, this is precisely the scale where climate is most difficult to simulate and where model errors and intermodel spread are largest. Climate scientists therefore often aggregate data to regions where the above problems are less severe, even though some spatial information gets lost in that processes. Here we have done this in a formal way and have demonstrated how the spatial variation σ^s, the model spread σ^m, the model error in climatology and the persistence of errors depend on the spatial scale of averaging. We have shown that the error and the intermodel spread can be significantly reduced by smoothing the data (consistent with earlier results by Räisänen 2001), however at the price of losing spatial detail (expressed in our case as an increase in the spatial spread).

Our results support typical scales between the grid-point and 2000 km depending on the variable, the location and the smoothing technique. Although there are of course various definitions of an optimal scale for different problems, we suggest that some form of spatial aggregation should be considered to provide a more robust estimate of climate change.

Our analysis also reveals that model resolution in CMIP3 seems to only affect performance in simulating present day temperature for small scales over land. Result may differ for other quantities, but given the limited advantages of high resolution even for reproducing present day climate, we speculate that pushing model resolution alone is unlikely to improve future predictions and reduce uncertainties, unless a more complete understanding of the physical and biogeochemical process is incorporated and the models are re-calibrated. This is consistent with the fact that uncertainties in climate projections have not decreased significantly in the last decade despite massive computational advances allowing for higher model resolution. As a consequence, regional high resolution models and downscaling may provide greater spatial detail but not necessarily higher confidence in local projections to determine the impacts of climate change.

Finally, we have shown that the initial model errors of 1960-1979 persist over time in particular at small spatial scales, justifying to some extent the common practice to focus on anomalies from a control simulation rather than absolute values (although that is unlikely to work well for more complicated quantities, see Buser et al. 2009). In agreement with earlier studies, there is a lack of correlation between straightforward measures of initial errors and future trends. We have shown the difficulty in relating model skill based on present day climatological errors (as characterized by Reichler and Kim 2008) to prediction skill in the future, whatever spatial scale is chosen.

Acknowledgments

We acknowledge the international modeling groups for providing their data for analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the model data, the WCRP/CLIVAR Working Group on Coupled Models (WGCM), and their Coupled Model Intercomparison Project (CMIP) and Climate Simulation Panels for organizing the model data analysis activity, and the IPCC WG1 TSU for technical support. The CMIP3 Multi-Model Data Archive at Lawrence Livermore National Laboratory is supported by the Office of Science, U.S. Department of Energy.

Chapter 4

Constraining climate sensitivity from interannual variability: an illustration of tuning in climate model ensembles

Constraining climate sensitivity from interannual variability: an illustration of tuning in climate model ensembles

David Masson and Reto Knutti

Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland

(in revision in Journal of Climate)

Abstract

Climate sensitivity has been remarkably difficult to constrain by comparing the simulated mean climate state from different models with observations, in particular for small ensembles with structurally different models like CMIP3. In this study, an empirical relationship between the interannual variation of seasonal temperature is found to be strongly correlated to climate sensitivity. The resulting 95% confidence interval for climate sensitivity is $1.9 - 4.9$ K with the best agreement for 3.3 K. In order to understand the nature of this relation, a multi-thousand-member ensemble of climate models from the climateprediction.net (CPDN) project and a pattern recognition algorithm called "random forest" are used. While the random forest algorithm successfully predicts the CPDN climate sensitivity values, the relation is non-linear and complex. Yet, if simulations are tuned for better agreement with a reference dataset, a simpler linear relation emerges. This supports the hypothesis that the CMIP3 models are tuned for a better agreement with the observations. As a possible consequence, the estimated uncertainty in climate sensitivity is smaller than in an ensemble without tuning. However, despite strong correlations with climate sensitivity, interannual variability does not constrain the range of climate sensitivity beyond the initial distribution of CMIP3 climate sensitivities. We speculate that this is partly because the feedbacks determining interannual variability as well as climate sensitivity are those that are already tuned through the mean climate and seasonal cycle, such that the data has effectively been used before in the model development processes and is of little value to further constrain climate sensitivity.

4.1 Introduction

General circulation models (GCMs) are tools to understand climate processes and are often applied to make projections for the next decades to centuries. The discretization of the equations of motion on a grid is subject to several choices (resolution, numerical schemes, hard- and software, etc.) and the need for parameterizations leads to structural uncertainties that are difficult to explore systematically by perturbing parameters in a single model (Stainforth et al., 2007; Knight et al., 2007). At present, the data from the Coupled Model Intercomparison Phase 3 (Meehl et al., 2007a), collected for the Intergovernmental Panel on Climate Change Fourth Assessment Report (IPCC AR4) (IPCC, 2007) forms the largest group of structurally different global models. However, this ensemble amounts to only 24 different models, and these models are not independent (Jun et al., 2008b,a; Knutti et al., 2010b). It is also not clear how these models should be averaged in a multimodel mean to produce a projection of the global temperature increase (e.g., Knutti et al., 2010b).

Models show somewhat different patterns of warming, but also disagree about the global magnitude of the projected changes. A central quantity to express their overall response to increasing greenhouse gases is climate sensitivity, the global mean equilibrium near-surface warming for doubling the atmospheric CO_2 concentration. It is the single most important quantity determining the impacts of climate change to a given forcing on long time scales. The CMIP3 climate sensitivity values range from approximatively 2 to 5 K. A more reliable estimate of the confidence interval can be obtained by constraining climate sensitivity from observations. Unfortunately, very few observational variables can strongly constrain climate sensitivity (see Knutti and Hegerl (2008) for a review). For the CMIP3 ensemble, Shukla et al. (2006) found a relationship between climate sensitivity and the ability of a model to simulate the present-day annual cycle and seasonal anomalies for surface temperature, where both the magnitude and the pattern are expressed in terms of a relative entropy between the simulated and observed distribution. The correlation is -0.74 and their best estimate favors a true climate sensitivity close to the highest simulated value (i.e. approximatively 5 K). Based on Covey et al. (2000), Wu et al. (2008) found a relation between the seasonal cycle amplitude and climate sensitivity in CMIP3 but the agreement is not perfect and explains only 40% of the variance in climate sensitivity. In the climateprediction.net (CPDN) ensemble, the climate sensitivity values could be approximated using the patterns of the seasonal cycle as input in a neural network pattern recognition technique (Knutti et al., 2006). Similar ideas of relating observed features of climate to feedbacks and climate sensitivity have been used in other studies based on the CPDN ensemble (Piani et al., 2005; Murphy et al., 2004; Sanderson et al., 2008) but in all cases the uncertainty ranges reflect only the parametric uncertainty of the HadSM3 model. Wu and North (2003) proposed a measure of interannual variability in surface temperature that was correlated with climate sensitivity in an earlier ensemble of structurally different GCMs, but their work got relatively little attention. The analysis is repeated here for the CMIP3 ensemble and a significant correlation is found (section 4.3.1). In order to understand this relation, the larger CPDN ensemble is in turn analyzed in section 4.3.2. The relation to climate sensitivity is no longer linear and follows a complex pattern that can be approximated with a random forest technique. We propose that this difference is due to the fact that the CMIP3 ensemble is tuned in order to reproduce the observations, in contrast to the CPDN ensemble where the parameters values are freely combined. Evidence for tuning in the CMIP3 ensemble has already been reported in the past. For example, the total radiative forcing is a large source of uncertainty because of unknowns mainly related to aerosol effects. Kiehl (2007) and Knutti (2008b) have shown that the CMIP3 models compensate high climate sensitivity values by low total radiative forcing without any obvious physical reason, except to match the warming of the 20^{th} century. The quantification of climate sensitivity depends critically on the feedback processes (Knutti et al., 2008b). Huybers (2010) showed that the albedo, lapse rate, water vapor, cloud feedback parameters strongly co-vary within the CMIP3 ensemble possibly as a result of tuning the energy balance, indicating that the $2-5$ K climate sensitivity range has already been reduced in comparison to what it might have been without covariance. On the other hand, the lack of correlation where it would physically be expected is surprising but might in fact support the tuning hypothesis. For example, the CMIP3 warming trends over the 20^{th} century and climate sensitivity values are not significantly correlated although one would expect that models with higher climate sensitivity would simulate higher transient climate response (Tebaldi and Knutti, 2007). The reason is that different aerosol forcings can obscure the picture, and different combinations of forcings and feedbacks yield similar trends (Knutti et al., 2002; Knutti, 2008b; Kiehl, 2007; Forest et al., 2002). But if the 20^{th} century warming trends have already been integrated in the models, it is likely that this information cannot be used again to constrain climate sensitivity because the model spread of transient warming has become too narrow.

What impact does tuning have on the uncertainty range of climate sensitivity and to which extent is it problematic? Tuning is sometimes seen as interfering with the falsification theory (Randall and Wielicki, 1997). In the philosophy of science, falsifiability is the possibility for a model to be falsified by observation (Popper, 1959). If all observations have already been used to tune a model, its refutation will be harder to achieve and the model results might give a false impression of validity (Oreskes et al., 1994). Alternatively, the equivalent term ”model calibration” is more neutral but refers in fact to the same process.

The goal of this paper is twofold: first we constrain the CMIP3 climate sensitivity range using interannual temperature variability. Second, we illustrate how tuning can play a role in either creating or masking a correlation with climate sensitivity within an ensemble of simulations. In section 4.2 we introduce the observational temperature references used and two simulation ensembles: the CMIP3 ensemble and the CPDN ensemble. In section 4.3.1, the range of climate sensitivity constrained by observation is presented for the CMIP3 ensemble. In section 4.3.2, the relevance of this constraint is justified using a pattern recognition algorithm in order to capture the CPDN climate sensitivity values. The reason of the linear relation between interannual variability and climate sensitivity is illustrated in section 4.4 by tuning the CPDN simulations. A discussion and conclusion are given in section 4.5.

4.2 Models and data

This study uses results from the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project Phase 3 (CMIP3, Meehl et al., 2007a), a coordinated project to gather and compare the different of GCMs used for the IPCC AR4 report (IPCC, 2007). The CMIP3 experiment used here is pre-industrial temperature control simulations with no external forcing and for the models listed in Table 4.1.

The variable of interest is surface air temperature and is compared to four observational and reanalysis datasets: the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis (ERA-40; Uppala et al., 2005), the Hadley Centre Climatic Research Unit instrumental dataset (HadCRUT3; Brohan et al., 2006), the National Centers for Environmental Prediction National Center for Atmospheric Research reanalysis (NCEP-NCAR; Kalnay et al., 1996) and the MERRA (GMAO, 2009) reanalysis. Because individual models and observations come at different spatial resolutions, the data is interpolated using bilinear interpolation to a common T42 grid (Gaussian grid associated with spectral truncation, 128 latitudinal by 64 longitudinal grid-points).

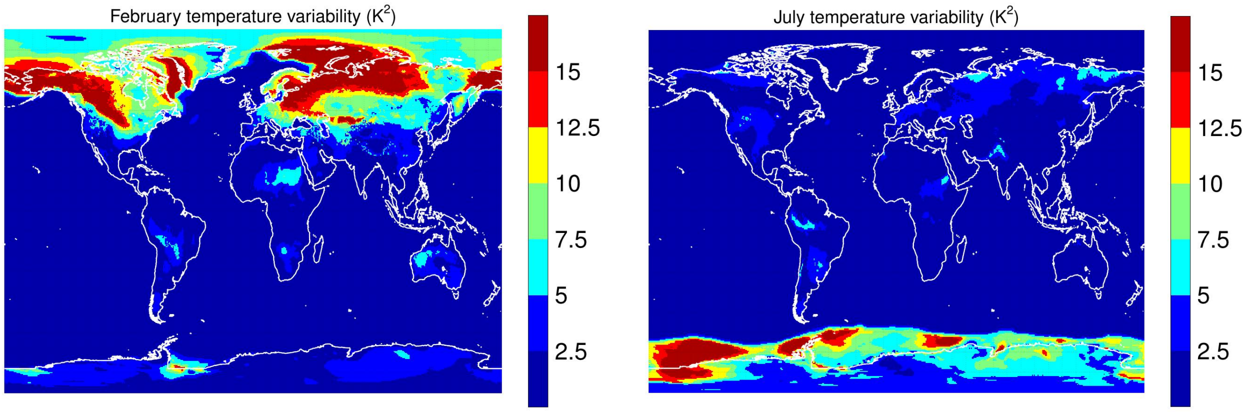

February temperature variability (K^2)
July temperature variability (K^2)
15
12.5
10
7.5
5
2.5

Figure 4.1: *Interannual temperature variability (expressed as variance) for (a) February and (b) July calculated for the MERRA dataset for the period 1979-2008.*

A direct calculation of the CMIP3 climate sensitivity values is computationally expensive because the climate system has to reach a thermal equilibrium. Rather than simulating a fully coupled transient evolution of the ocean, which takes several centuries to equilibrate with the surface forcing, a more efficient method is to use an atmospheric GCM coupled to a slab ocean, yielding a slab ocean equilibrium climate sensitivity. Another method uses the regression of the radiative flux at the top of atmosphere against the global average surface temperature change (Gregory et al., 2004) and produces an ”effective climate sensitivity”. To maximize the number of models, the climate sensitivity definition used here for the CMIP3 ensemble is the mean value of the two methods (or one of them if only one is available). The results do not depend on this choice.

In order to further investigate climate sensitivity uncertainty, about 5000 CPDN control simulations are used. The CPDN ensemble consists of several thousand simulations that were designed to explore parametric uncertainty (Stainforth et al., 2005). More than 50 parameters of the HadCM3L fully coupled model were perturbed in a range determined plausible by experts. Monthly temperature averages in 32 selected different regions were analyzed. The regions consist of the usual regions defined by Giorgi and Francisco (2000) plus additional regions covering a large portion of the oceans and the remaining land surface. Full maps of temperature at monthly resolution were not archived in the CPDN experiment. The CPDN equilibrium climate sensitivity values were estimated for 154 physics perturbations categories from a separate ensemble slab model experiments, i.e. using the atmosphere GCM coupled to a slab ocean.

