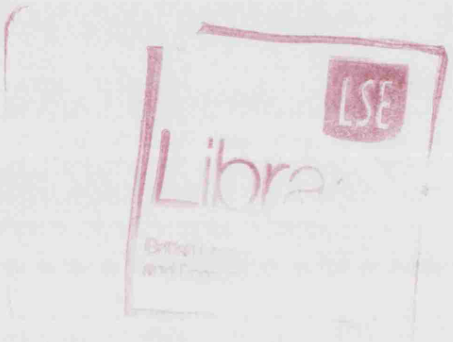
On the Evaluation of Uncertainties in Climate Models

Edward Tredger

London School of Economics and Political Science

Submitted to the Department of Statistics of the London School of Economics for the Degree of Doctor of Philosophy, London, 2009

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346



Declaration

I certify that the Thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

Acknowledgements

The research presented in this Thesis has been carried out with the generous support of an EPSRC Studentship and an LSE Statistics Departmental Scholarship. I am grateful for data provided by the Coupled Model Intercomparison Project and the climateprediction.net experiment. Without the contribution of members of the public to the climateprediction.net experiment many aspects of this Thesis would not be have been possible.

I am happy to acknowledge the guidance and support of my Supervisor, Lenny Smith and fruitful discussions with David Stainforth and all the members of CATS.

Abstract

The prediction of the Earth's climate system is of immediate importance to many decision-makers. Anthropogenic climate change is a key area of public policy and will likely have widespread impacts across the world over the 21st Century. Understanding potential climate changes, and their magnitudes, is important for effective decision making. The principal tools used to provide such climate predictions are physical models, some of the largest and most complex models ever built. Evaluation of state-of-the-art climate models is vital to understanding our ability to make statements about future climate. This Thesis presents a framework for the analysis of climate models in light of their inherent uncertainties and principles of statistical good practice. The assessment of uncertainties in model predictions to–date is incomplete and warrants more attention that it has previously received. This Thesis aims to motivate a more thorough investigation of climate models as fit for use in decision–support.

The behaviour of climate models is explored using data from the largest ever climate modelling experiment, the climateprediction.net project. The availability of a large set of simulations allows novel methods of analysis for the exploration of the uncertainties present in climate simulations. It is shown that climate models are capable of producing very different behaviour and that the associated uncertainties can be large. Whilst no results are found that cast doubt on the hypothesis that greenhouse gases are a significant driver of climate change, the range of behaviour shown in the climateprediction.net data set has implications for our ability to predict future climate and for the interpretation of climate model output. It is argued that uncertainties should be explored and communicated to users of climate predictions in such a way that decision-makers are aware of the relative robustness of climate model output.

Contents

List of Figures

Nomenclature

Roman Symbols

IC Initial Conditions

AR4 IPCC Fourth Assessment Report

CO_2 Carbon Dioxide

CPDN climateprediction.net

GCM General Circulation Model

GHG Greenhouse Gas

GMST Global Mean Surface Temperature

HFA Heat Flux Adjustment

ICE Initial Condition Ensemble

ICU Initial Condition Uncertainty

IPCC Intergovernmental Panel on Climate Change

NWP Numerical Weather Prediction

PDF Probability Density Function

PPE Perturbed Physics Experiment

SOTA State of the Art

SST Sea Surface Temperature

Chapter 1

Introduction

In recent years *climate change*[1] has become a significant issue in science, politics and the media. Whilst theories of man-made global warming have been around since the 19th Century Arrhenius (1896); Tyndall (1861) only within the last 30 years has climate science become a major scientific focus. Furthermore, it is only over the past 10 years that climate change has become a significant political issue, driven by an increasing awareness of its potential impacts. Assessing the human impact on the Earth's climate has become a major area of research and is of importance to many different decision–makers Association of British Insurers (2005); Parry *et al.* (2007); Stainforth *et al.* (2007b); Stern (2006).

Whilst certain details may be still under dispute, it has become widely accepted that the changes in climate that have occurred over the past 100 years are largely *anthropogenic* (man–made) Oreskes (2004); Solomon *et al.* (2007a) and that anthropogenic factors will continue to have a significant effect throughout the 21st Century. Focus has turned to predicting the details of how the climate will change over the next Century. This Thesis examines the robustness of these details and evaluates uncertainties in climate simulations.

Accurate climate prediction would be useful for at least three reasons:

1. **Mitigation Decisions.** The likely results of *mitigation* decisions can be better understood in the light of reliable climate predictions. For example, when considering whether (and by how much) to cut CO_2 emissions, it is important to know how the climate might change for given *scenario* of future CO_2 emissions Schellnhuber *et al.* (2005). When setting targets for emissions

[1]Terms defined in the Glossary are shown in italics at their first use.

it is helpful to know how different emission scenarios will relate to climate change on both local and global levels in a number of meteorological variables.

2. **Adaptation Decisions.** The planning of *adaptation* measures (e.g. the building of flood defences) would benefit from insight to how the climate will change on national or finer length scales. For example, suppose a government is considering building a dam to prevent future flooding. The optimal design and placement of this dam depends on several climatic factors. These include the frequency, intensity, spatial and temporal patterns of precipitation in the future, how much sea levels might rise, whether storms are likely to be more frequent and intense and the correlations between these factors.

3. **Impacts Assessment.** Even where no adaptation or mitigation decisions are planned directly, it is of interest to industry and government to know how the climate will change in the future e.g. when considering how energy demand might change in the coming years to decades. Take the example of a Life or General insurance company. It might be of great use to an insurer to know whether there is likely to be an increasing trend in extreme weather events, changes in mortality rates and thus estimate what additional capital might be needed to protect against future climate–related claims.

One of the key questions looked at in this Thesis is "How might climate models inform such decisions?". In order to provide support to decision–makers, projections of future climate have been provided by state–of–the–art *climate models.* These models are known as General Circulation Models (*GCMs*). GCMs represent a large investment of scientific research and resources McGuffie & Henderson-Sellers (2006); Solomon *et al.* (2007a); Thorpe (2005). The subject of this Thesis is the evaluation of GCMs for decision–support.

The problem of climate prediction poses many new and interesting challenges to statisticians. Increased study on the evaluation of complex models with little *out–of–sample* data would be of value in a number of different areas of applied statistics. In the case of climate predictions, statistical methods of model evaluation based on a comparison of out–of–sample predictions with observational data are hampered by lack of data – the long lead times of forecasts (10+ years) and the relative novelty of the field limit the potential for such methods of evaluation.

This Thesis uses alternative methods to assess the potential value of climate models to decision makers. The methodology used is briefly introduced in Section 1.1. Important results are highlighted in Section 1.2.

1.1 Methodology

This Thesis explores and analyses the richest source of climate model data available to date – the CPDN data set. The aim of this Thesis is to evaluate the uncertainties in climate model projections for decision–support and to establish statistical good practice relevant to this field.

The reliability of a model's out–of–sample predictions can never be verified in the sense of establishing a model as "true" Oreskes (1998); Oreskes *et al.* (1994). Methods of model evaluation can show where predictions are likely to be inadequate, but a model's out–of–sample predictions can never be proven to be accurate. Evaluation of potential model skill is especially difficult in the case of climate prediction where out–of–sample observations are lacking Smith (2002); Stainforth *et al.* (2007a). In light of this, the approach adopted in this Thesis is to check climate models for consistency of information rather than seeking to verify the models in any sense. Model output is said to be consistent where differences between simulations are not critical for decision–makers. These differences can be analysed **a)** across different structural models, **b)** in the same structural model across different parameter values or **c)** within a particular model across the starting state used to initialise a particular model simulation, the model's *initial conditions* (an *Initial Condition Ensemble*, or ICE). A set of Initial Condition Ensembles, each run under different model structures or parameter values (as in cases **a)** and **b)**) together form a *grand ensemble.* The diversity of output across model projections places a limit on the utility of model output in decision–support.

There have been attempts to attach *probabilities* to climate changes Annan & Hargreaves (2006); Giorgi & Mearns (2003); Pittock *et al.* (2001). This Thesis does not attempt to attach probabilities to climate impacts for four reasons;

1. Due to *model inadequacy* Kennedy & O'Hagan (2001b), there is no reason that the details of a model's climate distribution will hold for the Earth's climate system Smith (2002).

2. The effects of arbitrary choices of experimental design and parameter specification can affect the probabilities attained significantly Frame *et al.* (2005, 2007).

3. It is difficult to see how to probabilistically combine output from different models given the lack of a reliable metric in model space Allen & Stainforth (2002).

4. It may not be necessary to express climate projections as probabilities for model output to be useful Dessai & Hulme (2004); Judd (2008a).

Despite attempts to create Bayesian probability distributions for climate Annan & Hargreaves (2006); Goldstein & Rougier (2006), the Bayesian approach faces significant theoretical and practical difficulties, as outlined above. Rather than adopting a framework for uncertainties based on probability distributions, this Thesis quantifies different types of known uncertainties present in climate models and the level of consistency between models. Checking for consistency of information across an *ensemble* of models provides a direct evaluation of the robustness of model projections.

Two important methods for evaluating the reliability of climate models are: **1)** In–sample consistency with observations and the **2)** the range of predictions produced by different models out–of–sample, as given in Raisanen (2007). Further evaluation methods that can be used to gain confidence in model projections listed in Solomon *et al.* (2007a) are the **3)** simulation of present–day climate and **4)** the fact that models are based on well-understood physical principles. Question **1)** is looked at in Thesis in Chapter 3, although the main focus is on evaluation method **2)**, the diversity of model output provided by current models, which is investigated in detail in Chapters 6,7 and 8.

The uncertainty analysis applied in this Thesis is only possible with a large set of climate model simulations. At present, the only source of sufficient data was the *climateprediction.net* (CPDN) experiment, from which 45644 simulations are analysed in this Thesis. In comparison, the Intergovernmental Panel on Climate Change (IPCC) AR4 uses an ensemble of 58 simulations to evaluate climate models in their Summary for Policymakers Solomon *et al.* (2007a,b). Similar numbers of simulations were used to make projections under various scenarios in the IPCC Report. The CPDN data set allows new approaches to the quantification of uncertainty. The availability of a large set of data provides a unique opportunity to test the robustness of climate models.

New results and methods that are presented in this Thesis are given in Section 1.2.

1.2 Key results and new approaches

In this Thesis a framework for the evaluation of climate models is laid out and the types of uncertainties present are demonstrated and explored. The structure of this Thesis is as follows. Chapter 2 introduces a framework for understanding and evaluating the uncertainties in climate simulations. Chapter 3 analyses data used

in the IPCC AR4 in terms of GCMs' in–sample fit and motivates a more thorough assessment of uncertainty. Chapter 4 gives details of the CPDN experiment and the data sets analysed in Chapters 5 through 8 of this Thesis. Chapter 5 examines the heat flux adjustments (artificial adjustments of energy applied to the model's ocean) that are applied to the GCM used in the CPDN experiment. Chapter 6 presents results for the largest Initial Condition Ensemble analysed to date (64 simulations of the Hadley Centre's HadSM3 model). Chapter 7 investigates the range of behaviour across the CPDN data set of 45644 simulations and discusses the use of methods to reduce the range of behaviour shown. Chapter 8 looks at the utility of global mean temperature as a basis for decision–making and the uncertainty present on regional scales for sets of simulations with very similar global mean temperature response. Chapter 9 summarises this work and discusses the implications of new results for decision–makers.
The main advances in this Thesis, by Chapter are:

1.3 Chapter 2

Three tests are proposed that can be used to evaluate whether climate models might be fit for decision support. These tests are based on **1)** The in–sample fit of climate models, **2)** The range of model behaviour within an Initial Condition Ensemble and **3)** The diversity of model behaviour across an ensemble of different models. These strawman tests are presented as means to test the consistency of information in climate models. Uses of models that fail one or more of these tests are discussed.

1.4 Chapter 3: How reliable are climate models?

Chapter 3 examines climate model data presented in the Intergovernmental Panel on Climate Change Fourth Assessment Report (IPCC AR4). It is shown that:

1. There are significant differences between different GCMs' global mean temperature of up to 3 degrees Celsius in their 1901–1950 base climates. Such large differences could affect physical properties of these models that are relevant when comparing model simulations to observations.

2. The effect of taking different types of anomalies is shown to give significantly different presentations of GCMs' in–sample fit. In particular, when taking anomalies for each simulation results in a tighter multi–model ensemble than when taking anomalies with respect to each model (averaging the bias correction over each model's constituent simulations). It is shown that the use

of anomalies in the IPCC AR4 (taking anomalies for each simulation) is less physically meaningful than taking anomalies for each model and the former method distorts the variability of simulations both within individual models and across different structural models.

3. Model output is compared to observed global mean temperatures over the 20th Century. Residuals are compared on a model by model basis and it is shown that **1)** There can be considerable structure in the residual time series and **2)** The magnitude of residuals can be large (up to 0.5 degrees Celsius) in comparison to observed 20th Century global warming ($\sim$0.74 degrees Celsius).

4. The CMIP3 (the third Coupled Model Intercomparison Project Covey *et al.* (2003)) GCMs used in the IPCC AR4 are shown not to be exchangeable, calling into question the relevance of many methods of statistical analysis for climate model output. This is shown by calculating the number of simulations in one ensemble that exceeds the maximum member of another. This empirical statistic is then compared to the theoretical expectation based on the ensemble sizes. Results show that GCMs can not be assumed to be sampling from a common distribution. These initial results were confirmed by a Kruskal–Wallis test.

5. A new method for estimating the temporal correlation within GCM time series is proposed. This method requires an Initial Condition Ensemble and is based on the typical amount of time taken for an extremal simulation (maximum or minimum) to cross the median of the ensemble. It is shown that the mixing time for some GCMs is not significantly different from 1 year (the higher frequency of data used here), but that it can be higher for other GCMs.

1.5 Chapter 4: Intro to CPDN

1. A new method of quality control is presented, correcting for problems identified in previous quality control methods Stainforth *et al.* (2005). An unphysical local feedback in the East Pacific is detected using a local anomaly statistic. This method is shown to eliminate simulations with significant global cooling which fail to be detected when using global mean statistics.

2. Features of the CPDN experiment are documented for the first time e.g. availability of data, experimental design and issues in data analysis. Such documentation is important for other studies based on CPDN data sets.

1.6 Chapter 5: Heat Flux

The HadSM3 climate model used in the CPDN experiment analysed here requires the use of heat flux adjustments (HFA). The variability and effect of the HFAs on a grand ensemble of climate model simulations are looked at for the first time. It is shown that:

1. Perturbation of Initial Conditions has little effect on the global mean HFA (the greatest difference in global mean HFA between IC members across 418 model versions is $0.419W/m^2$). Perturbation of Initial Conditions can lead to differences of up to $40W/m^2$ ($\sim$100 times the greatest global difference) on a grid box level.

2. Parameter perturbed model versions of HadSM3 can require significantly different global HFAs. This is shown by carrying out a Singular Value Decomposition on Initial Condition Ensembles of HadSM3 model versions. The leading Singular Vector is shown to explain significantly more variability in the HFA fields where model simulations share parameter values than where simulations are drawn at random. It is also shown that whilst perturbing Initial Conditions makes less than $0.5W/m^2$ difference on the global mean scale, perturbing parameters can lead to changes of up to $70W/m^2$. It is argued that the HFA should then be calibrated for each set of parameter values.

3. There are shown to be significant seasonal variations in the HFA both globally and regionally. This effect is likely mimicking the seasonally-dampening effect of a deep ocean and means HadSM3's seasonal cycle might not respond to rising CO_2 in a physical way.

4. A relationship is shown between global mean HFA and *climate sensitivity* (CS), a statistic representing the estimated extent of warming that will occur in a model when CO_2 concentrations are doubled. Simulations with higher values of CS tend to have strong negative global mean heat fluxes (less than $-10W/m^2$). This relationship is potentially important for interpreting simulations with very high values of CS (greater than 8 degrees Celsius).

5. Relationships between HFA and model drift are investigated with the use of global mean and refined local statistics. No discernible pattern between global mean HFA and model temperature drift is found. The same model version can produce simulations that either drift or do not drift, suggesting drift is not dependent solely on parameter perturbation.

1.7 Chapter 6: Initial Condition Ensembles in Climate Modelling

1. The availability of a large Initial Condition Ensemble allows for a quantification of the HadSM3's internal variability, which is shown to be significant on length scales relevant for impact studies and adaptation decisions. The various roles of ICEs are discussed and their increased use is encouraged. It has typically been assumed that the effect of perturbing Initial Conditions on climate simulations was negligible Tebaldi & Knutti (2007).

2. It is shown here for the first time that Initial Condition perturbation can have a significant effect on model behaviour on relevant length and time scales in temperature and precipitation. The sign of the change in 8 year mean seasonal precipitation under a doubling of CO_2 is unanimous in only $\sim$ 3% of grid boxes. In temperature, 8 year mean seasonal differences within an Initial Condition Ensemble are shown to be as large as 10 degrees Celsius in some grid boxes. Such large differences are not usually considered possible and could affect the experimental designs and the interpretation of model variability. The effect of perturbing Initial Conditions is explored using bounding boxes (the maximum, minimum and range of values across an ensemble) and democracy plots (each member of an ensemble is given a vote and the number of votes counted).

1.8 Chapter 7: Constraining New Results from the CPDN grand ensemble

1. The range of behaviour shown in an ensemble of 45644 of GCM simulations is unprecedented, with estimated CS ranging from from 0.9 to over 16 degrees Celsius. Uncertainties are even larger on sub-global length scales.

2. Three methods are presented, as examples, to constrain the range of climate simulations and these methods are discussed in light of statistical good practice. These methods are based on **1)** Constraining the values for a key parameter, the Entrainment Coefficient, **2)** Global mean heat flux adjustment and **3)** Observational constraints in 7 variables. It is shown that for methods **1)** and **2)** simulations with CS of over 8 degrees can not be ruled out and that, for method **3)**, the distribution of CS is dependent on the choice of variable.

3. Simulations are selected based on the three different values of the Entrainment Coefficient (the most significant parameter perturbed in the CPDN experiment) to look at the effect on the distribution of simulated CS. The Entrainment Coefficient is shown to have an effect on the distribution of simulated CS but simulations with over 8 degrees CS exist for low, standard and high value of this parameter. For method **3)**, applying the constraint in multiple variables rules out more simulations, less than 0.1% of simulations pass the test applied in 7 variables simultaneously, whereas typically 10% pass in any individual variable.

4. It is shown that the shape of the distribution of CS is not an inevitable feature resulting from an approximately Gaussian distribution of feedbacks, as was suggested in Roe & Baker (2007). The distribution of CS can be changed substantially by a different choice of experimental design.

5. The relationship between strong negative global HFA and high CS simulations might be used to constrain the distribution of simulated CS. This can be done by considering a sub–set of simulations with a global mean HFA of magnitude less than the largest global mean HFA used in the standard HadSM3 ensemble. Global mean HFA can be used to change the distribution of CS, but simulations with CS over 8 degrees are still admitted. The use of such post-hoc filters is criticised on the basis of bad statistical practice.

6. When comparing model performance in–sample in 7 different variables, results depend on the choice of variable. For example, constraining in temperature tends to admit more high CS simulations and constraining in precipitation more low CS simulations.

1.9 Chapter 8: On the relevance of Model Means for Decision–Support

In Chapter 8, three sets of simulations are analysed, with global mean temperature rise of 2, 3 and 4 degrees respectively; it is shown that

1. Regional changes can differ significantly (over 6 degrees Celsius for some regions in 8 year mean seasonal temperature) for simulations with the same global mean temperature change.