4.3 Constraining climate sensitivity from observed interannual variability

4.3.1 CMIP3 ensemble

Natural variability is a crucial variable for detection and attribution studies of climate change (Wigley et al., 1998; Barnett et al., 2005) and climate model development is often

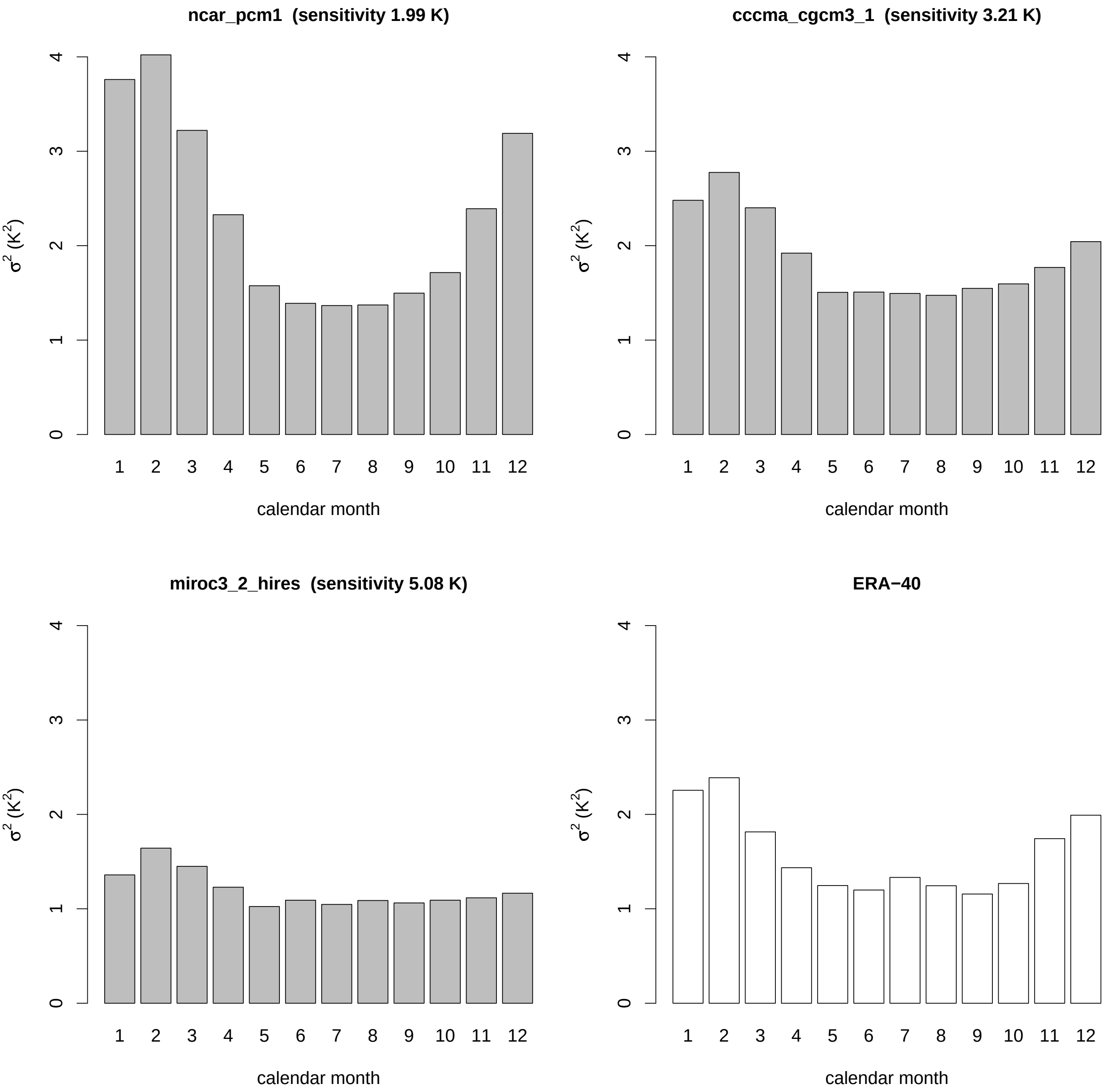


Figure 4.2: *The globally averaged variances σ^2 of the temperature anomalies for each calendar month. Three different CMIP3 models with low, medium and high climate sensitivity values (shown in parenthesis) and an observational dataset (ERA-40) are displayed for illustration.*

Table 4.1: *List of GCMs used, the length of their control simulations and climate sensitivity values. For the CMIP3 models, the climate sensitivity value is the average between "equilibrium climate sensitivity" and "effective climate sensitivity" values. See main text for details.*

Model	Length (yrs)	Climate sensitivity (K)
cccma_cgcm3_1	1001	3.21
cccma_cgcm3_1_t63	350	3.4
cnrm_cm3	500	2.45
csiro_mk3_0	380	2.65
gfdl_cm2_0	500	2.62
gfdl_cm2_1	500	2.84
giss_model_e_h	400	2.87
giss_model_e_r	500	2.63
iap_fgoals1_0_g	350	2.13
inmcm3_0	330	2.19
ipsl_cm4	320	4.11
miroc3_2_hires	100	5.08
miroc3_2_medres	500	3.96
miub_echo_g	341	3.10
mpi_echam5	506	3.63
mri_cgcm2_3_2a	350	3.08
ncar_ccsm3_0	230	2.53
ncar_pcm1	350	1.99
ukmo_hadcm3	341	3.18
ukmo_hadgem1	240	3.51
CPDN (HadCM3L)	160	0.13 - 9.27

focused as much on getting an adequate representation of variability as getting the mean state right. Wu and North (2003) have reported a relation between interannual variability and climate sensitivity in an earlier set of GCMs. The idea behind Wu and North (2003) relies on the cyclo-stationarity of interannual variability (Kim and Wu, 2000), i.e. the fact that in a control climate each individual calendar month has its own interannual variability that can be considered constant in spite of being different from the eleven other months. Monthly mean surface temperatures are considered first at the grid-point scale. The interannual variability is individually calculated at each grid-point and for each month. Fig. 4.1 illustrates those for February and July. Thereafter, the global average of monthly interannual variability is calculated for each CMIP3 model and the observational and reanalysis datasets. Fig. 4.2 shows examples of the globally averaged interannual variability for models with high, middle and low climate sensitivity as well as for the ERA-40 reference. Note that the globally averaged variability is higher during the boreal winter than during the boreal summer. The larger winter variability is due to different land mass distributions in the Northern and the Southern Hemispheres and is predominantly caused by snow cover (Kumar and Yang, 2003) and sea-ice variability. Apparently, the models with the highest and the lowest climate sensitivity do not match the shape of the reference dataset, leading to the hypothesis that the shape of the histograms could be used to evaluate some aspects of climate models. Wu and North (2003) proposed to quantify the asymmetry of the calendar month variances by the ratio σ_s^2/σ_w^2, where σ_s^2 is the global average of the smallest variance for the summer months June to August and σ_w^2 the global average of the largest variance for the winter months December to February.

One key result of our study is the linear relation between the ratio σ_s^2/σ_w^2 and climate sensitivity in the CMIP3 ensemble which is empirically found and shown in Fig. 4.3. The Pearson correlation coefficient is 0.78 and is statistically significant using a two-sided T test (p-value of $4.8 \cdot 10^{-5}$). Moreover, the probability for this correlation to be spurious is negligibly small, since Wu and North (2003) found approximately the same correlation (0.79) but with an older set of 13 GCMs. Using the linear relation above, it is possible to constrain climate sensitivity by the ratio σ_s^2/σ_w^2 derived from the observational references. Several sources of uncertainty have to be incorporated: first, the optimal linear regression is uncertain due to the scatter of the data points. Second, the real climate is always subject to changing external forcing (anthropogenic, volcanic and solar) which are not represented in the control simulations used in this study. Alternatively, the ratio σ_s^2/σ_w^2 derived from the CMIP3 transient simulation could be used instead of the control runs. However, this inflates the uncertainty because not all the models simulate the natural forcing. Also, the control simulations are longer and provide a more accurate estimate of σ_s^2/σ_w^2. Third, the observations are uncertain due to measurement uncertainty, model uncertainty in the reanalysis procedure, and due to sampling uncertainty caused by the relatively short time period for which observations are available. The 95% confidence interval of the linear regression is delimited by the two black dashed curves in Fig. 4.3. The uncertainty of the observational references ratio σ_s^2/σ_w^2 (marked by horizontal lines) results from external forcing and the limited time series available. In order to calculate the interannual variability of the calender months, the observational data is first detrended for each individual month in order to remove of the 20^{th} century transient warming. For each dataset the unforced natural variability is approximated by bootstrapping 30 individual years (20 for MERRA) from the total number of years. Since MERRA only has 29 years of data, a 20 years observation period is chosen instead. The interannual variability is then calculated for the twelve calendar months at each grid-point. This is repeated 1000 times to derive the 95% confidence interval for the ratio σ_s^2/σ_w^2.

The ERA-40 and MERRA reanalysis reference dataset shows ratios σ_s^2/σ_w^2 of similar magnitude and widths. The HadCRUT3 reference uses geographically scattered instrumental data which contributes to increase the uncertainty of σ_s^2/σ_w^2. The marine and land data have to be blended together and the station data are interpolated to a common grid. This homogenization artificially brings more variability over grid boxes with fewer observations (Brohan et al., 2006). We speculate that because fewer observations exist over oceans, the winter variability is overestimated and the HadCRUT3 ratio σ_s^2/σ_w^2 is shifted towards smaller values. The σ_s^2/σ_w^2 confidence interval based on the NCEP-NCAR reanalysis is larger and higher than for the other reference datasets.

The final 95% confidence interval for climate sensitivity combines the regression uncertainty with the reference uncertainty by projecting the observation 95% confidence interval on the linear regression 95% confidence interval. The result is given in Table 4.2. All reference

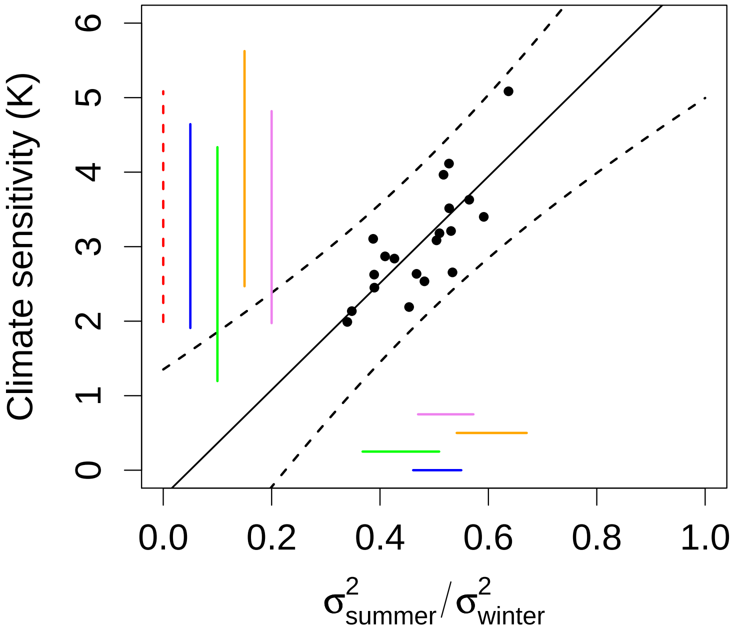


Figure 4.3: *Scatter plot of CMIP3 climate sensitivity versus the ratio of summer to winter interannual variability σ_s^2/σ_w^2. The correlation is 0.78, the dashed lines represent the 95% prediction confidence interval for the linear regression. The horizontal colored lines correspond to the 95% confidence interval calculated for the references, ERA-40 (blue), HadCRUT3 (green), NCEP-NCAR (orange) and MERRA (purple). The vertical lines on the left side summarize the results: the dashed red line is the original CMIP3 minimum to maximum range, the other solid colored lines are the 95% confidence interval for the climate sensitivity constrained by the corresponding reference dataset. The climate sensitivity uncertainty range takes into account both the regression and observational uncertainty.*

dataset are assumed to be equally likely, but the conclusions do not depend strongly on this assumption. The best estimate climate sensitivity value averaged over the 4 references is 3.3 K and the corresponding 95% confidence interval is 1.9-4.9 K, which closely corresponds to the ”likely” range of 2-4.5K given in the recent IPCC report (IPCC, 2007, Box 10.2).

Table 4.2: *Constraint on climate sensitivity from observed interannual variability ratio σ_s^2/σ_w^2, using the linear relation within the CMIP3 ensemble shown in Fig. 4.3. The columns are the length in years of the observational datasets, the best estimate of the predicted climate sensitivity and the 95% confidence interval.*

Dataset	Length (yrs)	Climate sensitivity	95% C.I.
ERA-40	44	3.22	1.91 - 4.64
HadCRUT3	159	2.75	1.20 - 4.33
NCEP-NCAR	61	3.96	2.47 - 5.62
MERRA	29	3.40	1.97 - 4.82
Combined	—	3.33	1.89 - 4.86

4.3.2 CPDN ensemble

The same analysis is repeated for the CPDN ensemble, and the ratio σ_s^2/σ_w^2 is computed for about 5000 CPDN control simulations. Because of this large amount of simulations, one would expect to see a clearer picture of a linear correlation between σ_s^2/σ_w^2 and the CPDN climate sensitivity values. Note that the CPDN simulation output is not given at the full grid-point resolution but is regionally aggregated and 32 regions have been chosen in order to cover most of the globe surface. As a consequence, the global average of the interannual variability is necessary less accurate than if computed directly from the grid-point scale but this should still provide useful information. The small grey dots in Fig. 4.4 show climate sensitivity versus σ_s^2/σ_w^2 for CPDN. In order to make a comparison between CMIP3 and CPDN possible, the CMIP3 ensemble was in turn regionally aggregated for the same regions to ensure consistency and is represented by black circles. The global average of the regional ratios σ_s^2/σ_w^2 is shifted towards smaller values because part of the grid-point information is lost, but the correlation still exists for the CMIP3 models (Pearson correlation of 0.68 and a p-value equal to 0.001 using a two-sided T-test). Surprisingly, no linear correlation exists for the CPDN ensemble.

This result might seem disturbing at first, but it does not imply that no relation exists. In order to seek for a possible relation, a pattern recognition algorithm is used to predict the CPDN climate sensitivity values with the regional ratios σ_s^2/σ_w^2 as input. The algorithm is called ”random forest” and is based on multiple regression trees (Breiman et al., 1984). The random forest is trained to match climate sensitivity values with 60% of the data and the remaining 40% of the data is used as validation set. The true climate sensitivity versus the sensitivity predicted from the regional σ_s^2/σ_w^2 ratios is shown in Fig. 4.5 and shows that the fitting explains more than 81% of the climate sensitivity with a root mean square (RMS) prediction error of 0.92 K. For comparison, the fit of the climate sensitivity values by Knutti et al. (2006) using regional seasonal cycles and a neural network technique explains 77% of the true climate sensitivity variance with a RMS prediction error between 0.5 K and 1.5 K for small to high sensitivities, respectively.

It appears that a relation exists in CPDN between climate sensitivity and the ratio σ_s^2/σ_w^2, even if this relation is complex and non-linear. Next, the CMIP3 climate sensitivity values are predicted using the random forest trained on CPDN. The results are shown in Fig. 4.5 by black circles and a red dot for the unperturbed CMIP3 HadCM3 model sharing the same structure as the CPDN simulations. The random forest predicts the HadCM3 climate sensitivity successfully but fails for four out of twenty CMIP3 climate sensitivity values that lie outside one CPDN standard error deviation, suggesting that these models are structurally too different from the HadCM3L model.

The real world climate sensitivity can also be predicted using the CPDN random forest and the observational references. The results are given in Table 4.3. The climate sensitivity 95% confidence interval averaged over the four reference datasets is 0.81-8.22 K. In contrast to the CMIP3 case, the confidence interval is larger and does not exclude low or high climate sensitivity values. The best estimate for climate sensitivity obtained for different observational dataset is between 3.27 K and 3.46 K, in close agreement with the results in section 4.3.1.

At least two interesting questions remain: first, why is the relationship between climate sensitivity and the ratio σ_s^2/σ_w^2 complex and non-linear while a simple linear regression is

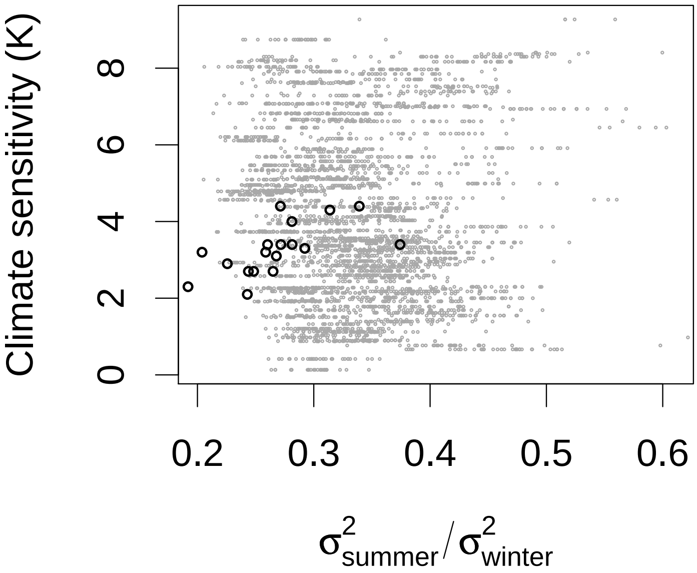


Figure 4.4: *Scatter plot of climate sensitivity versus the globally averaged regional ratios σ_s^2/σ_w^2. The grey dots represent about 5000 CPDN perturbed simulations and do not show any correlation, in contrast to the CMIP3 ensemble (black circles) which still shows a significant correlation of 0.68.*

sufficient in CMIP3? And second, why is the confidence interval derived from CPDN much larger in CPDN compared to CMIP3? Our hypothesis is that the tuning of models plays a major role and this point is discussed in the next section.

Table 4.3: *Constraint on climate sensitivity (best estimate and 95% confidence interval) from regional interannual variability ratio σ_s^2/σ_w^2, using the random forest fit on the CPDN ensemble shown in Fig. 4.5.*

Dataset	Climate Sensitivity	95% C.I.
ERA-40	3.37	0.88 - 7.60
HadCRUT3	3.27	0.98 - 8.74
NCEP-NCAR	3.46	0.98 - 8.17
MERRA	3.37	0.41 - 8.37
Combined	3.37	0.81 - 8.22

4.4 Origin of the linear relation and tuning

The reason why the asymmetry of the calendar month interannual variability (expressed as σ_s^2/σ_w^2) should correlate with climate sensitivity is not evident. If the globe had a symmetric

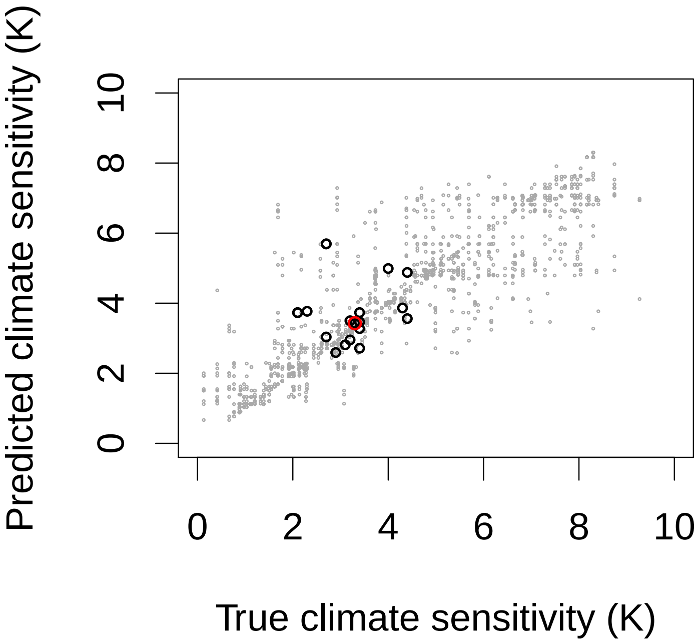


Figure 4.5: *Climate sensitivity predicted with a random forest algorithm using regional variability ratios* σ_s^2/σ_w^2 *as input vs true sensitivity. The scatter plot represents a subset of 40% CPDN data not used during the training phase (approx. 2000 simulations, grey dots). The explained variance is 81%. The black circles represent the CMIP3 models, the red circle is the CMIP3 HadCM3 unperturbed model.*

land-mass distribution from the equator, one would expect the ratio σ_s^2/σ_w^2 to be one and no relation with climate sensitivity should appear. As the dominant signal in temperature interannual variability is stronger over regions with snow cover and sea-ice, the representation of the albedo feedback in the models may explain the relation to climate sensitivity. This is consistent with recent work that demonstrates a relation between the seasonal and long term feedbacks related to snow albedo (Hall and Qu, 2006). It is also consistent with a correlation of -0.65 measured between the albedo feedback estimated by Soden and Held (2006) for 12 CMIP3 models and the ratio σ_s^2/σ_w^2 over the Northern Hemisphere. One may argue that the albedo feedback contribution and its spread across models is not the largest contribution to the magnitude and uncertainty of the total feedback (Soden and Held, 2006). On the other hand, Huybers (2010) has recently demonstrated correlations of different feedbacks across models, implying that the albedo feedback in a particular model could in fact explain more of the total feedback because it is linked to other feedbacks through tuning. In other words, a particular value of the albedo feedback favors certain values or ranges for other more important feedbacks. Wu and North (2003) propose a conceptual model of an energy balance with noise and damping as an alternative explanation of the relationship. However, further investigation is needed to explain the linear relation found in the CMIP3 ensemble. For the moment, it is important to keep in mind the fundamental difference between the CPDN and the CMIP3 ensembles. Although the CPDN parameters are perturbed within a realistic range set by experts, their combinations are not necessary plausible and some simulations will therefore significantly deviate from a realistic representation of the climate system. In contrast, the CMIP3 ensemble collects the best GCM solutions provided by various institutions. The performances of these models are often gauged by their ability to reproduce observations. It is thus likely that the CMIP3 models are tuned, whether unconsciously or not, to produce a plausible realization of the climate system. Our hypothesis is that this tuning simplifies the complex pattern found by the random forest to an approximately linear relation. In order to test this idea, we perform a simple idealized tuning procedure with the CPDN ensemble.

The CPDN ensemble is useful to test the impact of tuning because of its large ensemble size and the access to the parameter values of each simulation. From the CPDN perspective, ”tuning” is the selection of simulations according to a certain quality criterion. It is assumed that the design of the CPDN experiment did not try to reproduce the observations and the complete ensemble represents an essentially un-tuned state. This is not entirely true, since the CPDN HadCM3L ensemble is already a subset of earlier HadAM3 simulations that performed reasonably well, but the range of responses covered by the CPDN ensemble is still very broad, and no strong observational constraints were explicitly placed on any set of parameters.

Two cases of tuning can easily be tested: the regional mean temperature as well as the regional seasonal cycle compared to a reference dataset. A subset of 50 simulations are selected to build a subset of tuned simulations. The sample size is a subjective choice, but is a compromise between enough models to get a robust result and a restrictive criteria. If a subset of 50 simulations is selected for which the mean state of regional temperature mean state is closest (determined by the root mean square error) to the CMIP3 HadCM3 model, a linear relation emerges from the CPDN data cloud with a correlation of -0.79 (not shown). If the regional seasonal cycle defined as the boreal summer (June to August) minus winter (December to February) temperature is chosen as the tuning criterion the correlation is -0.76 (not shown) and if models are tuned for both regional mean state and regional seasonal cycle together, the correlation is -0.79. The latter subset is shown in Fig. 4.6 by the red dots and the red 95% prediction confidence interval. Tuning to observations results in lower correlations due to the additional structural difference between the observations and all CPDN models.

Based on a simplified energy balance model, Wu and North (2003) have shown how to get close to the observed ratio σ_s^2/σ_w^2 by tuning climate sensitivity. In GCMs however, tuning σ_s^2/σ_w^2 is hard to achieve because it implies that the complete statistic of regional seasonal interannual variability is correctly simulated. On the other hand, getting a reasonable seasonal cycle value for the GCMs is more feasible and is in fact considered as a necessary condition (Covey et al., 2000) for confidence in their projection. In addition, Covey et al. (2000) found a correlation of 0.4 between the amplitude of the seasonal cycle values in global mean surface temperature and climate sensitivity in an earlier set of GCMs. Later on, Wu et al. (2008) repeated this analysis for the CMIP3 ensemble and have found a correlation 0.63. Comparing the two results, it is evident that not only have the correlation improved, but also the range of the seasonal cycle amplitudes have narrowed. Because the seasonal cycle, the ratio σ_s^2/σ_w^2 and climate sensitivity are not independent and largely determined by the same feedbacks, tuning the models according to the seasonal cycle will also indirectly tune σ_s^2/σ_w^2. The reason for a negative slope in CPDN as opposed to a positive slope in CMIP3 remains unclear at this point. In summary, we find that an observational constraint can introduce a correlation, consistent with recent works (Huybers, 2010; Kiehl, 2007; Knutti, 2008b).

4.5 Discussion and conclusion

We have empirically demonstrated a near linear relationship between climate sensitivity and the asymmetry of the interannual variance in winter and summer surface temperature in the CMIP3 climate model ensemble. Even though the underlying mechanism is not fully understood, the correlation is very unlikely to be spurious, as it is very similar to the one found in a previous generation of models. The observations constrain climate sensitivity to $1.9 - 4.9$ K ($5 - 95\%$), with a most likely value of 3.3 K, similar to the mean and range of the CMIP3 ensemble.

Interestingly however, no relationship between variability and climate sensitivity is seen in the perturbed physics ensemble of CPDN. But when a subset of model versions from CPDN is selected to match the observed regional temperature distribution or seasonal cycle, this introduces a structure and correlation between interannual temperature variability and climate sensitivity, although of opposite sign. Huybers (2010) has shown that by placing a constraint on the top of atmosphere energy budget, correlations between seemingly unrelated feedbacks may occur; each model has to satisfy the global energy balance, i.e. a constraint is placed on the sum of all feedbacks, but not necessarily on any individual feedback. Kiehl (2007) and Knutti (2008b) have shown that matching the 20th century warming introduces a similar correlation between the total radiative forcing and climate sensitivity across the models; models with a higher sensitivity compensate the strong warming by a strongly negative aerosol effect.

The question is whether such tuning is problematic and whether it results in a model spread that is too small. Tuning the simulations decreases the possible configurations of a model by

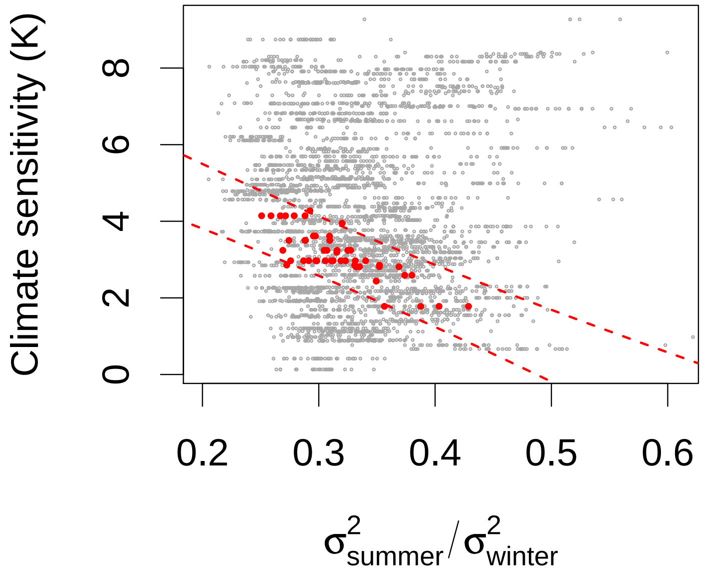


Figure 4.6: *Scatter plot of climate sensitivity versus the globally averaged regional ratios* σ_s^2/σ_w^2*. The grey dots represent about 5000 CPDN perturbed simulations. The red dots represent a subset of 50 CPDN simulations that are both close to the regional temperature mean state and the seasonal cycle of the unperturbed HadCM3 model as reference. The linear correlation is significant (-0.79). The dashed lines represent the 95% prediction confidence interval of a linear regression.*

eliminating parameter values that do not match observations, and therefore usually narrows the confidence interval. This can explain why the CMIP3 confidence interval based on observations is smaller than the original distribution of climate sensitivities in CPDN. Constraining the range of plausible models by observations could be considered as beneficial. But as a consequence, if the CMIP3 models are tuned to match the 20th century warming trend, the temperature mean state or seasonal cycle, of course these variables are no longer available as an observational constraint. This might explain why so few constraints on the CMIP3 climate sensitivity range are found, and why some constraints lead to posterior ranges similar to the original ones (Huber et al. 2010, in press). Interannual temperature variability as used here is difficult to tune and does show a relation to climate sensitivity, but since interannual variability, the mean state and the seasonal cycle are largely determined by the same processes, the range of climate sensitivities is again similar when applying observational constraints on interannual variability. The important question is thus not whether a particular dataset has been used for evaluation or tuning, but whether it provides additional independent information beyond what was already available beforehand.

One interpretation of the recent work on multimodel ensembles is that the CMIP ensemble might be considered as an approximate and ad hoc posterior distribution given observations on the climate mean, variability and trends; an ensemble however that was produced in an uncoordinated way, and of which we never knew the prior distribution nor the likelihood. The implications of such an interpretation are that one should be careful in using observations to weight models, since many observations were already used for model development and evaluation (Tebaldi and Knutti, 2007; Knutti et al., 2010b; Knutti, 2010). Poorly constrained weights are likely to do more harm than good (Weigel et al., 2010). Another implication is that such an ensemble is not only conditional on the observations but also on the model structure; if most or all models are missing or misrepresenting certain processes due to structural deficiencies, the tuning might erroneously lead to other problems, e.g. other feedbacks compensating for what is missing in a model. Such issues may not necessarily be apparent in a simulation of the present day climate, but could be important in the future. Understanding and quantifying uncertainty in a multimodel ensembles therefore remains challenging.

Acknowledgments

We acknowledge the international modeling groups for providing their data for analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the model data, the WCRP/CLIVAR Working Group on Coupled Models (WGCM), and their Coupled Model Intercomparison Project (CMIP) and Climate Simulation Panels for organizing the model data analysis activity, and the IPCC WG1 TSU for technical support. The CMIP3 Multi-Model Data Archive at Lawrence Livermore National Laboratory is supported by the Office of Science, U.S. Department of Energy.

We also thank all the individuals who have contributed their computer time for CPDN and Oxford University for providing the data.

Chapter 5

Climate model genealogy

Climate model genealogy

David Masson and Reto Knutti

Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland

(published in Geophysical Research Letters)

Abstract

Climate change projections are often given as equally weighted averages across ensembles of climate models, despite the fact that the sampling of the underlying ensembles is unclear. We show that a hierarchical clustering of a metric of spatial and temporal variations of either surface temperature or precipitation in control simulations can capture many model relationships across different ensembles. Strong similarities are seen between models developed at the same institution, between models sharing versions of the same atmospheric component, and between successive versions of the same model. A perturbed parameter ensemble of a model appears separate from other structurally different models. The results provide insight into intermodel relationships, into how models evolve through successive generations, and suggest that assuming model independence in such ensembles of opportunity is not justified.