2. The space and time scales of model diversity are quantified. The magnitude of regional uncertainties for a given global mean temperature change varies with length scale and variable. For example the range of DJF 8 year mean temperature change for simulations with 2 degrees Celsius global mean temperature rise rise is (0.96, 3.00) degrees in Australia and (1.32, 6.31) degrees in Northern Europe. The distribution of regional change is used to present diversity in sub–global response on a variety of length scales - global, hemispheric, tropical and extra-tropical, regional and local. Uncertainties in model precipitation response to doubled CO_2 are large – the sign of change is uncertain on length scales as large as many nations in most regions looked at.

3. The distributions of regional change are contrasted between the 2 and 3 degree global mean temperature sets. The magnitude of overlap between these distribution is shown to be large; in some regions and variables this overlap is over 20%, indicating that it is not always possible to define a unique relationship between global and regional changes.

4. Even if global mean temperature is constrained to within 0.2 degrees Celsius, significant regional uncertainties would remain. It follows from this result that global mean constraints are of limited use and that methods based on the patterns of change might be preferable.

At the end of each Chapter new results or methods will be listed as bullet points.

Chapter 2

Uncertainty and the use of State–of–the–Art climate models in decision–support

2.1 Overview

This Chapter looks at the problem of climate prediction and some of the challenges faced in using State–of–the–art (SOTA) climate models for decision support. The term SOTA models is used in this Chapter to refer to the set of models that are currently used to make climate predictions. In the context of the applications presented in this Thesis, SOTA models relates to GCMs, although this need not be the case (the SOTA necessarily changes over time). A climate prediction is defined here as a statement regarding the the future of Earth's climate system on time–scales typically of 10+ years. Particular focus is placed on the uncertainties inherent in climate prediction and possible methods to establish SOTA models as potentially fit for the purpose of decision support.

Section 2.2 explains the problem of climate prediction. The inherent uncertainties in climate prediction are qualified into four categories; Forcing Uncertainty, Initial Condition Uncertainty, Model Uncertainty and Model Inadequacy. Difficulties in the evaluation of climate predictions are discussed. The use of ensembles in the evaluation of these uncertainties is explained in Section 2.3.5.

Following on from the difficulties in establishing climate models as fit for decision-support, Section 2.4 presents three tests for consistency of information in climate prediction. These tests are not expected to prove that SOTA climate models are useful in a particular case, rather to show one some of the ways decision support can be sensitive to inherent uncertainties in climate simulations. The consequences of failing one of these tests and subsequent uses of a model whose predictions fail

are explained.
Section 2.5 discusses the use of climate model output in decision–support. Section 2.6 looks at the different needs of decision–makers in the context of adaptation to or mitigation of climate change. It is argued here that adaptation decisions often require more information and are subject to greater uncertainty than mitigation decisions. Section 2.7 looks at the communication of uncertainties and why transparent and effective communication is vital both for decision makers and providers of climate predictions.

2.2 The problem of climate prediction

There are a number of obstacles to accurately predicting future climate. Two of these obstacles are:

1. The Earth's climate system is highly complex and exhibits *non–linear* behaviour. This is also true of the SOTA models used to simulate the Earth's climate system. The presence of non–linearities in a complex system mean that precise prediction is impossible in practice due to inexact specification of the initial state (the starting state of a model simulation) or observational noise Judd & Smith (2001a); Lorenz (1963); Smith (2007). This is apparent in the case of *Numerical Weather Prediction* (NWP), where uncertainty in the initial state results in a loss of predictive skill typically after 2 weeks Orrell *et al.* (2001). Since exact deterministic prediction of either weather or climate is impossible, a distribution of climatic states should be forecast to represent this uncertainty. Obtaining this distribution is the goal of climate prediction Palmer *et al.* (2005); Smith (2002); Stainforth *et al.* (2007a).
   The evaluation of the distribution of future climate faces a number of difficulties. One such obstacle is that small perturbations to the climate system can result in disproportionately large (or small) effects on the distribution of future states due to *feedback* processes. Interactions between components of a non–linear system are dependent on each other and so the system can not be considered simply as the sum of its parts. In order to understand the important interactions within the climate system, complex models have been developed that can not be readily understood analytically. Since an analytical approach to understanding climate models is not possible, it is necessary to gain insight to the workings of the models from computer simulations (running the model and looking at the output).

Since there are non–linear *feedbacks* in the Earth's climate system, it is impossible to know whether some as yet unconsidered process or forcing will change future climate in an unexpected way. Thus, any climate prediction is conditional on the absence of any undiscovered feedbacks that would significantly affect its conclusions. Therefore, it is important to note that no climate prediction can be final but might be updated in light of future model development or an improved understanding of the climate system. Whilst this is true of scientific models in general, the conditional nature of climate predictions is particularly prevalent due to the inherent nature of climate predictions as extrapolations.

2. The climate system changes over time, thus any attempt at prediction must not be critically dependent on the assumption that the future will resemble the past; climate prediction is fundamentally a problem of extrapolation. The climate system is being altered due to anthropogenic influences with a potential magnitude of climate change greater than any present in *observational* data. Since the future climate is expected to be outside the range of data available, attempts at prediction are dependent on the dynamics of the change in climate occurring in a way consistent with our current understanding of climate science. A key motivation for the use of physical models in climate prediction depends on our ability to understand the physical processes driving future climate change Solomon *et al.* (2007a). This approach assumes that, having captured all the important drivers of climate change from past data, it is possible to extrapolate since climate will continue to react to forcings in a similar way.

The non–linearity of the Earth's climate system, coupled with a lack of analogues in the observational record, make climate prediction difficult. These problems mean that climate predictions are most relevant when the uncertainties in predictions can be reliably estimated. As explained in Section 2.4, some common methods of model evaluation are not possible for long–term climate prediction. Alternative methods that can be used to qualify and quantify the uncertainty in climate predictions are presented.

There are some important aspects of climate prediction that require the judicious application of statistical good practice. One example of this is that climate models are so complex that they can not be considered parsimonious in any conventional sense. With hundreds of tunable parameters, and perhaps a little over 100 years of reliable observational data, it might be argued that it is not surprising that models can re–produce the past well. In fact, it would be surprising if a model with more

tunable parameters than data points would not produce a good in–sample fit, unless parameter values were heavily constrained. Evaluation of climate predictions can not be definitive and should take the form of tests for consistency of information. These tests, described in Section 2.4, provide a means of sanity–checking climate predictions as potentially fit for purpose. Section 2.3 introduces the four categories of uncertainties and how these might be explored using ensembles.

2.3 Uncertainties and Ensembles

The evolution of future climate is highly uncertain for several reasons Stainforth *et al.* (2007a). Despite the huge research investment and fundamental physical theory that goes into SOTA climate models, it is an open question to what extent decision–relevant information can be extracted from these models. In order to answer this question it is necessary to understand the different types of uncertainty present in SOTA models; some are reducible, others not Smith (2002).
There are a variety of different uncertainties in climate prediction resulting from the problem of extrapolating using *dynamical systems*. These uncertainties can be classified in different ways Giorgi & Francisco (2000); Stainforth *et al.* (2007a). The four categories of uncertainty used here are:

1. Forcing Uncertainty
2. Initial Condition Uncertainty
3. Model Uncertainty
4. Model Inadequacy

Forcing Uncertainty is not dealt with in detail in this Thesis, because it does not relate to how climate models represent the climate system itself, and depends on political decisions and unknown natural forcings. Errors in the model's representation of the system being modelled are divided into three categories; Initial Condition Uncertainty, Model Uncertainty and Model Inadequacy. The four categories are now discussed in turn.

2.3.1 Forcing Uncertainty

Factors that affect the distribution of climate, not arising from internal variability of the system[1], are known as *forcings*. Natural forcings include changes in solar

[1]Internal, or "Natural", variability usually means variations in climate not due to some external forcing factors, such as anthropogenic effects.

luminosity and volcanic eruptions. Volcanic eruptions are an important driver of climate variability on timescales of months to years are volcanic eruptions, leading to significant drops in regional and GMST Robock & Oppenheimer (2003); Yang (1999). We might expect our models to be inaccurate in the future at least to the extent that they do not take into account the potential effect of large volcanic eruptions and other forcings we know that are unaccounted for.

Whilst naturally occurring forcings have an effect on the climate system, the most important changes in forcings for the climate of the 21st century are anthropogenic Cubasch *et al.* (2001); Stott & Kettleborough (2002). Anthropogenic forcings include the emissions in GHGs, such as CO_2. In climate models the analysis of forcings is often restricted to Greenhouse Gases (GHGs), or GHGs as CO_2 equivalent. The effect of forcings can be measured in terms of their radiative effect (in W/m^2).

Uncertainty in what the future forcing will be depends, in part, on policy and the actions of mankind over the next 100 years. There are large uncertainties in the emission of GHGs and other anthropogenic sources Nakicenovic *et al.* (2000). The uncertainty in future anthropogenic forcings is, partially, in the control of global policy and is not an issue for climate science, as such. Climate models can only hope to provide insight into each forcing scenario under consideration, given the other forms of uncertainty so that policy can be made in a more informed manner. In so far as emissions are policy–dependent, Forcing Uncertainty is partially controllable since mankind can choose what level of GHGs to emit over the course of the 21st Century. Unlike mitigation decisions, for many decisions relating to climate impacts or adaptation, Forcing Uncertainty must be included, as discussed in Section 2.6.

Assessing the likelihood of various forcings is not addressed in this Thesis. This Thesis will treat the problem of climate prediction as conditional on a scenario of future forcings.

2.3.2 Initial Condition Uncertainty

Climate models need to be initialised with some starting state. Where the model and the system are identical (this is also known as the Perfect Model Scenario Judd & Smith (2001a)), this might be done by using the most accurate estimates of the system's current state. Given that no observations are perfectly accurate, there will be a set of possible starting states in whatever high–dimensional space the model lives in. Propagating this set of possible starting states using a model results in a distribution of possible states at each time point in the simulation. This distribution can be thought of as a manifestation of Initial Condition Uncertainty (ICU).

With a Perfect Model and the ability to fully explore ICU, the distribution of future climate would be representable by a reliable Probability Density Function (PDF)[1] where the model's PDF provides a fully accurate guide to the frequency of real world events. Fully propagating ICU (every possible Initial Condition) in a climate model gives the model's PDF, also known as the model's *climatology.*

It is a fundamental property of non–linear systems that simulations can show sensitive dependence on initial conditions (ICs). In systems displaying sensitive dependence on ICs, even arbitrarily close starting states diverge on long time scales Lorenz (1963). An example of the growth of ICU can be taken from Numerical Weather Prediction; there are fundamental limits to the predictability of weather given noisy observations since very similar current states can lead to widely divergent forecasts on timescales of a week or more.

In the case of climate, the initial growth of ICU is not of interest *per se*, since interest lies in the behaviour of the model on timescales beyond which ICU is thought to affect the distribution of model behaviour. Reducing ICU should have no significant effect on the model climate distribution on long time scales Stainforth *et al.* (2007a). The presence of ICU in climate modelling means that we are always dealing with a distribution of climatic states. Climate modelling aims to understand how this distribution will change; weather modelling aims to provide forecasts on short lead–times, conditioned on the estimated current state of the climate system. Where the model and the system are distinct, there is no unique correct model starting condition. Yet all models must be initialised in order to be run. A model can be initialised using an *analysis* (observations projected into model space through a model) although there would still be a set of different Initial Conditions with which the model could be equivalently initialised. This set of possible initial states, when transformed into model space, need no longer directly represent the uncertainty in observations. It is critical to understand that a climate model's PDF may not bear any useful resemblance to future climate. It is the role of climate scientists to evaluate the potential similarity between climate models' PDFs and the future climate. ICU is analysed in detail in Chapter 6.

2.3.3 Model Uncertainty

There is also uncertainty in how to represent our knowledge of the climate system. This uncertainty is present both in the structure of the model and its *parameter*

[1] A PDF is reliable when the event occurs with a frequency consistent with the value indicated by the PDF Bröcker & Smith (2007a)

values. The case of parametric uncertainty is discussed first, where the *model structure* is held constant. Unlike ICU, changing parameter values in a dynamical system can lead to dramatically different behaviour Cuellar Sanchez (2006); Sprott (2003). There are a number of different, plausible, values that each parameter can take. Furthermore, parameter values may not be fixed and can vary over time Kennedy & O'Hagan (2001b).

For many (if not all) parameters in a SOTA climate model, there are a set of values that might be used, each potentially resulting in a different set of model dynamics. An example of such an undetermined parameter is the drag coefficient, relating to the frictional retardation of the atmospheric flow due to the roughness of the Earth. This parameter value is thought to be known empirically to about $\pm 10\%$ Thorpe (2005).

The same structural model with a different set of parameter values is called a *model version.* These different sets of parameters can be explored in a given model, in a Monte Carlo–type approach. How the propagation of parametric uncertainty should be done is an open question – the resultant distribution of model output depends critically on subjective choices such as: which parameters were perturbed, which intervals to vary the parameters within, the parameter sampling strategy, any prior distributions that might be used and how to interpret the model output Frame *et al.* (2005, 2007); Stainforth *et al.* (2007a). The effect of these prior choices for parameter perturbation differs from the case of perturbing ICs. Whereas perturbing ICs allows a sampling of a single distribution, changing parameter sampling strategies or other prior distributions changes the distribution of model output itself. Unlike ICU, model behaviour under different sets of parameter values does not sample from a common distribution; the model will behave differently for each different set of parameter values. Each model version has its own distribution with its own internal variability. There is therefore a hierarchy of uncertainty: ICU represents the distribution for a given model with certain parameter values and parametric uncertainty represents the different model versions that are possible for a given structural model.

When considering the uncertainty of how to represent our physical understanding structurally, it is clear that there is no way to sample objectively from "model space" Allen & Stainforth (2002). It is only possible to consider existent models rather than consider all the models that might potentially exist. It has been shown that the differences between parametrically perturbed model versions can be larger than the difference between structurally distinct models Solomon *et al.* (2007a); Stainforth *et al.* (2005). This suggests that the current range of different structural models are not fully representative of total Model Uncertainty.

The significance of Model Uncertainty is that there are a number of different model versions that can be used to predict future climate; differing either by parameter values or model structure. We do not know which model version, if any, will produce the most accurate forecast[1] for a particular variable and future time period.
It is important to note that the concept of a single "best" model or set of parameter values is irrelevant in the context of climate prediction since there are a number of different qualities that can be looked for in a climate model. In the case of model structure, there are a number of different ways of compiling the same physical understanding into a model. Furthermore, no single model is likely to be most useful in every way and it is not always possible to tell which model will provide the most useful predictions; thus a set of models must be considered.
The diversity of forecasts arising from different models forms a lower bound on the precision with which climate predictions can be made. Such Model Uncertainty can be explored using ensembles, as explained in Section 2.3.5.

2.3.4 Model Inadequacy

Every model is imperfect; this is true by a model's very nature. In a theoretical sense, a perfect model ceases to be a model and becomes a restatement of the system itself. In the real world models are inevitably imperfect. As such, we do not have access to a perfect climate model. In particular, complex computer simulation models are inadequate i.e. they are an incomplete or flawed representation of the system being modelled Chatfield (2002); Kennedy & O'Hagan (2001a); Oreskes *et al.* (1994); Smith (2002). Whilst there is no doubt that models are imperfect there is equally no doubt that some are useful. The relevant question is in what way their inadequacies render them useful and in which ways they are useless (or worse, misleading).
In the case of climate models, there are a number of known inadequacies. Despite their great complexity, there are still missing processes that are important for modelling climate change (e.g. atmospheric chemistry, the carbon cycle, vegetation models etc.). Many models do not have an explicit stratosphere or deep ocean (including HadSM3, the model used in the CPDN experiment). Furthermore, the grid resolution in current models is relatively coarse, leading to unrealistic simulation of important processes such as clouds and precipitation Dai (2006); Karl & Trenberth (2003).
Model Inadequacy is a major problem in the use of models to understand the climate. Unlike many other fields, such as NWP, the lack of out–of–sample data makes

[1]The most relevant measure of forecast accuracy will depend on the user.

it difficult to evaluate in which ways GCMs are adequate for use. Given a sufficient amount of relevant out–of–sample data it is possible to evaluate a model and understand its strengths and weaknesses. Systematic biases can be detected, models can be improved and the likely future out-of-sample skill of the model estimated. Where there is insufficient out–of–sample data, it is not possible to tell how the model will perform out–of–sample; but in–sample fit can provide a lower bound on future model accuracy, as explained in Section 2.4.

The problem of extrapolation in light of Model Inadequacy can be related to the story of Russell's chicken Russell (1946); Stainforth *et al.* (2007a). A chicken is fed each day by the farmer, and believes the farmer will continue his benign behaviour in the future, then one day is suddenly slaughtered. For the chicken such an event was unthinkable based on prior data. The problem has also been characterised as the black swan effect Hume (1748); Taleb (2008) (prior to the discovery of Australia it might never have been thought that there may be non–white swans.). The point of such examples is that it is never possible to know whether the future will resemble the past; indeed in the case of climate prediction we know that it will not.

Since we do not have the opportunity to compare climate model results to out-of-sample verifications, it is not possible to know how a climate model will go wrong. This problem is discussed in more detail in Section 2.4.

2.3.5 Ensembles

In the case of simple chaotic models, such as the logistic map May (1976); Sprott (2003), it is possible to evaluate some of the uncertainties categorised in Section 2.3 analytically. For example the growth of ICU over time can be understood mathematically from the model equations Sprott (2003). In the case of SOTA climate models, this is not possible; the models are far too complex to study mathematically and computations too laborious. The method most widely used to understand climate model behaviour involves using *ensembles* Collins & Knight (2007). In the case of climate modelling, an ensemble can be thought of as a dynamical system version of a Monte Carlo simulation. Ensembles come in a variety of types, four of the most important for climate modelling are discussed in this Section; ICEs, perturbed physics ensembles (PPE) Murphy *et al.* (2004); Stainforth *et al.* (2005), *multi–model ensembles* Solomon *et al.* (2007a); Tebaldi & Knutti (2007) and grand ensembles.

2.3.6 Initial Condition Ensembles

In order to evaluate ICU and understand the internal variability of a climate model, an ICE can be run. ICEs are formed by running the same structural model several times, with the same parameter values, with different starting states. In the case of the CPDN project the different ICs are formed by small perturbations about a common temperature field. An ICE enables a quantification of the distribution of a model's climate. ICEs are discussed in detail in Chapter 6.

Perturbed Physics Ensembles

Model Uncertainty can be explored using a Perturbed Physics Ensemble (PPE). A PPE consists of a set of model simulations, using the same structural model, but with different parameter values. Due to the large number of parameters in a climate model, the number of possible levels and combinations of parameter perturbations it is not possible to fully explore this type of uncertainty given finite computational constraints. In Murphy *et al.* (2004) 53 model versions were used to explore Parametric Uncertainty for 29 parameters chosen by climate modelling experts to be important. PPEs of this kind can give an estimate of Model Uncertainty but the particulars of the results inevitably depend on subjective choices, as described earlier in this Section. The problem of how to sample parameter space effectively has not been solved; linear factor analysis Murphy *et al.* (2004) and Latin hypercube designs have been used Annan & Hargreaves (2007). Other work, such as Sanderson *et al.* (2008), has shown how data from the CPDN ensemble can be used to inform choices of future parameter values. In particular, Sanderson *et al.* (2008) proposes that a neural net can be used to detect the likely choices of parameter values that would yield a wide range of model behaviour, hence a smaller ensemble could be used to achieve a similar range of behaviour than methods not informed by existing perturbed physics ensembles. These methods have been chosen since they require far less computational resources than fully exploring parameter space.