5.1 Introduction

Uncertainty in climate model projections is often characterized by some measure of spread across an ensemble of simulations (Furrer et al., 2007; Tebaldi et al., 2005; Tebaldi and Knutti, 2007). The results thus depend on the range of responses covered by the models, and the distribution of the models within that range. The most recent coordinated ensemble used here is from the World Climate Research Project (WCRP) Coupled Model Intercomparison Project Phase 3 (CMIP3, Meehl et al., 2007a). A common assumption in such ensembles is that the set of models reflects the uncertainty in how to best describe the climate system in a model, arising partly from the difficulty in defining a unique model quality metric (Parker, 2006; Tebaldi and Knutti, 2007; Knutti, 2008a; Knutti et al., 2010b,a). Whether models span the full uncertainty range is hard to verify or falsify. The other assumption is that the models can be considered as independent in the sense that every model contributes additional information (Annan, 2010a). All models of course contain common elements (e.g. the equations of motion) because they describe the same system, and they produce similar results. But if they make the same simplifications in parameterizing unresolved processes, use numerical schemes with similar problems, or even share components or parts thereof (e.g. a land surface model), then their deviations from the true system or other models will be similar. In the extreme case, a model run at two resolutions, or the same model run twice with two initial states provide very little additional information about climatology or a decadally averaged projection. We qualitatively define an additional model as dependent if it provides little insight into why and how models differ from each other in the existing ensemble, and from observations. While statistically convenient, the assumption of independence is unlikely to be fully justified. Successful concepts in models are often copied or inherited, some institutions have used whole components from other models. Models are evaluated against the same observations, often using similar metrics. In CMIP3 several modeling groups have submitted two or three models. Most intercomparisons are thus ensembles of opportunity in which the sampling and dependence in the model space is unknown.

5.2 Method

The metric used to quantify the distance between unforced control simulations of two models is based on the Kullback-Leibler divergence that takes into account the full spatial field of monthly values in a control simulation. It thus considers the mean state, the seasonal cycle, the interannual variations, as well as the spatial correlation. A hierarchical clustering applied to the distance matrix of pairwise model dissimilarities and produces a 'family tree' of the models. The position at which the tree connects two models (relative to zero) characterizes the disagreement between the simulated control climate of two models, the vertical ordering of the branches is arbitrary. In addition, three reanalysis datasets (ERA, NCEP and MERRA) and two precipitation datasets (GPCP, CMAP) are treated like additional models. The details of the statistical method as well as the CMIP3 and the reanalysis and observation datasets are described in the supplementary material. Note that in contrast to earlier work (Jun et al., 2008a; Knutti et al., 2010b; Pennell and Reichler, 2010) this method does not analyze pairwise correlation of model errors to observations, but simply the similarity of two models as expressed by the simlilarity of their temperature and precipitation fields (see also Annan, 2010a). Observations are included here as 'additional models' just for illustration. The method and all conclusions are independent of whether the ensemble is interpreted as models being centered around truth or models and truth being indistinguishable, because the method only uses pairwise distances between models (Knutti et al., 2010a; Annan, J. D. and Hargreaves, J. C., 2010; Annan, 2010a).

5.3 Results and discussion

The tree in Fig. 5.1 shows that models from the same institution in almost all cases are very similar (e.g. GISS, MIROC, CCCMA, GFDL, UKMO, CSIRO). The degree of similarity varies and is more pronounced for example for GFDL than for GISS. Some of these pairs are not surprising, e.g. the two CCCMA models only differ in resolution. Others like the two UKMO for temperature are more surprising. Some characteristics seem to be preserved that keep these two models close, despite significant changes that were made to most components of the model. But relationships go beyond the "same modeling center" attribute. MIUB-ECHO-G and INGV-ECHAM4 (both based on an ECHAM4 atmosphere) cluster for temperature, INGV-ECHAM4 and MPI-ECHAM5 (both ECHAM based but with different versions) cluster for precipitation. A less evident pair, BCCR-BCM2_0 and CNRM-CM3, is identified for both temperature and precipitation. These models share the same atmosphere and land components.

Pennell and Reichler (2010) performed a similar analysis using hierarchical clustering but with a different distance metric based on model biases and 35 climate variables. While their results for CMIP3 are similar to those presented here, we show that a single variable (thus avoiding normalization) is sufficient to reveal most of the dependency structure, and that the key elements of dependence are similar for both surface temperature and precipitation. Observation and reanalysis datasets are not needed for the analysis, but when included like additional models they also cluster together, with some distance to the models, but well within the bulk of the simulations.

The picture gets even more interesting when the QUMP perturbed physics ensemble (Collins et al., 2010) and the previous generation of models in CMIP2 is included, shown in Fig. 5.2. The CMIP2 and CMIP3 models from the same institution also tend to cluster. For precipitation for example, the old NCAR CSM, PCM1 and the NCAR_WM models are close. The newest NCAR CCSM3 in CMIP3 however was developed almost independently from earlier NCAR models and appears separated. Qualitatively, the history can be traced back further for most models (Edwards, 2010). But given the rapid development, the increase in resolution in the models, the inclusion of new processes and the availability of more observations, we believe the connections between successive model versions are unlikely to persist over more than one or two generations.

In most of the trees, there is no clear separation into two or three clusters that are far apart, i.e. there is no evidence for multiple classes of models, different mutually exclusive theories or philosophies in how to build a model, or a clear separation between CMIP2 and CMIP3. The climate model landscape rather resembles an evolutionary process. Individual models take small steps compared to the size of the model space, successful pieces of a model are kept, inherited and copied and less successful parts go extinct. Existing models adapt to new environments (computer architecture and capacity, new observations, improved understanding of the climate system), although by deliberate rather than random modifications. New models rarely are written from scratch but evolve from combining, modifying and improving existing parts and ideas.

The perturbed versions of the HadCM3 (Collins et al., 2010) model separate themselves from the rest of the CMIP models. For some aspects, a large PPE can span a ”model space” similar or larger than CMIP3, e.g. for the range of feedbacks and climate sensitivity (Sanderson et al., 2010; Collins et al., 2010; Stainforth et al., 2005). However, if the full spatiotemporal fields are considered, the underlying model structure (grid, numerical scheme, parameterizations, resolved processes) appears to be important. Note that parameter perturbations in the QUMP ensemble are chosen to maximize the spread in feedbacks but ensure good agreement with climatology for each member (see supplementary material). Very different unconstrained model versions are likely to exist, and those may well fall outside the QUMP cluster.

5.4 Conclusions

Our analysis of spatial and temporal variations in surface temperature and precipitation shows strong similarities in groups of two or three models, supporting earlier claims that the effective number of independent models is smaller than the actual number of models in the

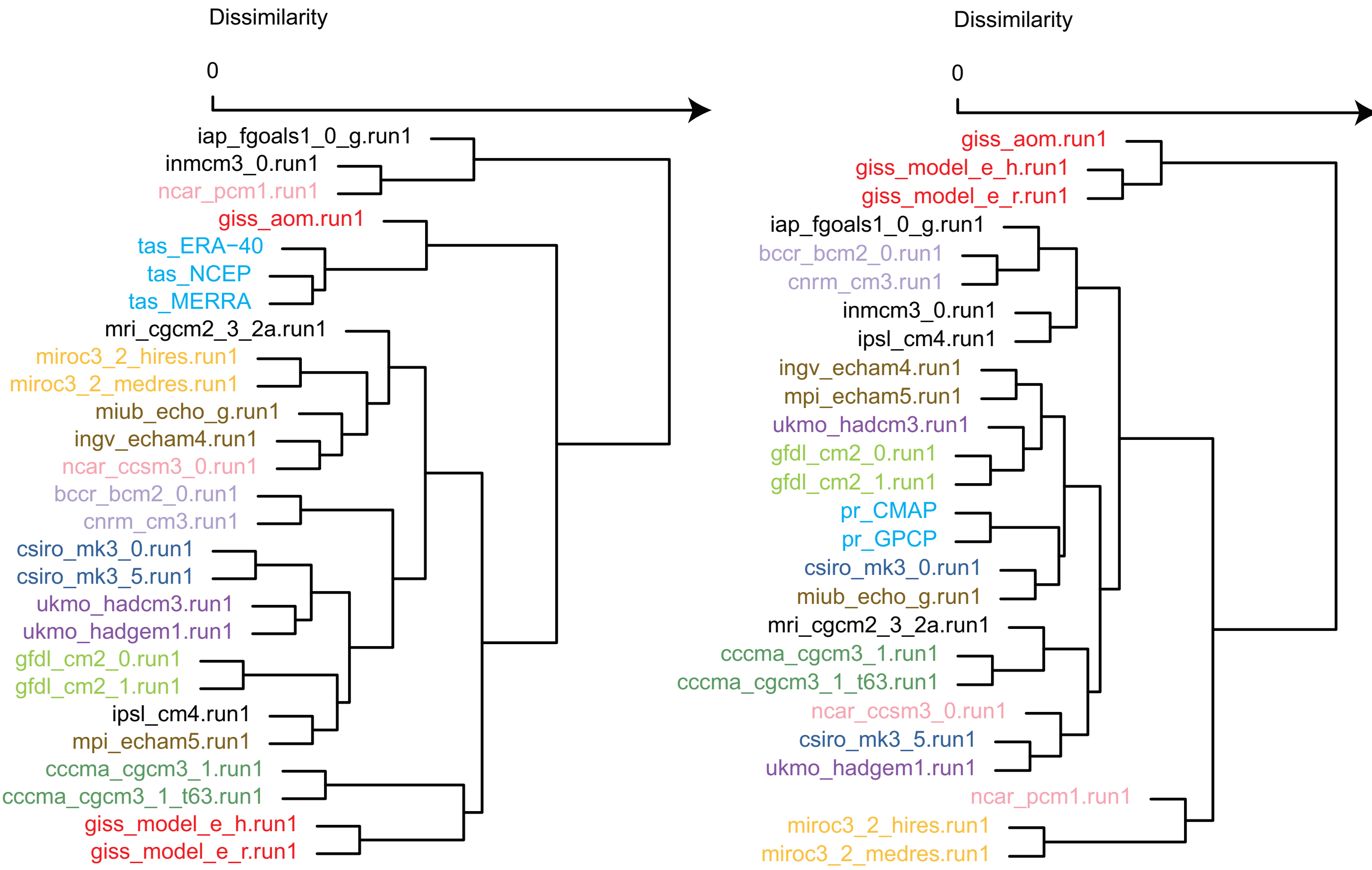


Figure 5.1: *Hierarchical clustering of the CMIP3 models for surface temperature (left) and precipitation (right) in the model control state. Models from the same institution and models sharing versions of the same atmospheric model are shown in the same color. Observations also are marked by the same color. Models without obvious relationships are shown in black.*

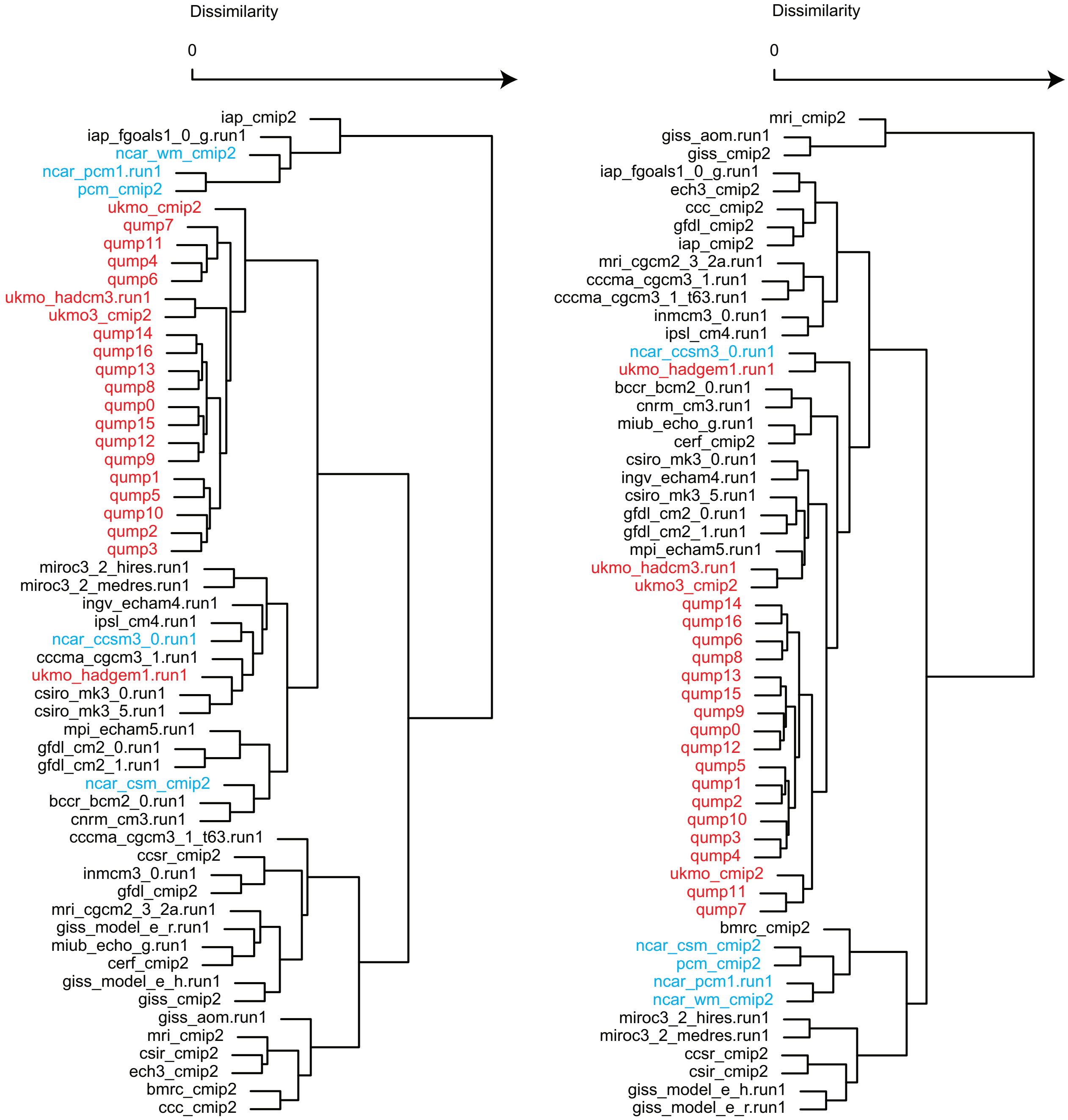


Figure 5.2: *Hierarchical clustering of the CMIP3, CMIP2 and QUMP perturbed physics ensemble for surface temperature (left) and precipitation (right) in the model control state. Models developed by the UK Metoffice Hadley Centre are shown in red, models developed by NCAR are marked in blue.*

multi model ensemble (Pirtle et al., 2010; Pennell and Reichler, 2010; Jun et al., 2008b; Tebaldi and Knutti, 2007; Knutti, 2008a; Knutti et al., 2010b; Knutti, 2010; Annan, 2010a). Models developed at the same institution show the most striking similarities, but dependencies are even visible between two models that use different versions of the same atmosphere. Ensembles of different generations as well as observations largely overlap, suggesting a gradual evolutionary development and refinement of models. For the metric and variables chosen here, structural model differences seem to be important. We interpret this as an indication that sampling different model structures is important to capture the full range of model behavior.

Correlations between the control state and the projected change across models are generally weak (Knutti et al., 2010b), implying that a mapping of the dependence structure into projections is difficult (see supplementary material for a discussion of clustering projections), i.e. two models that have similar biases in the present may not have similar projection errors in the future. It is therefore unclear whether the dependence in the control state implies that uncertainty in projections is underestimated, or that the number of effective models in the future is the same as in the present.

Using equal weights for all models to create a most likely projection fails to take into account model dependencies. If a group of similar models is part of the ensemble, either from small changes in parameters or resolution, this poses a risk of double counting and giving undue weight to the structure underlying the group. Given the large cost of model development, the availability of open community models and the broader availability of supercomputers, model variations of existing models and perturbed physics ensembles (Sanderson et al., 2010; Collins et al., 2010) may become more common in the future, making this dependence a much bigger issue.