Multi–model Ensembles

A multi–model ensemble is a set of simulations made using a set of different structural models. There are at least 20 different modelling centres that have developed GCMs; the range of results produced is a guide to the diversity present in how current physical understanding is represented. These models share many similar properties and should be thought of as highly inter–related. Furthermore, we can not know how the diversity of GCMs currently available reflects uncertainties in the structure of the model itself. The range of current SOTA climate models could be

treated of as a lower limit on uncertainty in the model structure. A multi–model ensemble is analysed at in Chapter 3.

Grand Ensembles

A *grand ensemble* is a set of ICEs run under different structural models or model versions. In the CPDN experiment, a number of ICEs are run under a number of different model versions where parameter values are varied. Grand ensembles allow a comparison of model behaviour, allowing for more sophisticated uncertainty assessment than single simulation a multi–model or PPE would allow. In a grand ensemble the internal variability of each model used can be evaluated and used to compare model behaviour. Since a comparison of climate distributions is relevant for robust decision analysis grand ensembles are an important area of current research. The CPDN grand ensemble is introduced in Chapter 4 and is studied in Chapters 7 and 8.

Duplicate Simulations

A duplicate simulation is an exact copy of an another simulation. Therefore, in theory, the results obtained from duplicate simulations should be identical. Duplicate simulations can be used to verify certain aspects of experimental design. Due to differences in computing architecture, processor or numerical errors in computation there can be differences between duplicate simulations in practice Knight *et al.* (2007). Evaluation of these differences is useful so that other sources of variability and uncertainty might be attributed.

2.4 Model Evaluation and straw–men for climate

Prior to using model output to inform decisions, it is statistical good practice to evaluate for which purposes the model is fit for use. This poses the question: "How can a climate model be evaluated given that it is extrapolating decades into the future?". As previously stated, out-of-sample comparison of model output to observations is not accessible in the case of climate prediction Reichler & Kim (2008). Often models are use to make predictions on lead times of 10+ years and have a working life of about 5–10 years. Once they are updated with a new model the old model is no longer studied in detail. A notable exception to this is Hansen *et al.* (1988, 2006) in which an older climate model is checked against out–of–sample data for the years 1988-2005. Even where observations are available, only a single time series can be used as a verification, placing limits on our ability to ever assess the

probabilistic reliability Bröcker (2005) of models. By virtue of their very newness, current SOTA climate models are unevaluated out–of–sample. Other methods that do not rely on out–of–sample data for evaluation must be used.
Whilst it is not possible to know that an extrapolation of the future state of the climate system is justifiable Frame *et al.* (2007); Hume (1748); Oreskes *et al.* (1994), it is possible to assess whether certain predictions are likely to be unreliable. This "predictive falsification" can be achieved in a number of ways; three such *straw-man* tests are presented in this Section. Failure in any one of these tests is tantamount to considering those predictions as decision–irrelevant, although it is expected that a model might be able to pass all these tests with respect to one decision and not another. Passing all tests does not mean that the model is useful; these tests do not verify a model but should be thought of as sanity checks. Each test will be presented in the following format; **1)** the rationale behind the test is first introduced, **2)** the test itself presented, **3)** the justification for the test and **4)** limitations of the test.

2.4.1 In–sample fit

The degree to which a model can re–produce past observations, in–sample, provides a lower bound on its ability to predict out–of–sample (future forcings are unknown and models can not be over–fit out–of–sample as they can be in–sample). Whilst accurate in–sample performance does not imply useful out–of–sample performance it is a requirement that a model can produce *skillful* simulations in–sample, if the model is to be useful out–of–sample. Such in–sample fit should properly be assessed using an ICE of simulations that take account of the model's internal variability. The ability of a set of climate models to re–produce the GMST time series of the 20th Century is examined in detail in Chapter 3.

- The Test: In the variable(s) of interest, evaluate the model's in–sample fit for the predictor chosen e.g. 10 year mean August temperature in Southern England. By comparing model output to observations using a relevant measure of model skill, a limit can be set on likely out–of–sample performance. Model output should comprise a set of simulations, preferably including ICEs to include the effects of each model's internal variability, whereas the observations will comprise a single time series. The measure of model fit should be relevant to the end–use of the model. Methods to evaluate in–sample model performance include the use of *bounding boxes* Weisheimer *et al.* (2004), shadowing the observations to within observational uncertainties Judd & Smith (2001a,b); Smith (2001), or by using some proper skill score in the case of probabilistic model output Bröcker & Smith (2007b).

- Justification: The consistency of model behaviour with observations in–sample is a lower limit on its consistency with out–of–sample verifications. If it can be shown that the model does not produce simulations consistent with observations when the verification is known prior to making the prediction, it is difficult to see why the predictions should be more useful where the verifications are not known.

- Limitations: This test can only provide a lower limit on the accuracy of a model's predictions but can not set an upper limit. Thus, it can not be directly inferred from passing this test that the model is producing the right in–sample fit for the right reasons. An example where this test would be less relevant would be if the model being tested is over-fit on the available data, giving a misleadingly close in–sample fit.

2.4.2 Initial Condition Test

The magnitude of uncertainty seen across an ICE provides a lower bound on uncertainty in that simulation, as explained in Chapter 6. ICU can be quantified in a climate simulation and used as a straw–man as a test for robustness of model predictions.

- The Test: Run a large ICE over the period of interest. By examining the range of behaviour using the distribution of ICE members on various length scales and variables it is possible to judge whether it is possible to inform a particular decision e.g. if an ICE disagrees on the sign of precipitation change on all relevant length and time scales it would be dangerous to use such a model to inform decisions critically dependent on the sign of precipitation change. A user–specified level of uncertainty could be set before conducting this test if used to reject model predictions as failing to provide consistent information.

- Justification: ICU is an irreducible source of uncertainty Stainforth *et al.* (2007a), that represents the internal variability of the model. ICEs provide a means to evaluate this type of uncertainty and reflect the robustness of model behaviour.

- Limitations: Whilst this test provides an irreducible lower bound on uncertainty, this type of uncertainty taken alone under–estimates the full uncertainty present in climate predictions – this test may not be very powerful.

This test becomes more powerful as the number of members in the ICE increases; as more uncertainty is explored, the test becomes more stringent, although there is likely to be a saturation in the value of adding extra IC members after a certain size of ICE. As such, passing this test can not provide much confidence in a climate prediction but it can rule out predictions that are subject to high levels of internal variability.
Determining the critical level of ICU is subjective and must be decided on a case by case basis, preferably with consultation with the decision maker.

This test is discussed in the context of regional simulations of temperature and precipitation in Chapter 6.

2.4.3 Model diversity Test

There are a number of different models that can be used to make climate predictions, as stated in Section 2.3. Since any of these models might be used to make climate predictions, the diversity of results across available models should be assessed as a lower bound on Model Uncertainty. Model Uncertainty can be thought of as arising from the existence of a set of possible models. This type of uncertainty is discussed in more detail in Chapter 7 which the results of the CPDN grand ensemble are presented.
This test is a more stringent version of the previous test using ICEs. The basic principles are the same, but since it is expected that the diversity of different models' predictions will be greater than the magnitude of ICU this test will likely be more powerful than the test based on ICEs.

- The Test: Run a multi–model ensemble over the future period of interest, where models differ by either parameter value (as in a PPE) or model structure. The diversity of behaviour across models can be used to assess the magnitude of uncertainty arising from Model Uncertainty using a bounding box methodology. A user–defined level of uncertainty could be specified before conducting this test if used to reject models.

- Justification:

  The range of model output represents the range of possible predictions that could be obtained given a set of models and possible parameter values. Where this range is very large (too large for useful decision support), it is not possible to provide useful predictions without further information.

- Limitations:

  The diversity of climate models is limited in the sense that it is only possible to run models that are available; these models are not independent and represent an ensemble of opportunity rather than an objective spanning of model space Allen & Stainforth (2002).
  It is necessary to show some judgement in what constitutes an acceptable model – it is possible to deliberately include models with a particularly low (or high) level of diversity. As a possible guideline, only models whose predictions which pass the first two tests described in this Chapter should be used in this test i.e. only models that provide adequate in–sample behaviour and for which ICU is not too great in the variables, length and time scales of interest.
  In order for this test to be most powerful, as many and as different models as possible should be used. The use of this test is discussed in Chapter 7 in the context of a large grand ensemble of simulations.

2.4.4 Utility of models that fail these tests

If a particular model prediction fails to pass one or more of the above tests, this model's output, in the variables tested, is inadequate for quantitative use in decision support. That is not to say that the model itself is useless, nor should a failure to pass a test be seen as a purely negative result. Four uses of a model that fails one of the above tests are presented here.

1. If a model prediction fails a test, it does not mean that the model as a whole is invalid. It could be that a model will provide more robust predictions in some variables and length scales than others e.g. failing to correctly simulate *local* seasonal precipitation does not necessarily mean the model is useless for predicting global annual mean temperature.

2. It is advantageous to know the cases in which predictions are highly uncertain. This could save considerable expense on developing, interpreting or purchasing detailed model output. Furthermore, decisions based on over–confident information are unlikely to be optimal.
   In the context of decision–support, it may be very important to remain flexible rather than making a decision using pre–mature or incorrect information. To act on climate predictions known to be unreliable risks a considerable over–commitment on the decision-makers' behalf and a loss of reputation and trust on the behalf of the provider of climate predictions. It should be noted that

the lifetime of a prediction, and the rate at which confidence in climate science would be lost, is not necessarily on the same timescale as the predictions themselves. Rather, loss of confidence in climate models' predictions might occur as soon as a new model, or set of predictions, provides conflicting evidence. Where it is believed that model predictions will change with future model improvement, it would be misleading to present them as final. This concept is related to the idea of Stable Inferences from Data Allen & Stainforth (2002), in which it is proposed that we should have more trust more in those aspects of model behaviour that are robust over different models and over time than in those that are prone to change.

3. The model can be still be used heuristically or for the purposes of scenario generation. Whilst the model output may not be directly usable, quantitatively, in the context of a particular decision, model output can help in conceptualising a problem. Using the model to generate various scenarios can be helpful even if such output will not play an explicit quantitative role. A climate model may suggest avenues of investigation not previously considered despite its inability to provide robust predictions. Furthermore, it has been suggested that an invalidated (or yet to be evaluated) model can be used to encourage appropriate data collection, or as an important illustrative device Hodges (1991). A model that provides no conclusive quantitative evidence will support policy that is flexible and responsive. A similar point is made in Allen & Frame (2007), in which it is argued that we can adapt mitigation policy to global warming as we observe it to occur. Such "wait and see" strategies do not exclude pre–emptive mitigation, depending on the decision–makers risk preferences and the relative costs of under or over–shooting mitigation targets. An uncertain model would add weight to the argument for adaptive climate policy since it may prove costly either to over or under commit on emissions targets (or on local adaptation plans) based on current evidence.

4. Model output is useful in understanding the current state, and progress of, climate science. Using models to diagnose courses of future research and model improvement is an essential use of models, regardless of whether they can provide decision makers with useful predictions. In fact, detecting model failures is key to model improvement. Whilst today's models may not provide robust predictions, it is possible that the next generation of models will. In inter–disciplinary areas such as climate research, model output provides an opportunity to assess the state of the science and inter-compatibility of the various contributing areas.

Based on the above four points, it would be wrong to interpret the failure to provide direct quantitative evidence as a reason to ignore SOTA climate models. Significant qualitative understanding can be gleaned from models failing tests for consistency of quantitative information as well as a basis for further scientific progress.

2.5 Decision support

This Thesis presents results that are aimed at informing decisions and climate change policy. Any attempt to use model output to inform forecasts of climate change requires some ability to translate information in model simulations to information in the real world. Otherwise, model results will remain in "model land" and do not relate to Earth's climate or contain any direct relevance to real life decisions. In order to improve the decision–relevance of climate predictions it is important to work with the users of climate model output. Through partnership with decision makers, climate experiments can be made more relevant and can be interpreted in a more practical way. A framework for an iterative process between the providers and users of climate model output has been suggested in Stainforth *et al.* (2007b). As a first step towards evaluating the decision–value of climate models it is important for decision makers to determine how climate affects their decision and to frame questions in terms of statistics that might be generated by climate models. If a particular decision is robust to the uncertainties present across ensembles of climate model simulations climate model output might be useful. It is likely that decision makers will not have access to all the information they require and will either have to settle for uncertain and provisional model results or not to use climate model output at all. It is important here that the uncertainties inherent in climate prediction are communicated effectively in order to differentiate the varying degrees of confidence that should be attached to different aspects of model output.

An important question is how to represent model output to users. Various suggestions have been made including giving the range of model diversity Stainforth *et al.* (2007b), probability density functions Jenkins *et al.* (2007) or a more complex Bayesian framework Goldstein & Rougier (2006). It is an open question how best to communicate the information in imperfect models; other techniques are also being suggested include the use of probabilistic odds instead of probabilities Judd (2008a). In order to assess which methods should be pursued it is necessary to consult the users of climate model results so that the most relevant and understandable method is applied that can transmit uncertainty information.

In this Thesis, model output is not interpreted as probabilistic statements about future climate, rather as giving insight to the workings of the climate model(s) used.

2.6 Uncertainties in Adaptation and mitigation decisions

Two important types of climate policy that might be informed by climate predictions are **1)** adaptation and **2)** mitigation. Adaptation can be seen as changes that are made to reduce the costs or exploit the benefits of a changing climate. Adaptation decisions are subject to all four types of uncertainty explained in Section 2.3, including forcing uncertainties. In order to make effective adaptation decisions it is important to know not only the sign of climate change, but the possible magnitude of the change i.e. when deciding how high to build a flood barrier, it may be insufficient to know only that it will rain more in the future; it is also necessary to know by how much precipitation will change and statistics on how the magnitude and frequency of precipitation events will change. It is possible to make adaptation decisions based on the relative risk of climate changes e.g. the risk of flood waters exceeding a certain level, rather than planning for specific changes. Using the relative risk of an event of interest has the advantage that it treats future climate as a distribution considering a range of possible events.
Mitigation decisions target reductions in the extent of future climate change. Mitigation decisions can take a different approach to uncertainties to adaptation planning. Two differences in the treatment of uncertainties between mitigation and adaptation decisions are:

1. Since the most important drivers of climate change are themselves the subject of mitigation decisions, Forcing Uncertainty is, in part, reducible. Whilst there are sources of Forcing Uncertainty that are not dependent on decision–makers (such as variations in solar luminosity, or volcanic eruptions), to a large extent the decision–making process relies only on providing robust predictions conditional on a particular mitigation decision. Based on such conditional predictions, usually of the form of scenario analysis Nakicenovic *et al.* (2000), decision–makers can influence the future path of climate forcings.

2. Since GHGs and other important forcings are global in their effects, mitigation is largely a global issue whereas adaptation is a local one. Thus, mitigation decisions are not dependent on local climate changes in the same way adaptation decisions are. The uncertainties associated with global means are typically smaller than for regional means Solomon *et al.* (2007a) (shown in Chapter 8) and mitigation decisions need not always consider the distribution of specific costs and benefits on local scales. On the other hand, it is important to understand the local impacts of climate change in order to estimate

the likely costs of a given level of GMST rise. Thus mitigation decisions can not exclude regional variations in climate response in the case of deciding economically optimal policy but do not depend on this information as critically as adaptation decisions.

2.7 Communication of uncertainties

It is important for climate scientists to communicate the uncertainties in climate predictions transparently. Despite the importance of such communication, there have been few attempts to provide decision-makers with comprehensive information on the uncertainties in climate predictions. An example of the presentation and communication of model output for decision-makers will be discussed in detail in Chapter 3, taken from the IPCC AR4 Solomon *et al.* (2007a), where the performance of SOTA models in simulating the time series of GMST over the 20th Century is analysed in detail. This example will highlight the need for robust and relevant model evaluation methods and provides motivation for investigating more relevant methods of evaluation of climate predictions.
It is important not to mislead decision makers as to the uncertainties of climate prediction because:

1. Decision–makers are likely to make worse decisions if they are not informed of the full range of possibilities or are overconfident in the predictions of climate models.

2. Climate scientists risk losing credibility if the uncertainties in current models are not fully disclosed and the next generation of climate models produce different forecasts. Such loss of credibility could be irretrievable and it will take a very long time for sufficient verification data to be obtained in order to establish new models as trustworthy. Exposure of misrepresentation of uncertainty could be used by skeptics of anthropogenic global warming to cast doubt on the better understood aspects of climate change.

There remain issues of how best to communicate uncertainty to decision makers. An important issue is that many decision–makers lack the quantitative background to understand technical aspects of climate statistics or model output. Decision-makers would benefit from an understandable communication of uncertainties. It is argued in this Thesis that the best way to do communicate climate science is to present the uncertainties in climate prediction as fully as possible. Over the past 20 years, due to increased research in climate science, there has been a great increase in

our understanding of the climate system; this knowledge has decreased uncertainty in whether observed warming is anthropogenic but has increased our uncertainty in its magnitude Andronova & Schlesinger (2001); Stainforth *et al.* (2005). It would be wrong to interpret uncertainty as incompatible with knowledge, but rather as a result of critical and honest evaluation of our understanding.

2.8 Conclusion

The problem of climate prediction has been presented together with a categorisation of the uncertainties involved. Such categorisation is important since the four types of uncertainty discussed in this Chapter require different treatment: Forcing Uncertainty is reducible in mitigation decisions, Initial Condition Uncertainty (representing internal model variability) is quantifiable using ICEs, a lower bound can be placed on Model Uncertainty using perturbed physics or multi–model grand ensembles and an assessment of Model Inadequacy is only accessible with sufficient comparison to observations (preferably out–of–sample).
Communication and transparency of uncertainty is argued to be of key importance if climate science and science–based policy are to follow a successful and mutually beneficial partnership. This Thesis aims to motivate a more complete understanding of uncertainties in the light of statistical principles of good practice.
New methods presented in this Chapter are:

- Three methods to evaluate the consistency of information in climate simulations have been proposed using in–sample data, ICEs and multi–model ensembles. These tests allow the providers and users of climate output to weed out predictions with no robust predictive skill.

- The possible uses of quantitatively inadequate models (including models that fail one or more of the tests presented) has been described. Failing one of the tests presented has been shown to be a potentially useful result in itself.

Chapter 3

How reliable are the models used to make projections of future climate change?

3.1 Introduction

The main models used to provide projections of climate today are GCMs, some of the most complex models ever built. For these projections of future climate to be used in decision–support, it is important to evaluate GCMs as fit for purpose. As discussed in Chapter 2, this evaluation is non–trivial due to the nature of the problem (long-term extrapolation of a complex, physical system) and a lack of relevant past analogues with which to compare model output.

This Chapter looks at an example of the evaluation of some of the climate models used in the recent IPCC AR4. In particular, the ability of climate models to reproduce the observed GMST time series over the 20th Century is looked at in detail. It is shown that the presentation of model output in the AR4 is both limited and misleading in its reflection of model performance. This example motivates a more comprehensive analysis of climate models' likely predictive *skill*.