The goal for an ensemble should be to maximize diversity in models yet ensure good performance for all members, and minimize dependency. In principle, this could be achieved with a sufficiently large and broad ensemble to start with, and an appropriate weighting that takes into account two distinct factors: performance metrics measuring model skill, and the dependence to other models to account for sampling problems demonstrated above. In practice, this proves to be very difficult. For the former, the obvious problem is the lack of repeated verification to define skill for the forecast quantity of interest (Knutti et al., 2010b; Tebaldi and Knutti, 2007; Knutti et al., 2010a). Skill therefore has to be indirectly determined by relating the forecast to metrics based on past trends and climatology. With few exceptions (e.g., Stott et al., 2006) such relationships are weak (Knutti et al., 2010b), making the definition of model weights ambiguous. If weights are incorrectly specified, the forecast is likely to be worse than if no weighting was used, in particular for small ensembles (Weigel et al., 2010). Accounting for model performance may be possible in some cases, e.g. in the Arctic where several metrics of present day climate and past trends are clearly related to future warming and sea ice decline, and where the underlying processes are well understood (Boé et al., 2009b,a). If the identified model biases lead to biased predictions (Stroeve et al., 2007), it would seems stupid not to consider the observed evidence to improve projections and estimate uncertainties.

Even if the general formulation of an unambiguous weighting scheme for various regions, variables and timescales that takes into account model performance and dependence appears to be a long way off, a few conclusions are obvious. First, there is a lively debate in the community on the point of model weighting (Knutti, 2010), but the issue of sampling in ensembles has received very little attention. Second, diversity is critical. The number of structurally different models is small, and maintaining a sufficiently large set of reasonably independent models that span a wide range of plausible assumptions and scientific viewpoints is important both to quantify uncertainty and to understand model differences (Knutti, 2010). Eliminating a model from an analysis is easy, extrapolation beyond the range covered by the ensemble is nearly impossible. Third, models are rarely built with lasting value as a primary goal (Held, 2005) and are superseded by newer models. Yet to understand why models and their projections differ, archiving results from older model versions and common scenarios would help. Fourth, conclusions drawn from ensembles should at least test the sensitivity to how models are selected in the ensemble. Current coordinated model experiments are like asking the same question to a small number of people, without thinking about how to select those people, how many to ask, and how to account for the fact that they may have similarly biased opinions. This undoubtedly makes the interpretation of the answers challenging.

Acknowledgments

We thank Thomas Reichler for CMIP2 and Mat Collins for the Hadley Centre QUMP data. We acknowledge the modelling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy (DOE).

5.5 Supplementary material

5.6 Data

The data consists of simulated and observed monthly surface air temperature and precipitation fields from pre-industrial control experiment with no external forcing. Three sets of models are used. The first set is the ensemble ”Quantifying Uncertainty in Model Predictions” (QUMP) and was generated by perturbing the parameters of the HadCM3 climate model (Collins et al., 2010). Multiple parameters are perturbed simultaneously such that the ensemble covers a wide range of feedbacks, yet all models are required to be close to the climatological mean state as measured by the CPI score (Murphy et al., 2004; Webb et al., 2006). Thus the QUMP PPE is not a random perturbation of a base model, but can be interpreted as a PPE constrained by climatology, in much the same way as CMIP2/3 is designed (or tuned in some informal way) to agree reasonably well with climatology. If the requirement of agreeing with the observed mean state is relaxed, other model versions are likely to be possible that look very different from those used here. The second and third set belong to the phase 2 (CMIP2, Covey et al., 2003; Meehl et al., 2000) and phase 3 (CMIP3, Meehl et al., 2007a) of the World Climate Research Program (WCRP) Coupled Model Intercomparison, a coordinated project to gather and compare the different of GCMs used for the third and fourth IPCC assessment reports, respectively IPCC (2001) and IPCC (2007). The observational-based datasets for surface air temperature are the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis (ERA-40; Uppala et al., 2005), the National Centers for Environmental Prediction / National Center for Atmospheric Research reanalysis (NCEP-NCAR; Kalnay et al., 1996) and the Modern Era Retrospective-analysis for Research and Applications (MERRA; GMAO, 2009) reanalysis. For precipitation, the CPC Merged Analysis of Precipitation (CMAP; Xie and Arkin, 1997) and the Global Precipitation Climatology Project (GPCP; Adler et al., 2003) datasets are used. Because individual models and observations come at different spatial resolutions, the data is interpolated to a common T42 grid (Gaussian grid associated with spectral truncation, 128 latitudinal by 64 longitudinal grid-points) using bilinear interpolation.

5.7 Methods

The degree of dissimilarity between pairs of individual models and observational datasets is based on a multivariate description of natural variability. The method is inspired from Shukla et al. (2006). We choose the Kullback-Leibler divergence to quantify the dissimilarity between the two simulations. While simpler metrics such as root mean square error or pattern correlation are possible, those consider only mean states and the information contained in the variability is ignored. As natural variability is a crucial variable, model development is often focused as much on variability as on the mean state (Covey et al., 2000). We therefore expect similarities to become clearer when comparing models on the basis of the mean climate plus natural variability.

In this study, the seasonal cycle (four successive seasons starting from December-January-February) is defined at each grid-point as the seasonal mean value minus the average annual mean value of the entire simulation. The interannual variation in the seasonal cycle is defined at each grid-point as the seasonal mean value minus the climatological value of that season. To avoid biased estimates of variability, both the control simulations and the observational-based datasets are linearly detrended for each individual season. Trends in control simulations occur because of internal model drifts. The 20^{th} century transient climate change is removed from the observations by linear detrending to allow a comparison with the control simulations. The method thus considers the seasonal and interannual variations with their spatial structure, but not the global mean state and trends.

Each seasonal cycle is fitted with a multivariate normal distribution $\mathcal{N}(\vec{\mu}_i, \Sigma)$ where $\vec{\mu}_i$ is a vector of length N representing the average seasonal cycle over N grid-points. The index i stands for boreal winter, spring, summer or fall. The covariance structure between the grid-points expressed in the $N \times N$ covariance matrix Σ is calculated using all successive seasons $t = 1, \ldots T$ available in the model or observation-based dataset.

The dissimilarity between two multivariate normal distributions A and B is expressed by the Kullback-Leibler divergence $J(A, B)$:

$$\begin{aligned} J(A,B) &= \frac{1}{2}\mathrm{Tr}\left\{(\Sigma_A - \Sigma_B)(\Sigma_B^{-1} - \Sigma_A^{-1})\right\} \\ &\quad + \sum_{i=1}^{4} \frac{1}{2}\mathrm{Tr}\left\{(\Sigma_A^{-1} + \Sigma_B^{-1})(\vec{\mu}_{i,A} - \vec{\mu}_{i,B})(\vec{\mu}_{i,A} - \vec{\mu}_{i,B})^t\right\} \end{aligned} \tag{5.1}$$

This metric was first described in (Kullback, 1968) in the context of information theory and statistics. $J(A, B)$ is a positive and symmetric measure. However, it is not a distance in the mathematical sense since the triangle inequality condition is not fulfilled. It should also not be confused with the relative entropy used in Shukla et al. (2006) which is not a symmetric measure. The Kullback-Leibler divergence used here is the sum of two components, one part due to the difference in covariances and related to interannual variations, and the other due to the difference in the climatological seasonal cycle.

A difficulty arises when computing (5.1) because of the inverse covariance matrices Σ^{-1}. Since the number T of simulated seasons is generally smaller than the number of grid-points ($N = 8192$ in T42), the matrices Σ are often singular and do not have a unique inverse matrix. The solution is to reduce the spatial dimensionality (N) while preserving the maximum amount of information. We have chosen to project the data on a new coordinate system based on Empirical Orthogonal Functions (EOFs) of observed seasonal anomalies. While the number of simulated seasons is between 320 and 4004, the observational dataset have generally much shorter records. In order to get the best estimate of the observation-based covariance matrix, ERA-40, NCEP and MERRA are concatenated, resulting in 541 seasons. The same is done with CMAP and GPCP, resulting in 223 seasons for precipitation.

The new coordinate system uses 500 EOFs for temperature and 200 EOFs for precipitation and explains almost 100% of the original variance. This arbitrary choice is a compromise between retaining as much information as possible and avoiding singular matrices for the complete set of simulation and observations. According to North's Rule-of-Thumb (von Storch and Zwiers, 2004, section 13.5.5), about 20 EOFs should be sufficient to find the physically relevant patterns. While this is true for the interannual variations, the climatological seasonal cycle needs more basis vectors to be correctly reconstructed because the EOFs are based on the interannual variability. For example, the linear correlation between the original ERA-40 boreal autumn mean vector and its representation in the EOF basis is only 0.58 with 20 EOFs but 0.94 when 500 EOFs are used instead.

Next, a hierarchical clustering (see Hastie et al., 2001) is applied to map the structure of all data objects. The algorithm uses the dissimilarity matrix of pairwise data measured by the Kullback-Leibler divergence in the reduced space. The cluster analysis groups data objects such that pairwise dissimilarities between those data assigned to the same cluster tend to be smaller than those in different clusters. The hierarchical clustering is done separately for surface air temperature and precipitation. The position of a node in the tree on the horizontal axis measures the dissimilarity between its two daughters (which could itself be a single model or a cluster of models). The range of the dissimilarity metric depends on the number of EOFs and has no physical or climatological interpretation.

Note that the nearest neigbor in the tree can change if more models are added to the ensemble (i.e. going from Fig. 1 to Fig. 2), even if the distance of the models in the original ensemble does not change.

5.8 Robustness of results and additional discussion

The length of the control simulations varies for each model, but the results are insensitive to that. The same analysis with different segments of the control simulations shows very similar results; most of the relationships become evident already using a ten year segment of the control simulation for each model. The results are also robust to using the observational datasets separately rather than concatenating them to calculate the EOF basis. Similar results can in principle be obtained with other basis sets, e.g. spherical harmonics.

One may argue that the dissimilarity metric should combine multiple variables. However, it is unclear how to account for dependence between variables, and any combination of variables requires an arbitrary choice to normalize variables with different units. The results presented here show that many relationships are already obvious with a single variable and are consistent for surface temperature and precipitation.

Another obvious question is whether projected changes could be analyzed instead of control simulations. There are several difficulties in doing that. First, by considering anomalies from a base period, the information of the seasonal cycle is lost, which is important in identifying model relationships. Second, it is unclear whether the projected change patterns should be normalized to account for different climate sensitivities. Finally, if two models cluster it is not clear whether they behave similarly in terms of how they simulate processes or whether they just use the same particular sets of forcings. An analysis performed on the projected changes by the end of the twenty-first century relative to current climatology indicates that only a small fraction of the model relationships are preserved (not shown). This is also consistent with the general lack of correlation between the climatological mean state and the predicted changes (Knutti et al., 2010b). The control simulations are appealing because they really express how a model simulates feedbacks and processes, i.e. they tell us about the physical assumptions as well as the numerical schemes built into the model.

An obvious question is whether weights could be defined based on the hierarchical clustering. That appears difficult at this stage and is further complicated by the lack of correspondence of the similarity of the control states to the similarity in the projected changes, but this is certainly an important topic for further research.

Chapter 6

Model weighting, ensemble size and tuning: a conceptual study

Model weighting, ensemble size and tuning:
a conceptual study

David Masson and Reto Knutti

Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland

Abstract

Multi-model averages is a widespread technique to aggregate the results of an ensemble of simulations and has been often used to calculate the best estimate of future climate change. Because the general circulation models used in the IPCC AR4 report do not perform equally well, some studies have suggested that an optimum weighting would improve the accuracy of climate projections. We investigate this hypothesis with a conceptual experiment where the size of an ensemble and the calibration of the models can be controlled. The results are consistent with other studies on the risks of an inappropriate model weighting and we find that weighting is generally less beneficial than model calibration or increasing ensemble size.

6.1 Introduction

The Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections (Knutti et al., 2010a) emphasizes the limitations of combining multiple models for certain applications. Progress in this area is expected to depend on the variable, spatial and temporal scale of interest. In the case of temperature and precipitation change, models are sometimes treated equally as in Meehl et al. (2007b), or weighted according to their performance as in Giorgi and Mearns (2002) or Tebaldi et al. (2005).

Whetton et al. (2007) suggest to give higher weight to the best performing models because models close in simulating present-day patterns tend to remain close in the future (Meehl et al., 2007b). Räisänen et al. (2009) have studied this idea in details and show that weighting produces consistent results only when present-day performance is related to simulated future temperature change, a condition that is not fulfilled for temperature over most part of the globe. Some other studies emphasize the risk of an inappropriate weighting. For example, Weigel et al. (2010) use a conceptual model to show that more information may be lost by weighting than potentially gained. Equal-weighting appears therefore to be a safer solution for many applications.

Even giving the same weight to all the models is biased by the fact that models are not independent. Because several institutions have submitted more than one model to the Third Model Intercomparison Project (CMIP3 Meehl et al., 2007a) used in the Intergovernmental Panel on Climate Change Fourth Assessment Report (IPCC AR4 IPCC, 2007), some model concepts are implicitly given more importance. Another concern arises if models are tuned to match the observations: the information used to evaluate model performance should be as independent as possible of those used during the modeling calibration phase (Tebaldi and Knutti, 2007; Knutti, 2010; Knutti et al., 2010a,b).

This study investigates the impact of model weighting as a function of the ensemble size and model tuning to the consistency of global temperature projections. We use a cross-validation technique based on a conceptual model. The idea is whether two independent IPCC reports based on two model ensembles would agree, admitting that enough technical and financial resources were available. We expect the consistency between the two reports to be a necessary condition to trust the model projections.

Section 6.2 analyses the impact of weighting models in relation with the ensemble size and uses the CMIP3 models. Section 6.3 analyses the impact of weighting in relation with model calibration and uses artificial models generated by a toy-model. This study is rather a conceptual analysis than a systematic investigation of the CMIP3 model outputs. Nevertheless, some conclusions are likely to apply to existing climate ensembles.

6.2 Role of the ensemble size using CMIP3 models

This study uses a subset of the CMIP3 data. One ensemble member for each of the 24 Atmosphere Ocean General Circulation Models (AOGCMs) simulations under the A1B emission scenario (Nakicenovic et al., 2000) is used. The observation-based data sets is ERA-40 (Uppala et al., 2005) for surface temperature. Climatological averages over periods of 20 years (annual means) have been extracted from the original monthly averaged temperature fields. The evolution of the global temperature between 1900 and 2080 is shown in Fig. 6.1. The colored lines represent the CMIP3 simulations and the black line corresponds to the ERA-40 reanalysis. The first experiment stages two ensembles A and B made of CMIP3 models. We are interested in the consistency of the projected temperature increase $\Delta T = T_{2080} - T_{1980}$ between the ensemble A and B. Two effects are likely to change the degree of agreement between A and B: the ensemble size N and the model weighting λ inside each ensemble. The consistency is expected to increase with the number of models N because the variance of the sample mean decreases with $\sqrt{N}$ (see von Storch and Zwiers, 2004, section 4.3.1). However, the consistency between ensembles A and B is still unknown when the models are weighted.

The experiment is constructed as follows: N different CMIP3 models are randomly assigned to the group A and B. Every model m predicts an increase of global temperature change ΔT_m and is weighted according to its ability to reproduce the ERA-40 climatological mean between 1980 and 1999 (termed present-day here). The weighting for model m is proportional to $w_m(\lambda) = \exp(-\lambda \cdot E_m^2)$, where λ controls the weighting and E_m is the error of the model m in reproducing the present-day global mean temperature. The estimated future temperature change is assumed to be normally distributed according to $\mathcal{N}(\Delta\bar{T}, \sigma)$ centered on the multi-model weighted average $\Delta\bar{T}$ and with weighted variance σ^2 calculated as

$$\sigma^2 = \sum_{m=1}^{N} w_m(\Delta T_m - \Delta\bar{T})^2 \Bigg/ \sum_{m=1}^{N} w_m$$

where ΔT_m is the warming in model m. Fig. 6.2a shows the confidence range for the ensembles A and B as function of the weighting steepness λ for two 5-model ensembles whose members have been arbitrarily chosen. In this particular case, weighting according to present-day performance shifts the center of the distributions away from each other and does not result in better agreement between the two ensembles A and B. With increasing weighting, the width of the distributions gets smaller and the ensembles are more likely to become overconfident and to disagree with each other.

In the general case, the ensemble A is generated by choosing N models from the original CMIP3 ensemble of 24 models and the ensemble B by choosing N different models from the remaining $(24 - N)$ models. The disagreement between the two distributions $\mathcal{N}_A$ and $\mathcal{N}_B$ for different weighting is quantified with the Kullback-Leibler divergence. Because ΔT is assumed to be normally distributed, the divergence is given by

$$J(A, B) = \frac{1}{2}(\sigma_A - \sigma_B)(\sigma_B^{-1} - \sigma_A^{-1}) + \frac{1}{2}(\sigma_A^{-1} + \sigma_B^{-1})(\Delta\bar{T}_A - \Delta\bar{T}_B)^2 \tag{6.1}$$

(see Kullback, 1968). The Kullback-Leibler divergence is a symmetric measure of dissimilarity defined on the positive real axis. The closer $J(A, B)$ is to zero, the more similar are the distributions $\mathcal{N}_A$ and $\mathcal{N}_B$. 500 different pairs of ensembles are generated in order to provide many different combinations and increase robustness. The disagreement between the hypotheses A and B on the future warming is shown in Fig. 6.2b as function of the weighting and the number of models. This figure does not show values for weighting larger than $\lambda = 2$ because no overlap between the two distributions A and B produces large values (about 80) of Kullback-Leibler divergence. Two important features appear: first, larger samples increase the agreement between A and B. This is expected and follows the decrease with $\sqrt{N}$ of the sample mean variance. Second, model weighting is never profitable. This is a direct consequence of present-day performance being unrelated to the predicted change in the CMIP3 ensemble (see also Räisänen et al., 2009; Knutti et al., 2010b).

A strong weighting favors the models close to the observation. When the model sample is large, the best performing model in the ensemble A has a larger probability to lie next to those in the ensemble B. While weighting increases the confidence in the best performing models, it also reduces overlapping between the confidence intervals of A and B. The disagreement induced by weighting is more severe when the size of the ensembles is small. With two models in each sample for example, weighting is equivalent to selecting one of the two models with an increasing confidence, and the predictions provided by A and B are more likely to disagree.

6.3 Role of model calibration using artificial models

Weighting models does generally not improve the consistency of the CMIP3 global temperature increase, because the correlation ρ between present-day bias and future projection is weak (Tebaldi and Knutti, 2007; Räisänen et al., 2009). As the correlation ρ measured in the CMIP3 ensemble cannot be changed, artificial global temperature simulations are created to mimic the CMIP3 models while allowing the control over ρ by calibration. This section analyses how model consistency is affected by weighting when models have been tuned using the same variable as those used for model evaluation. Again, two ensembles A and B with N models are created, but here a simple toy model is used. The global temperature T is artificially

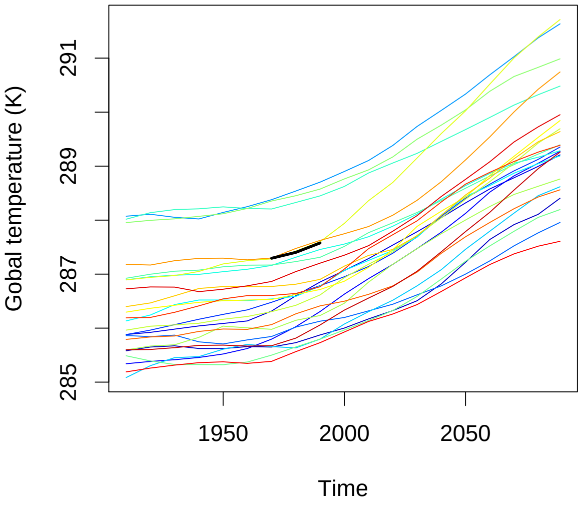


Figure 6.1: *Evolution of the global average temperature. Colored lines are CMIP3 model simulations and the black line represents the ERA-40 reanalysis.*

simulated by a linear function of time: $T(t) = T(t_0) + \mu \cdot t$. The boundary condition $T(t_0)$ is generated from a normal distribution $\mathcal{N}(\bar{T}_0, \sigma_0)$, with $\bar{T}_0$ the average CMIP3 simulated global temperature in the period $(1900 - 1919)$ and σ_0 corresponding to the CMIP3 standard deviation. The trend μ is generated from a Gamma distribution to avoid negative values implying a cooling. The Gamma distribution $\Gamma(\hat{\alpha}, \hat{\beta})$ is a function defined on the positive real axis and the parameters $\hat{\alpha} = \frac{\bar{\mu}^2}{\sigma^2}$ and $\hat{\beta} = \frac{\sigma^2}{\bar{\mu}}$ are estimated to fit the CMIP3 trends with mean $\bar{\mu}$ and standard deviation σ (see Wilks, 2006, section 4.4.3). Fig. 6.3 shows artificial simulations (colored lines) together with the ERA-40 reanalysis (black line). The model agreement can be tuned by arbitrarily shrinking or expanding the standard deviation σ measured in the CMIP3 ensemble. This action directly controls the correlation ρ between the present-day bias and future temperature increase. When the artificial simulations have the same initial conditions $\bar{T}_0$ but different trends, the relation is fully deterministic with a correlation equal to one. The initial uncertainty σ_0 in year t_0 is responsible to change the deterministic relation into stochastic behavior. When the relative importance of the future projection uncertainty $\sigma \cdot \Delta t$ is large compared to σ_0, the relation is more deterministic and the linear correlation approaches one.

In the same way as the previous experiment, two ensembles A and B are created. Each ensemble gets a fixed number of 12 artificial simulations. Fig. 6.4a shows one possible outcome whose correlation ρ is 0.8. In this particular case, a moderate weighting brings more agreement between the ensemble A and B. In order to get a general result, 10'000 ensemble pairs with each 12 artificial simulations have been generated with varying the tuning parameter σ between half to four times the original CMIP3 standard deviation of the 20^{th} century simulated trends. The disagreement is again quantified with the Kullback-Leibler divergence expression (6.1). The results are aggregated into several correlation bins from the smallest value to the largest. Finally, a detailed view in Fig. 6.4b shows the disagreement between the ensembles A and B as function of the weighting stress λ and the correlation ρ.

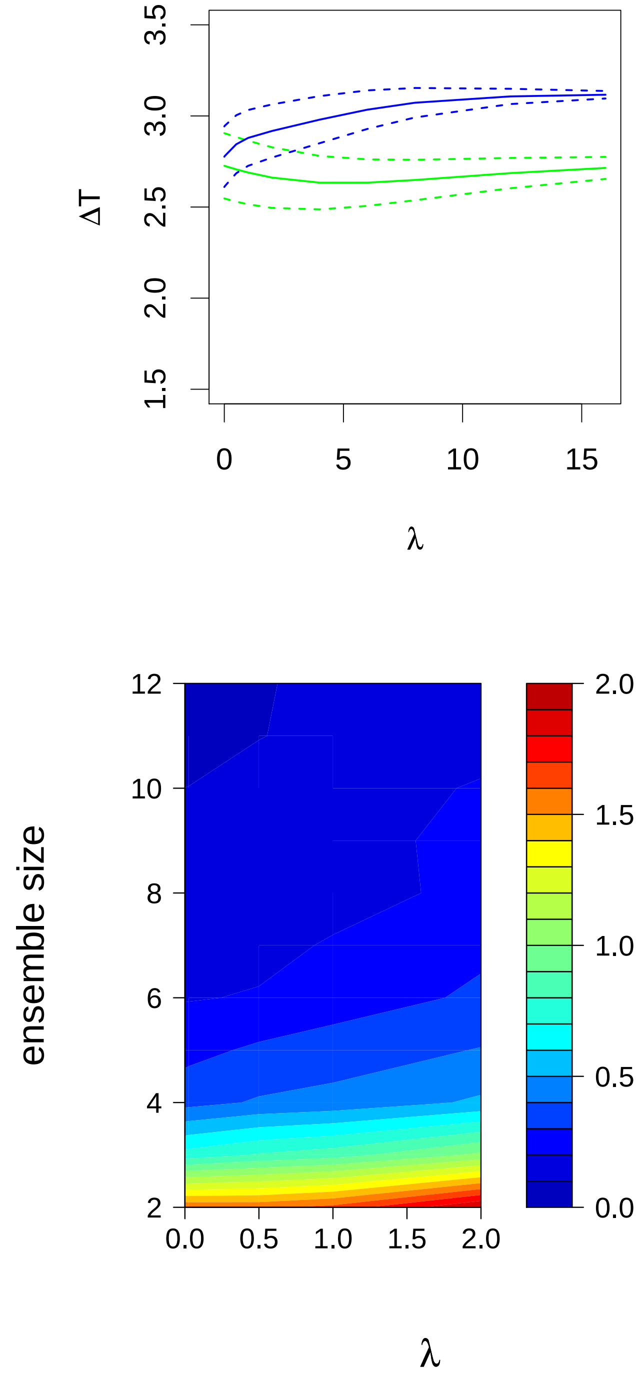


Figure 6.2: *(a) Projected temperature increase for two ensembles with 5 models as a function of the weighting λ. The solid lines are the ensemble mean values and the dashed lines represent one standard deviation. (b) Disagreement between the two ensembles as a function of the ensemble size and the weighing λ. The disagreement is measured with the Kullback-Leibler divergence and 500 model combinations from the 24 CMIP3 ensemble have been averaged to ensure robustness.*

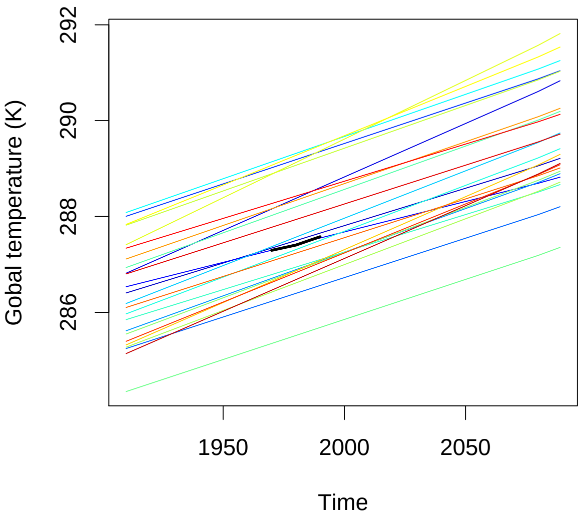
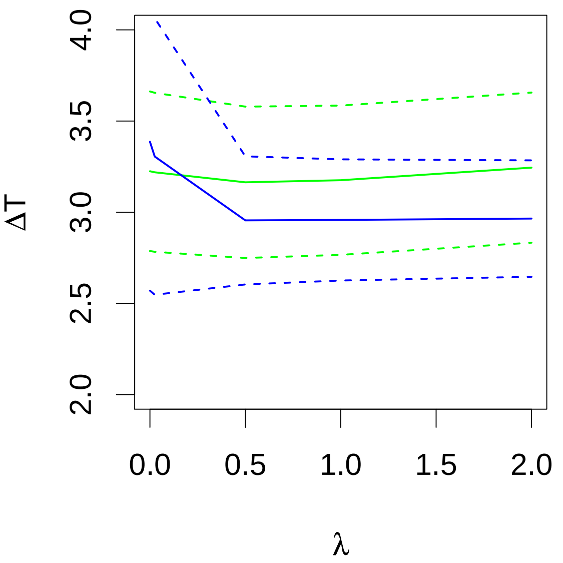


Figure 6.3: *Evolution of the global average temperature. Colored lines are artificial simulations, see description in text. The black line represents the ERA-40 reanalysis.*

The maximum model consistency is reached when the correlation is low and no weighting is applied. A low correlation is the result of tuning towards the warming consensus. It is therefore not surprising that the ensembles A and B agree when ρ is close to zero, consistent with Räisänen et al. (2009). Here, a higher correlation can be reached by relaxing the tuning constraint and allowing model trends to come from a wider distribution. By construction, the disagreement is larger simply because σ_A and σ_B are larger in the divergence expression (6.1). Consistent with Räisänen et al. (2009), a moderate weighting improves the agreement. A moderate weighting acts as a filter that eliminates outliers and inadequate simulations, while the bulk of the remaining models remains approximately equally weighted in comparison. This is visible in the particular case shown in Fig. 6.4a where the confidence interval of one of the two ensembles (in blue) is strongly reduced because an outlier is dismissed by weighting. The beneficial effect of weighting lasts as long as the distributions $\mathcal{N}_A$ and $\mathcal{N}_B$ of the future temperature increase do no get too overconfident and still overlap. This experiment suggests two possibilities to improve the consistence of multi-model projections: either the models are tuned towards a better agreement, or they get weighted according to their performance.

6.4 Discussion and conclusion

We have analyzed the impact of weighting models in multi-model projections as function of the ensemble size and tuning. The agreement between two ensembles of models is a consistency criteria and is a necessary condition for climate projection. It is nevertheless not sufficient to guarantee future skill since all models might share similar biases. The conceptual experiments done here suggest that a better model calibration and larger ensemble size improve consistency. In contrast, weighting models according to performance is useful only when it acts as a filter that eliminate deficient models.



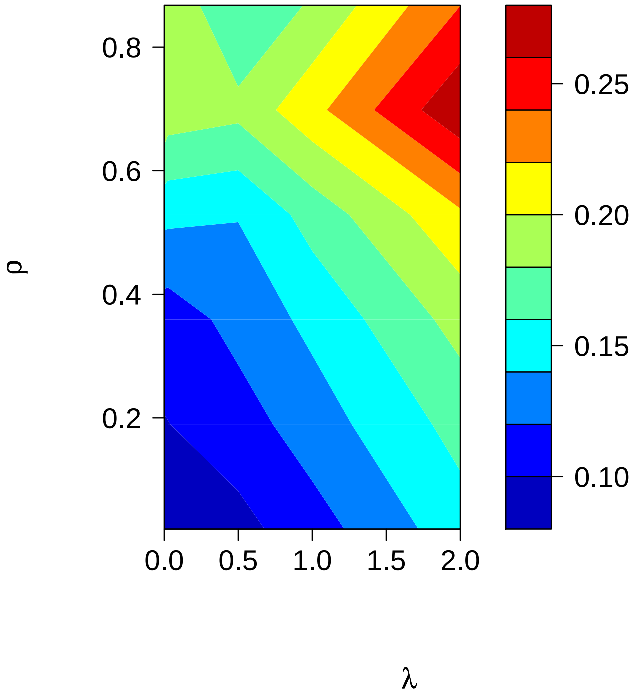


Figure 6.4: *(a) Projected temperature increase for two synthetic 12-models ensembles as a function of the weighting λ. The solid lines are the mean values and the dashed lines represent one standard deviation. (b) Disagreement between two ensembles as a function of the weighing λ and the correlation ρ between present-day bias and future temperature increase. The disagreement is measured with the Kullback-Leibler divergence and 10'000 synthetic models combinations have been averaged to ensure robustness.*

How does this result relate to a set of real GCMs? First, note that model calibration is also a form of weighting where inadequate model configurations are eliminated. Second, some studies have chosen to dismiss certain models by selecting a subset of CMIP3 ensembles in order to constrain future climate change (Walsh et al., 2008; Mahlstein and Knutti, 2010). In these studies, the constrained variable correlates relatively well with the present-day performance measure. With a moderate weighting equivalent to the elimination of poor models, a certain similarity exists with the conceptual model presented here. A high correlation between present-day performance and future climate change may indicate that the information contained in the observation have not been fully used to calibrate the models. On the other hand, the absence of correlation between model fidelity in simulating the 20^{th} century trend and future warming is a possible evidence of model tuning towards the observed trend (Knutti, 2008b). The agreement of the CMIP3 models to reproduce the observed warming together with the absence of correlation also fit the presented results.