Three criteria for the reliability of GCMs are presented in the IPCC AR4 (e.g. FAQ 8.1 in Working Group 1);

1. "One source of confidence in models comes from the fact that model fundamentals are based on established physical laws"

2. "A second source of confidence comes from the ability of models to simulate important aspects of the current climate... Models' ability to reproduce these and other important climate features increases our confidence that they repre-

sent essential physical processes important for the simulation of future climate change"

3. "A third source of confidence comes from the ability of models to reproduce features of past climates and climate changes ... One example is that the global mean temperature trend over the past century can be modelled with high skill when both human and natural factors that influence climate are included."(pages 600–601, Chapter 8, Solomon *et al.* (2007a)).

This Chapter focuses on the evidence for criteria **3)** - the in–sample performance of a set of GCMs in re–producing the time series of 20th Century GMST. The results presented are also relevant for criteria **1)**. It should be noted that to accurately simulate observed GMST is insufficient for many decisions Smith *et al.* (2008); Stainforth *et al.* (2007b). Nevertheless, the extent to which GCMs can re–produce observed 20th Century GMST in–sample places a lower bound on their ability to inform decisions on finer spatial scales and in other variables Smith *et al.* (2008). Furthermore, since models are tuned in–sample Bender (2008); Johnson (1997); Oreskes *et al.* (1994); Stocker (2004) the magnitude of in–sample residuals provides a lower bound on the potential out–of–sample accuracy, as explained in Chapter 2. The IPCC AR4 presents a plot of in–sample GCM simulations from 14 modelling centres around the world versus observed GMST. This plot appears in two different cases:

1. In Frequently Asked Questions 8.1 – "How reliable are the models used to make projections of future climate change?". The plot is used to suggest that GCMs produce a reliable representation of the time series of 20th Century GMST.

2. The plot also appears in Chapter 9, Figure 9.5, in a comparison of model's in–sample skill with and without anthropogenic forcings. This plot is used to show that GCMs can only match observations when anthropogenic forcings are included than using only natural forcings – "Figure 9.5 shows that simulations that incorporate anthropogenic forcings, including increasing greenhouse gas concentrations and the effects of aerosols and that also incorporate natural external forcings provide a consistent explanation of the observed temperature record, whereas simulations that include only natural forcings do not simulate the warming observed over the last three decades." (p. 684, Solomon *et al.* (2007a))

The presentation of model output in cases such as the IPCC AR4 Figures 8.1 and 9.5 is questioned here in the context of the reliability of climate models' projections. The IPCC Figure 8.1 is limited in terms of case **1)** demonstrating the reliability of models' projections and is better suited to case **2)** attribution of the causes of past warming. The agreement between models and observations shown in IPCC Figure 8.1 has been described as "remarkable" Knutti (2008b); it is shown here that the agreement is not as good as might appear at a first glance.
The ability of models to simulate GMST in–sample is looked at more closely in this Chapter and the presentation of data is discussed. This is done by comparing observations to GCM simulations in absolute space and an analysis of the particular way that each model responds to forcings. Presenting the raw model output, without first taking *anomalies*, or the residuals of each model with respect to the observations give a very different view of model skill than the IPCC Figure 8.1.
Section 3.2 briefly introduces some fundamental ideas underlying GCMs. Section 3.3 gives details of the data used in IPCC Figure 8.1 Solomon *et al.* (2007a). The IPCC Figure is re-produced in Section 3.4 and is looked at in detail. In particular, the use of anomalies is looked at in Section 3.4 and are shown to obfuscate important information regarding the reliability of model projections. The skill of individual models is examined in terms of their *residuals* in Section 3.5 and the lack of exchangeability between model simulations analysed in Section 3.6. Results and suggestions for improvements in the presentation of model output are discussed in Section 3.7.

3.2 General Circulation Models

The IPCC AR4 report uses data from a number of GCMs. GCMs are complex computational models based on our scientific understanding of the climate system. All the GCMs looked at in this Thesis are deterministic and require a large amount of computational power to run e.g. using a distributed computing set–up it takes $\sim$ 2–3 weeks to run a 45 year simulation of the Hadley Centres' HadSM3 model on a typical Pentium4 3.2 GHz PC. Due to the large investment of time taken to develop GCMs and the computational resources required to run climate modelling experiments, studies using GCMs are often restricted to large modelling centres around the world such as the UK's Hadley Centre or the National Centre for Atmospheric Research in the US.
In this Chapter the output of each of 11 different GCMs is condensed into a single dimensional time series of GMST. GMST is calculated by an area weighted average

of surface temperature over the globe and annual averaging. More details of this calculation and specific features of GCMs are presented in Chapter 4.

3.3 The IPCC Figures

The IPCC Figures 8.1 and 9.5 present simulations of 20th Century GMST from 14 structurally distinct GCMs. For some GCMs, multiple simulations were available, making an ICE. In total 58 simulations were plotted, together with the multi–model mean of the anomaly–adjusted model output (the model–mean is calculated by the arithmetic mean over all simulations). The observations, as well as each model simulation are plotted as anomalies with respect to the period 1901–1950. Each of these GCM simulations uses a set of anthropogenic and natural forcings (including, but not exclusively, Greenhouse Gases (GHGs), solar forcings and volcanoes) determined by each modelling centre. It should be noted that since the anthropogenic and natural forcings used in 20th Century simulations are determined by each modelling centre, different forcings are used in different GCMs Covey *et al.* (2003); Meehl *et al.* (2005). Further details of how these Figures were made in the IPCC AR4 can be found in the Supplementary material to Chapter 9 of Working Group 1 of the IPCC AR4, Appendix C. These methods are described briefly in Section 3.4.
This Chapter uses 47 model simulations, available from the Coupled Model Intercomparison Project[1], of the 20th Century are analysed from 11 structurally distinct GCMs, developed at 9 different modelling centres across the world. These 47 simulations correspond to simulations used in IPCC Figure 8.1; the remaining 9 simulations were not available at the time of analysis. The models used, the number of simulations and their respective modelling centres are shown in Table 3.1.

There are a number of ways to present the same model output that can give different indications of the model's reliability. Three methods are presented in Section 3.4, beginning with the raw model output. In all three plots, observations are shown in black, model simulations in yellow and the multi–model mean in red (averaged over all simulations). Vertical blue lines indicate the timings of four

[1]I 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.

Model I.D.	n	Modelling Centre	Country
ncar-ccsm3	8	National Centre for Atmospheric Research	USA
miub-echo-g	5	Max Planck Institute for Meteorology	Germany
gfdl-cm2-0	3	US Department of Commerce, NOAA, Geophysical Fluid Dynamics Laboratory	USA
gfdl-cm2-1	3	US Department of Commerce, NOAA Geophysical Fluid Dynamics Laboratory	USA
giss-model-e-h	5	NASA Goddard Institute for Space Studies	USA
giss-model-e-r	9	NASA Goddard Institute for Space Studies	USA
inmcm3-0	1	Institute for Numerical Mathematics	Russia
miroc3-2-medres	3	Center for Climate System Research, National Institute for Environmental Studies Frontier Research Center for Global Change	Japan
mri-cgcm2-3-2a	5	Meteorological Research Institute	Japan
ncar-pcm1	4	National Centre for Atmospheric Research	USA
ukmo-hadgem1	1	Hadley Centre	UK

Table 3.1: The models used in the CMIP 3 project used in this Chapter. The model I.D., number of simulations available (n), and the modelling centre that ran the experiment is shown.

major volcanic eruptions; Santa Maria (1902), Agung (1963), El Chichon (1982) and Pinatubo (1991)

3.4 Presentation of Model Output

The raw data from the 47 simulations analysed in this Chapter are presented in Figure 3.1. Figure 3.1 shows that the models range in their base GMST by about 3 degrees Celsius. If uniform, such a difference in temperature equates roughly to a difference in radiation emitted at the surface of a blackbody of $16W/m^2$ (this number is based on the Stephan-Boltzmann energy balance equation Boltzmann (1884); Stefan (1879), stating that the total energy radiated from a black body is directly proportional to the fourth power of the black body's thermodynamic temperature)[1], a factor of 10 larger than the estimated anthropogenic forcing over the 20th Century, estimated at $1.6W/m^2$ (with a 90% confidence interval of 0.6–$2.8W/m^2$) in

[1]Of course, GCMs do not model the Earth as a black body, and so the actual difference in surface radiation between GCMs will differ from $16W/m^2$. The point here is simply that 3 degrees is a large difference for models that rely on their physical coherence for their predictive skill.

Solomon *et al.* (2007a). Differences of this magnitude in the baseline GMST could affect temperature–dependent feedback processes e.g.the 0 degree Celsius ice–line could be significantly different in models differing by 3 degrees Celsius in GMST, thus affecting albedo feedback processes. It is not argued here whether such large differences can be robustly subtracted when comparing simulations to observations (or when comparing GCMs), nor that this fact invalidates the relevance of GCMs for simulating important aspects of climate change, rather that such differences are better acknowledged. In IPCC Figure 8.1 a linear offset is applied to each non–linear simulation, eliminating the differences in absolute GMST between GCMs. The assumption that a linear offset can be robustly applied out–of–sample to a non–linear model requires justification. Without any such justification, one might interpret the 3 degree difference in GMST as a violation of the physical basis of GCMs, stated as a reason to trust climate models in Chapter 8 of the AR4, as given in Section 3.1. It is not obvious that a 3 degree difference in baseline GMST does not have a significant impact on the physical properties of these GCMs. The same data used to make Figure 3.1 is now used to make a re–production of IPCC Figure 8.1. This re–production is shown in Figure 3.2.