Yet, the calibration of the parameters during the model development process might also be dangerous for at least two reasons. First, when many parameter combinations exist as in perturbed physics ensemble (see e.g., Murphy et al., 2004), calibration based on present-day performance is close to screening optimal predictors and systematic biases might appear in the future (DelSole and Shukla, 2009). Second, while calibration of models results in clearer consensus among simulations, it also affects diversity and reduces the number of truly independent models. However, these caveats seem to be less problematic than model weighting. With the exception of eliminating clearly identified outliers, weighting does not improve model consistency in the simplified experiments done here.

Chapter 7

Conclusions and outlook

7.1 Conclusions

This PhD thesis was primary focused on the statistical analysis of model ensembles. The strength of this approach is to avoid subjective choices as much as possible and to rely on the data only. However, different statistical approaches give different results and the way models are interpreted affects uncertainty estimates. The interpretation of models depends on the spatial and temporal scale considered as well as whether calibration or model-interdependence are taken into account. The following list summarizes the main findings and puts them into a wider perspective:

- **Optimal spatial scale:** Chapter 3 examined the spatial and temporal scales at which climate models can provide robust results. A definitive answer for an optimal spatial scale does not exist and depends on the variable, the topography and the time-horizon of interest. Moreover, different spatial aggregation techniques or definition of an optimal spatial scale lead to different results (see Räisänen and Ylhäisi, 2010, for a similar study but different conclusions). Overall, this analysis suggests that some form of spatial aggregation is necessary where grid-point information is known to be unreliable.
- **”Constant bias” assumption in regional climate modeling:** The present-day performance of GCMs at small spatial scales is preserved until the end of this century when the global bias is ignored at least for monthly mean values of surface temperature and precipitation. This is consistent with the common practice of removing the control bias from the regional climate change projections which is known as the *constant bias* hypothesis (see e.g., Buser et al., 2009). However, other hypotheses exist as the *constant relation* between models and observation, i.e. the future regional projections are scaled by a factor reflecting how much natural variability is over- or under-estimated by the GCM. An extension of this study is to verify whether the ”constant relation” scaling factor is also preserved through time.
- **Model resolution:** Model resolution in CMIP3 seems to affect present-day temperature performance only for small scales over land. This is consistent with the fact that uncertainties in climate projections have not decreased significantly in the last decade despite massive computational advances allowing for higher model resolution. On the other hand, higher resolution is indeed required for reliable estimation of future heat-wave,

droughts or changes of tropical cyclone frequencies which also matter for adaptation and mitigation planing (Palmer et al., 2008). This raises an interesting debate about future ensembles of high-resolved models. While such models produce significantly more data, network devices or storage capacity unfortunately does not scale at the same rate as computational capacity. From the data user perspective, this means loaded network bandwidth, more space required on storage devices and an increase of data pre-processing.

- **Model calibration:** Even if an observational constraint on climate sensitivity was found in chapter 4, it does not reduce uncertainty beyond the original CMIP3 range of $2-4.5$ K. As model calibration might have used all the information contained in the observation, it is not surprising that this information cannot be used a second time to constrain climate sensitivity. This suggests challenging interpretations for the estimation of future climate change. First, if models are misrepresenting certain processes due to structural deficiencies, tuning parameters might lead to other problems, e.g. other feedbacks compensating for what is missing in a model (Kiehl, 2007; Huybers, 2010). For the same reason, model weighting based on present-day skill might be risky. Such issues may not necessarily be apparent in a simulation of the present day climate, but could be important in the future.
- **Model inter-dependence:** Chapter 5 points to strong similarities between models developed at the same institution or sharing the same atmospheric component. Therefore, models cannot be considered as independent. Moreover, projections based on the CMIP3 multi-model mean are biased toward certain model designs and do not capture the true diversity of climate models. A hierarchical clustering based on control simulations suggests a gradual evolutionary development of climate models through successive generations. Models are rarely built with lasting value as a primary goal and when a new generation of GCMs is gathered in an ensemble, older models are not maintained any longer. Yet, in order to understand why models differ, comparing results from older model versions and common scenarios is critical. Therefore, older model simulations should be as accessible as those of the newest generation. In addition, the newer models should at least run a scenario shared by previous versions (e.g. the A1B scenario). This has the advantage of disentangling progress made in climate modeling from the use of different scenarios.
- **Model weighting:** The conceptual experiment done in chapter 6 has illustrated the risk of model weighting for the reliability of future climate projections. Consistent with Räisänen et al. (2009), we have found that model weighting is justified only when a relation between present-day performance and future climate projection exists. In this simplified experiment, model weighting is beneficial when models are not tuned or when some models are clearly deficient. To conclude this part, let us recall that a couple of years ago, the fact that all available models were often averaged with equal weight, regardless of their biases, might have seemed strange. The scientific community was building metrics in order make accurate climate projections (e.g., Giorgi and Mearns, 2003; Tebaldi et al., 2005; Greene et al., 2006; Furrer et al., 2007). In the meanwhile, other studies (Tebaldi and Knutti, 2007; Räisänen et al., 2009; Weigel et al., 2010) have suggested serious caveats associated with inappropriate weighting. Though imperfect, giving the same weight to all models seems less risky than model weighting.

7.2 Outlook

While the first GCM prototypes designed in the late 1940s were run on a computer able to perform 5000 operations per second, the typical capacity available in 2007 was in the order of 10^{12} floating point operation per second (one tera flop). This evolution has allowed the direct dynamical computation of certain processes, reducing uncertainty related to parameterization. The new generation of supercomputers (IBM Roadrunner, Cray Jaguar) already offer capacities in the order of 10^{15} flops. How should climate models make the best use out of it? This question is linked to another: how to make the best use of a collection of models? A simple solution does not exist, but trying to answer these question helps a better coordination of GCM experiments which clarifies the interpretation of future climate change. With the growing amount of data produced by GCMs, the role of statistics will be enhanced, either to extract relevant information or to provide an adequate frame to interpretation. The following aspects might be addressed by further studies:

- **Information theory:** Model skill is often evaluated using root mean square errors (rmse) of either climatological averages or variances for a collection of variables (e.g., Gleckler et al., 2008; Reichler and Kim, 2008; Pennell and Reichler, 2010). However, these metrics do not explicitely formulate how these quantities are related, nor are they invariant under changes of variables even though the same physics is being described (Majda and Gershgorin, 2010). A more complete description of climate is possible by using multivariate distributions together with information theory. The results found in chapter 5 and in other studies (e.g., Shukla et al., 2006; DelSole and Tippett, 2007; Majda and Gershgorin, 2010) suggest promising applications of empirical information theory into climate data processing.
- **Nature of model ensembles:** The recent past years have seen the formulation of two challenging statistical frameworks to quantify climate change uncertainty. The first is known as the *truth plus error* view and considers each simulation as an individual sample of a distribution centered around the truth (see e.g., Tebaldi and Knutti, 2007). A second viewpoint is known as the *statistically indistinguishable* paradigm, where both models and observations are sampled from the same distribution, but where observations do not necessary coincide with the center of the distribution (see e.g., Tebaldi and Sanso, 2009; Annan, 2010b). Current progress is made towards a reconciliation of both views and the resulting formulation might also impact the way we think about multi-model combination and weighting.
- **Model diversity:** George E. Box is credited for having said ”all models are wrong but some are useful”. In the context of model ensembles this might be rephrased as ”ensembles are useless when individual members all agree with each other”. Model disagreement reflects different hypotheses about the climate system. This diversity is crucial in order to sample all the uncertainty we think exists. Several solutions might improve model diversity. An idea is to generate perturbed simulations as different as possible (as in Stainforth et al., 2005), but whose proximity with the observation is maximized (as in Murphy et al., 2004; Webb et al., 2006). Assuming that the financial and technical capacity is available, each GCM institution would be encouraged to provide a dozen of such perturbed yet calibrated simulations. The experiment would be collected into

a coordinated ensemble and could sample a larger region of structural and parametric uncertainty.

- **Risk of model calibration and weighting:** We have explored in chapter 4 and 6 the role of model calibration in the CMIP3 ensemble. It might be interesting to study the risk of model calibration for future climate projections using perturbed physics ensembles with some cross-validation technique. Instead of artificial models where the interpretation is limited, perturbed ensembles or some model of intermediate complexity as the Bern2.5D model (Stocker et al., 1992; Knutti et al., 2002) could be used instead.

There is no doubt that quantifying uncertainty of future climate change is a crucial question for those planning decisions related to adaptation and mitigation. More reliable estimates reflect progress made in observation measurements, process understanding (physics of climate), statistical concepts and computing resources. Therefore, uncertainty should not be confused with ignorance. Whatever progress is made in climate science, uncertainty cannot be eliminated completely. Science is therefore most useful to society when it finds good ways in exploring and communicating uncertainty.

Acknowledgments

I would like to express my gratitude to my teachers and professors. Many among them gave their best to transmit not only their knowledge, but also tools for a deeper understanding of what they taught.

I wish to especially thank my supervisor Prof. Dr. Reto Knutti for fostering the stimulating atmosphere that prevails in our group, and for the fruitful discussions we had these past years. Reto introduced me to new concepts and points of view in a way that fueled both my curiosity and enthusiasm. I thank him also for his attention and patience.

I am also grateful to all of my colleagues of the ClimPhys group at IAC*ETH* for a pleasant working environment. While exchanging and debating ideas allowed for a vibrant intellectual climate, humor and kindness were often part of the whole atmosphere. How could constraining future climate uncertainty not be a pleasure with some cake during the coffee break or when listening to *Radio Bias* in the office? Thanks also to Dr. Urs Beyerle, for there is no more insurmountable barrier between humans and machines when he is around. Big thanks also to our secretaries Esther Jampen and Rosemarie Widmer for their administrative help, their good spirit and smile.

I owe a special thank to PD Dr. Diethelm Würtz for his encouragement and trust. He has not only introduced me to the exiting field of statistics, but have also played a positive role in the decisions that I have made.

My closest friends Gabriela d'Hondt, Yannick Städler and Charles Ravussin must also know how deep my gratitude goes to them. Generosity, support but also constructive debates brought me a lot during the time of my PhD.

Most important, I express my gratitude to all of my family for everything they did to encourage me. In particular, to my mother Elizabeth, my father Frank and my grandmother Marie-Lise Masson for their unconditional love and constant care during all these years.

Finally, many thanks to all those who have contributed to this thesis in any other way.

Curriculum Vitae

David Masson, Agnesstrasse 43, 8004 Zurich, Switzerland.
Born on 13 April 1979.
Swiss and French citizen.

EDUCATION AND PROFESSIONAL TRAINING

10.2007	-	01.2011	Ph.D. student in the group of Prof. R. Knutti, Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland.
05.2006	-	07.2007	Internship in financial engineering at Finance Online, Zurich, Switzerland.
07.2005	-	10.2006	Quality and testing in software engineering, Swiss Re, Zurich, Switzerland.
10.2004	-	06.2005	Study of the Russian language and culture at the Saint-Petersburg State University.
10.1998	-	04.2004	Studies in Physics at ETH Zurich, Switzerland. Diploma thesis in Theoretical Physics under the direction of Prof. J.-P. Derendinger (University of Fribourg) in collaboration with Prof. D. Wyler (University of Zurich) and Prof. J. Froehlich (ETH Zurich).
1995	-	1998	Maths-Physics high school degree in Lausanne.

AWARD

High school award in physics.

PUBLICATIONS

Masson, D. and R. Knutti, 2011: Spatial scale dependence of climate model performance in the CMIP3 ensemble. *Journal of Climate*, doi:10.1175/2011JCLI3513.1.

Masson, D. and R. Knutti, 2011: Constraining climate sensitivity from interannual variability: an illustration of tuning in climate model ensembles. In revision in *Journal of Climate*.

Masson, D. and R. Knutti, 2011: Climate model genealogy, *Geophysical Review Letters*, 38, L08703, doi:10.1029/2011GL046864.

ORAL PRESENTATION

Masson, D. and R. Knutti, 2010, Constraining climate sensitivity from interannual variability: an illustration of tuning in climate model ensembles. International Meetings of Statistical Climatology, Edinburgh, United Kingdom, July 2010.

POSTER PRESENTATIONS

8th NCCR Climate Summer School, Grindelwald, September 2009.
7th NCCR Climate Summer School, Monte Verità, September 2008.

Bibliography

Adler, R. F., et al., 2003: The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present). *Journal Of Hydrometeorology*, **4 (6)**, 1147–1167.

Annan, J. D., 2010a: Climate model independence and agreement. *Geophysical Research Letters*, submitted.

Annan, J. D., 2010b: Understanding the CMIP3 multi-model ensemble. *Geophysical Research Letters*, submitted.

Annan, J. D. and Hargreaves, J. C., 2010: Reliability of the CMIP3 ensemble. *Geophysical Research Letters*, **37**, L02 703.

Barnett, T., et al., 2005: Detecting and attributing external influences on the climate system: A review of recent advances. *Journal Of Climate*, **18 (9)**, 1291–1314.

Boé, J., A. Hall, and X. Qu, 2009a: Deep ocean heat uptake as a major source of spread in transient climate change simulations. *Geophysical Research Letters*, **36**, L22 701, doi: 10.1029/2009GL040845.

Boé, J. L., A. Hall, and X. Qu, 2009b: September sea-ice cover in the arctic ocean projected to vanish by 2100. *Nature Geoscience*, **2 (5)**, 341–343.

Breiman, L., J. Friedman, C. Stone, and O. R.A., 1984: *Classification and regression trees*. Chapman and Hall, 368 pp.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal Of Geophysical Research-Atmospheres*, **111 (D12)**.

Buser, C. M., H. R. Künsch, D. Lüthi, M. Wild, and C. Schär, 2009: Bayesian multi-model projection of climate: bias assumptions and interannual variability. *Climate Dynamics*, **33 (5)**, 849–868, doi:10.1007/s00382-009-0588-6.

Christensen, J. H., et al., 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, edited by S. Solomon et al. Cambridge University Press, 847–940.

Collins, M., B. B. B. Booth, B. Bhaskaran, G. Harris, J. Murphy, D. Sexton, and M. Webb, 2010: Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. *Climate Dynamics*, doi:10.1007/s00382-010-0808-0, published online.

Covey, C., K. M. AchutaRao, U. Cubasch, P. Jones, S. J. Lambert, M. E. Mann, T. J. Phillips, and K. E. Taylor, 2003: An overview of results from the coupled model intercomparison project. *Global And Planetary Change*, **37 (1-2)**, 103–133.

Covey, C., et al., 2000: The seasonal cycle in coupled ocean-atmosphere general circulation models. *Climate Dynamics*, **16 (10-11)**, 775–787.

Cox, P. and D. Stephenson, 2007: Climate change - a changing climate for prediction. *Science*, **317 (5835)**, 207–208, doi:10.1126/science.1145956.

DelSole, T. and J. Shukla, 2009: Artificial skill due to predictor screening. *Journal of Climate*, **22 (2)**, 331–345, doi:10.1175/2008JCLI2414.1.

DelSole, T. and M. K. Tippett, 2007: Predictability: Recent insights from information theory. *Reviews of Geophysics*, **45 (4)**, RG4002, doi:10.1029/2006RG000202.

Edwards, P. N., 2010: *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. The MIT Press, 528 pp.

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 (5552)**, 113–117.

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. *Geophysical Research Letters*, **34 (6)**, L06 711.

Giorgi, F. and R. Francisco, 2000: Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. *Climate Dynamics*, **16 (2-3)**, 169–182.

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. *Journal Of Climate*, **15 (10)**, 1141–1158.

Giorgi, F. and L. O. Mearns, 2003: Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophysical Research Letters*, **30(12)**, doi: 10.1029/2003GL017130.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *Journal Of Geophysical Research-Atmospheres*, **113 (D6)**.

GMAO, 2009: Modern era retrospective-analysis for research and applications. http://gmao.gsfc.nasa.gov/research/merra.

Greene, A. M., L. Goddard, and U. Lall, 2006: Probabilistic multimodel regional temperature change projections. *Journal Of Climate*, **19 (17)**, 4326–4343.

Gregory, J. M. and P. M. Forster, 2008: Transient climate response estimated from radiative forcing and observed temperature change. *Journal Of Geophysical Research-Atmospheres*, **113 (D23)**, doi:10.1029/2008JD010405.

Gregory, J. M., et al., 2004: A new method for diagnosing radiative forcing and climate sensitivity. *Geophysical Research Letters*, **31 (3)**, L03 205.

Hall, A. and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophysical Research Letters*, **33 (3)**, L03 502.

Hansen, J., D. Johnson, A. Lacis, S. Lebedeff, P. Lee, D. Rind, and G. Russell, 1981: Climate impact of increasing atmospheric carbon-dioxide. *Science*, **213 (4511)**, 957–966.

Hastie, T., R. Tibshirani, and J. Friedman, 2001: *The elements of statistical learning*. Springer, 746 pp.

Hawkins, E. and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, **90 (8)**, 1095–1107, doi:10.1175/2009BAMS2607.1.

Held, I. M., 2005: The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, **86 (11)**, 1609–1614, doi:10.1175/BAMS-86-11-1609.

Huybers, P., 2010: Compensation between model feedbacks and curtailment of climate sensitivity. *Journal Of Climate*, **23 (11)**, 3009–3018.

IPCC, 2001: *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 881 pp.

IPCC, 2007: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessement Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 996 pp.

Jackson, J. D., 1998: *Classical Electrodynamics*. Wiley, 808 pp.

Jun, M. Y., R. Knutti, and D. W. Nychka, 2008a: Local eigenvalue analysis of CMIP3 climate model errors. *Tellus Series A-Dynamic Meteorology And Oceanography*, **60 (5)**, 992–1000.

Jun, M. Y., R. Knutti, and D. W. Nychka, 2008b: Spatial Analysis to Quantify Numerical Model Bias and Dependence: How Many Climate Models Are There? *Journal Of The American Statistical Association*, **103 (483)**, 934–947.

Kalnay, E., et al., 1996: The NCEP/NCAR 40-year reanalysis project. *Bulletin Of The American Meteorological Society*, **77 (3)**, 437–471.

Kennedy, M. C. and A. O'Hagan, 2001: Bayesian calibration of computer models. *Journal Of The Royal Statistical Society Series B-Statistical Methodology*, **63**, 425–450.

Kiehl, J. T., 2007: Twentieth century climate model response and climate sensitivity. *Geophysical Research Letters*, **34 (22)**, L22 710.

Kim, K. Y. and Q. G. Wu, 2000: Optimal detection using cyclostationary EOFs. *Journal Of Climate*, **13 (5)**, 938–950.

Knight, C. G., et al., 2007: Association of parameter, software, and hardware variation with large-scale behavior across 57000 climate models. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, **104 (30)**, 12 259–12 264.

Knutti, R., 2008a: Should we believe model predictions of future climate change? *Philosophical Transactions Of The Royal Society A*, **366 (1885)**, 4647–4664.

Knutti, R., 2008b: Why are climate models reproducing the observed global surface warming so well? *Geophysical Research Letters*, **35 (18)**, L18 704.

Knutti, R., 2010: The end of model democracy? *Climatic Change*, **102 (3-4)**, 395–404.

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

Knutti, R., R. Furrer, C. Tebaldi, and J. Cermak, 2010b: Challenges in combining projections from multiple climate models. *Journal Of Climate*, **23 (10)**, 2739–2758, doi: 10.1175/2009JCLI3361.1.

Knutti, R. and G. C. Hegerl, 2008: The equilibrium sensitivity of the Earth's temperature to radiation changes. *Nature Geoscience*, **1 (11)**, 735–743.

Knutti, R., S. Krahenmann, D. J. Frame, and M. R. Allen, 2008a: Comment on "heat capacity, time constant, and sensitivity of earth's climate system" by s. e. schwartz. *Journal Of Geophysical Research-Atmospheres*, **113 (D15)**, D15 103.

Knutti, R., G. A. Meehl, M. R. Allen, and D. A. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *Journal Of Climate*, **19 (17)**, 4224–4233.

Knutti, R., T. F. Stocker, F. Joos, and G. K. Plattner, 2002: Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature*, **416 (6882)**, 719–723.

Knutti, R., et al., 2008b: A review of uncertainties in global temperature projections over the twenty-first century. *Journal Of Climate*, **21 (11)**, 2651–2663.

Kullback, S., 1968: *Information Theory and Statistics*. Dover Publications, 399 pp.

Kumar, A. and F. L. Yang, 2003: Comparative influence of snow and SST variability on extratropical climate in northern winter. *Journal Of Climate*, **16 (13)**, 2248–2261.

Levitus, S., J. I. Antonov, T. P. Boyer, and C. Stephens, 2000: Warming of the world ocean. *Science*, **287 (5461)**, 2225–2229.

Mahlstein, I. and R. Knutti, 2009: Regional climate change patterns identified by cluster analysis. *Climate Dynamics*, doi:10.1007/s00382-009-0654-0, published online.

Mahlstein, I. and R. Knutti, 2010: Ocean heat transport as a cause for model uncertainty in projected Arctic warming. *Journal Of Climate*, doi:10.1175/2010JCLI3713.1, early online release.

Majda, A. J. and B. Gershgorin, 2010: Quantifying uncertainty in climate change science through empirical information theory. *Proceedings of the National Academy of Sciences of the United States of America*, **107 (34)**, 14 958–14 963, doi:10.1073/pnas.1007009107.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor, 2007a: The WCRP CMIP3 multimodel dataset - A new era in climate change research. *Bulletin Of The American Meteorological Society*, **88**, 1383–1394.

Meehl, G. A., C. Covey, B. McAvaney, M. Latif, and R. J. Stouffer, 2000: The Coupled Model Intercomparison Project (CMIP). *Bulletin Of The American Meteorological Society*, **81 (2)**, 313–318.

Meehl, G. A., W. M. Washington, W. D. Collins, J. M. Arblaster, A. X. Hu, L. E. Buja, W. G. Strand, and H. Y. Teng, 2005: How much more global warming and sea level rise? *Science*, **307 (5716)**, 1769–1772.

Meehl, G. A., et al., 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, edited by S. Solomon et al. Cambridge University Press, 747–785.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, and M. Collins, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430 (7001)**, 768–772.

Nakicenovic, N., et al., 2000: *IPCC Special Report on Emissions Scenarios*. Cambridge University Press, 570 pp.

Oreskes, N., K. Shrader-Frechette, and K. Belitz, 1994: Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, **263 (5147)**, 641–646.

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. *Bulletin Of The American Meteorological Society*, **89 (4)**, 459–470.

Parker, W., 2006: Understanding pluralism in climate modeling. *Foundations of Science*, **11**, 349–368, doi:10.1007/s10699-005-3196-x.

Pennell, C. and T. Reichler, 2010: On the Effective Number of Climate Models. *Journal Of Climate*, doi:10.1175/2010JCLI3814.1, early online release.

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. *Geophysical Research Letters*, **32**, L23 825, doi:10.1029/2005GL024452.

Pirtle, Z., R. Meyer, and A. Hamilton, 2010: What does it mean when climate models agree? A case for assessing independence among general circulation models. *Environmental Science & Policy*, **13 (5)**, 351–361.

Popper, K. R., 1959: *The logic of scientific discovery*. Basic Books, 480 pp.

Räisänen, J., 2001: CO_2-induced climate change in CMIP2 experiments: Quantification of agreement and role of internal variability. *Journal Of Climate*, **14 (9)**, 2088–2104.

Räisänen, J., 2007: How reliable are climate models? *Tellus Series A-Dynamic Meteorology And Oceanography*, **59 (1)**, 2–29.

Räisänen, J., L. Ruokolainen, and J. Ylhäisi, 2009: Weighting of model results for improving best estimates of climate change. *Climate Dynamics*, **35**, 407–422, doi:10.1007/s00382-009-0659-8.

Räisänen, J. and J. Ylhäisi, 2010: How much should climate model output be smoothed in space? *Journal Of Climate*, doi:10.1175/2010JCLI3872.1, early online release.

Randall, D., et al., 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, edited by S. Solomon et al. Cambridge University Press, 589–662.

Randall, D. A. and B. A. Wielicki, 1997: Measurements, models, and hypotheses in the atmospheric sciences. *Bulletin Of The American Meteorological Society*, **78 (3)**, 399–406.

Reichler, T. and J. Kim, 2008: How well do coupled models simulate today's climate? *Bulletin Of The American Meteorological Society*, **89 (3)**, 303–311.

Roe, G. H. and M. B. Baker, 2007: Why is climate sensitivity so unpredictable? *Science*, **318 (5850)**, 629–632.

Sanderson, B., K. Shell, and W. J. Ingram, 2010: Climate feedbacks determined using radiative kernels in a multi-thousand member ensemble of AOGCMs. *Climate Dynamics*, **35**, 1219–1236, doi:10.1007/s00382-009-0661-1.

Sanderson, B. M., 2010: A multi-model study of parametric uncertainty in predictions of climate response to rising greenhouse gas concentrations. *Journal Of Climate*, doi:10.1175/2010JCLI3498.1, early online release.

Sanderson, B. M., et al., 2008: Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. *Journal Of Climate*, **21 (11)**, 2384–2400.

Santer, B. D., et al., 2009: Incorporating model quality information in climate change detection and attribution studies. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, **106 (35)**, 14 778–14 783.

Scherrer, S. C., 2010: Present-day interannual variability of surface climate in CMIP3 models and its relation to future warming. *International Journal of Climatology*, doi:10.1002/joc.2170, published online.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter, and D. Paolino, 2006: Climate model fidelity and projections of climate change. *Geophysical Research Letters*, **33**, L07 702, doi:10.1029/2005GL025579.

Soden, B. J. and I. M. Held, 2006: An assessment of climate feedbacks in coupled ocean-atmosphere models. *Journal Of Climate*, **19 (14)**, 3354–3360.

Stainforth, D. A., M. R. Allen, E. R. Tredger, and L. A. Smith, 2007: Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transactions Of The Royal Society A - Mathematical Physical And Engineering Sciences*, **365 (1857)**, 2145–2161.

Stainforth, D. A., et al., 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433 (7024)**, 403–406.

Stocker, T. F., D. G. Wright, and L. A. Mysak, 1992: A Zonally Averaged, Coupled Ocean Atmosphere Model For Paleoclimate Studies. *Journal of Climate*, **5 (8)**, 773–797.

Stott, P. A. and J. A. Kettleborough, 2002: Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature*, **416 (6882)**, 723–726.

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

Stouffer, R. J., 2004: Time scales of climate response. *Journal Of Climate*, **17 (1)**, 209–217.

Stroeve, J., M. M. Holland, W. Meier, T. Scambos, and M. Serreze, 2007: Arctic sea ice decline: Faster than forecast. *Geophysical Research Letters*, **34 (9)**, L09 501, doi:10.1029/2007GL029703.

Tebaldi, C. and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions Of The Royal Society A - Mathematical Physical And Engineering Sciences*, **365 (1857)**, 2053–2075.

Tebaldi, C. and B. Sanso, 2009: Joint projections of temperature and precipitation change from multiple climate models: a hierarchical bayesian approach. *Journal of the Royal Statistical Society Series A-statistics In Society*, **172**, 83–106, doi:10.1111/j.1467-985X.2008.00545.x.

Tebaldi, C., R. L. Smith, D. Nychka, and L. O. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *Journal Of Climate*, **18 (10)**, 1524–1540.

Uppala, S. M., et al., 2005: The ERA-40 re-analysis. *Quarterly Journal Of The Royal Meteorological Society*, **131 (612)**, 2961–3012.

von Storch, H. and F. Zwiers, 2004: *Statistical Analysis in Climate Research*. Cambridge University Press, 485 pp.

Walsh, J. E., W. L. Chapman, V. Romanovsky, J. H. Christensen, and M. Stendel, 2008: Global Climate Model Performance over Alaska and Greenland. *Journal of Climate*, **21 (23)**, 6156–6174, doi:10.1175/2008JCLI2163.1.

Washington, W. M., R. Knutti, G. A. Meehl, H. Y. Teng, C. Tebaldi, D. Lawrence, L. Buja, and W. G. Strand, 2009: How much climate change can be avoided by mitigation? *Geophysical Research Letters*, **36**, L08 703.

Webb, M. J., et al., 2006: On the contribution of local feedback mechanisms to the range of climate sensitivity in two GCM ensembles. *Climate Dynamics*, **27 (1)**, 17–38, doi:10.1007/s00382-006-0111-2.

Weigel, A., R. Knutti, M. Liniger, and C. Appenzeller, 2010: Risks of model weighting in multi-model climate projections. *Journal Of Climate*, **23**, 4175–4191, doi:10.1175/2010JCLI3594.1.

Whetton, P., I. Macadam, J. Bathols, and J. O'Grady, 2007: Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models. *Geophysical Research Letters*, **34 (14)**, L14 701, doi:10.1029/2007GL030025.

Wigley, T. M. L., R. L. Smith, and B. D. Santer, 1998: Anthropogenic influence on the autocorrelation structure of hemispheric-mean temperatures. *Science*, **282 (5394)**, 1676–1679.

Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences*. Elsevier, 627 pp.

Wu, Q. G., D. J. Karoly, and G. R. North, 2008: Role of water vapor feedback on the amplitude of season cycle in the global mean surface air temperature. *Geophysical Research Letters*, **35 (8)**, L08 711.

Wu, Q. G. and G. R. North, 2003: Statistics of calendar month averages of surface temperature: A possible relationship to climate sensitivity. *Journal Of Geophysical Research-Atmospheres*, **108 (D2)**, 4071.

Xie, P. P. and P. A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bulletin Of The American Meteorological Society*, **78 (11)**, 2539–2558.

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 (7152)**, 461–466.