In Figure 3.2, each individual model simulation (yellow) is plotted as an anomaly relative to its 1901–1950 average, as are the HadCRUT3 observations (black). These anomalies are taken for each simulation by subtracting the 1901–1950 simulation mean from the 20th Century time series. The multi–model mean (red) is calculated by the arithmetic mean over all simulations from the anomaly time series, and as such is expected to lie closer to the line $y = 0$ due to a reduction in variance. The multi–model mean shows less variability that individual model simulations (furthermore, the multi–model mean will have an improved RMSE over individual simulations through its lower variability independently of any improvement in its representation of observed dynamics; a point often not considered in studies comparing the value of a multi–model mean to constituent simulations e.g. Reichler &

Kim (2008)). Plotting the data as anomalies produces an immediate overlap between simulations and observations over the period 1901–1950, regardless of whether simulations and observations sharing any common dynamics or baseline global mean temperature level. In anomaly space, the model time series appear close to observations, showing a similar magnitude of change in GMST over the 20th century. This graph is presented as evidence that "the global mean temperature trend over the past century can be modelled with high skill" in Chapter 8 of the AR4 Solomon *et al.* (2007a). Certain shortcomings of GCMs are noted in Chapter 8 of the AR4 with respect to their ability to forecast future climate change, such as "deficiencies remain in the simulation of tropical precipitation" and "The ultimate source of most such errors is that many important small-scale processes cannot be represented explicitly in models, and so must be include in approximate form as they interact with larger-scale features." (p.601, Solomon *et al.* (2007a)), but the ability of GCMs to re–produce GMST changes is not included. Indeed, the ability of climate models to simulate 20th Century GMST is presented as a source of confidence in models. Figure 3.3 shows the same plot, but with anomalies calculated with respect to each model rather than each simulation (the same period, 1901–1950 is used). Using model–means as a basis for taking anomalies assumes there is a common bias within each model that should be subtracted before comparing simulations. In contrast, taking anomalies for each simulation, as in Figure 3.2, assumes there is a global bias to be subtracted from each simulation individually and that this bias differs between simulations produced by the same GCM. Taking anomalies for each simulation means that difference bias corrections are applied for simulations produced by the same GCM. Since simulations produced by the same model differ only by IC, and not in their dynamics (the model structure and parameters are identical across all members of each model's ICE), taking anomalies for each simulation reduces intra–model variability and leads to an artificially tight agreement between simulations and observations over the period during which anomalies are taken. The

smallest difference in offset used for the same model is 0.045 degrees Celsius for the **ncar-pcm1** model and the largest is 0.294 degrees Celsius for the **gfdl-cm2-0** model. It should be noted that the 0.294 degree difference in offset is of the same order of magnitude as the observed warming over the 20th Century (~0.74 degrees Celsius). Taking a single offset for each GCM, Figure 3.3 distorts the variability between the simulations, when compared to Figure 3.2. Over the period 1901–1950, the variance across simulations is, on average, 20% higher for simulations with model–means subtracted compared to simulations with means subtracted for every simulation. Conversely, the variance over the second half of the 20th Century is about 10% lower when taking a single offset for each model, rather than each simulation. These differences in variance might partly be due to any difference in offset at the beginning of the 20th Century disappearing over time in the model–mean anomaly case; there will be no difference, on average, between members of an ICE. Thus, taking anomalies for each simulation rather than for each model has the effect of distorting the variability in the ensemble since systematic differences in GMST are introduced between ICE members where none should exist.
The variance of the ensemble increases greatly towards the end of the 20th Century; the variance across the 47 simulations is approximately 3 times greater in the last 10 years of the 20th Century, taking anomalies either as simulation or model means, than for the first 10 years. This could be an indication of the variable response to GHG forcings across models or the most recent observational data not being available for "tuning" the model (pp.596, Working Group 1, AR4, also Bender (2008); Murphy *et al.* (2004); Stocker (2004)) during development.

It is not clear that temperature anomalies are sufficient for decision makers than absolute temperatures. Absolute temperatures are also very relevant for the purposes of many decisions. In the case of impact studies, it can be very important to consider the absolute temperature since various events of interest are linked to a specific temperature e.g. water freezing at 0 degrees has an impact on agriculture,

sea–ice extent, and even planetary *albedo*. Other examples of impacts sensitive to absolute temperature are crop failure, heat mortality and water vapour feedbacks. Note that these impacts are not global; knowing GMST is insufficient to uniquely deduce regional impacts Smith *et al.* (2008). The relationship between GMST and regional climate response is looked at in detail in Chapter 8.

Presenting the model output in absolute temperature space gives a very different picture of the model's ability to re–produce 20th Century observations. In Figure 3.2 the models appear to re–produce the dynamic changes seen in observations with significantly different base GMSTs. Section 3.5 looks at the residuals of each GCM individually in order to understand how well each GCM captures the dynamics of the observed GMST time series.

3.5 Residual Analysis of Model Output

Section 3.4 illustrated significant and systematic differences in GMST between the GCMs that make up IPCC Figure 8.1. Even if presenting GCM output as anomalies is justified, understanding how well the individual GCMs capture the dynamics of the climate system being modelled is useful for assessing their likely out–of–sample skill. The extent to which GCMs can simulate the dynamics of observed GMST change provides an estimate of the limit of their predictive skill out–of–sample. This model–observation comparison can be done by looking at the residuals for each GCM. Residuals are defined by subtracting the time–series of observations from each model simulation. This differs from the model anomalies; anomalies are calculated by subtracting a single number from the entire time series, in this case the 1901–1950 mean.

Figures 3.4, 3.5, 3.6 and 3.7 show the model output for the 11 GCMs used in residual space (the time series of observations are subtracted from each simulation). The simulations are presented by GCM, with the line $y = 0$ representing the observations. There are three notable points about these residual plots:

1. Residuals are often as large as 0.3 degrees and in some cases are greater than 0.5 degrees. This is large in comparison to the magnitude of warming seen over the 20th century (~0.74 degrees) and provides a lower limit on these GCMs' likely accuracy out–of–sample.

2. There is clear structure in the residuals, indicating that model errors can not be assumed to be independent over time and identically distributed. For the **giss-e-r**, **miub-echo-g**, **miroc3-2-medres** and **mri-cgcm2-3-2a** models there is a pattern in the residuals inverse to that of observations – observations warm from 1901–1945, then cool or level out to 1960 following by sustained warming to 2007. When these four GCMs' residuals are linearly regressed against observations, each simulation has a significantly negative slope coefficient, as shown in Table 3.2. That these patterns are seen inversely in some models suggest these GCMs under–react in response to rising levels of GHGs i.e. these GCMs warm, but by systematically less than observations. These results indicate that the model error in these GCMs is systematic and relevant for their ability to simulate 21st Century climate. If these GCMs are not responding to forcings in the same way as the observations, their projections will be increasingly unreliable for extrapolations further into the future.

3. The GCM ICEs do not always "capture" truth; often IC members are too hot or too cold and react in a common way; this effect can be most clearly seen in Figure 3.5, top graph (**giss-e-r model**), where the 9 member ICE captures the observations in only 43 years of the 20th Century. If the model were accountable Smith (2001), we would expect to see some IC members above the observations and some below (allowing for the ensemble size Judd *et al.* (2007)). For those GCMs with more than 1 ICE member (a 1 member ICE never captures), between 39 and 65 years of observations are captured. Testing the ability of GCMs to capture observations in-sample allows systematic model

Model I.D.	n	Min/Max Slope Coefficient	Min/Max St. Dev.
miub-echo-g	5	(-0.77,-0.43)	(-2.92, -6.23)
giss-model-e-r	9	(-1.28,-0.97)	(-6.12, -12.6)
miroc3-2-medres	3	(-1.07,-0.86)	(-8.8, -9.2)
mri-cgcm2-3-2a	5	(-0.86, -0.77)	(-11.2, -12.9)

Table 3.2: The range of slope coefficients and their significance using a simple linear regression of GCM residuals against observations for four GCMs. The number of simulations available for each GCM is denoted by n.

errors to be diagnosed.

When multi-GCM output is shown, without distinction for the constituent GCMs, it is no longer possible to readily understand the details of individual simulations. The presentation of model output in IPCC Figure 8.1 obscures important information regarding uncertainties and makes the simulations appear in better agreement with observations than suggested by the two alternative methods presented here, or by analysis of the residuals of individual GCMs.

3.6 Exchangeability

This Section answers the question of whether GCM output can be considered exchangeable. Exchangeability is defined here as the case where two models produce output that are identically distributed, a critical assumption underlying many statistical methods. This definition is adapted from Galambos (1995) – "exchangeable random variables are identically distributed.". The exchangeability of models is analysed using two statistical methods; **1)** Using order statistics to evaluate where one model is significantly hotter than the other and **2)** The non–parametric Kruskal-Wallis test Kruskal & Wallis (1952) for equality of medians is applied across 5 models with at least 5 available simulations. A new method, based on order statistics, is used to estimate the temporal correlation of model output in order to test the assumptions that the analysis exchangeability relies on.

The residual plots shown in Figures 3.4, 3.5 and 3.6 in Section 3.5 suggest a lack of exchangeability between GCMs (different GCMs appear to have different dynamics) - it can not be assumed that models are drawing from a common distribution. This lack of exchangeability can be shown by plotting all GCM residuals simultaneously, as in Figure 3.8, and highlighting two models for comparison. The same model output as shown in Figure 3.2 is shown in two different presentations - the top graph shows models and observations as anomalies from their respective 1901–1950 means and the bottom graph shows the residual time series of the model output. The time series from two GCMs are coloured in red (8 simulations from the **ncar-ccsm3** GCM) and in blue (5 simulations from the **giss-model-e-h** GCM). Colouring the simulations of these two GCMs highlights the fact that these models appear not to be sampling from a common distribution. During the last ~30 years of the 20th Century the NCAR model is frequently hotter, for every IC member, than all simulations from the GISS model. The two GCMs match closely over the first half of the 20th Century then differ in the latter part of the Century, suggesting that these models are responding to rising GHGs in a different way. It might be thought that the different responses can be explained by the different *climate sensitivities* [1] of the models, with varying model response to other factors such as aerosols and volcanoes also accounting for some of these differences. In this case CS refers to the amount of GMST rise that eventually results from doubling CO_2 concentrations in a climate model. The equilibrium CS of the **ncar-ccsm3** and **giss-h** models both have an equilibrium CS of ~2.7 degrees Celsius Kiehl *et al.* (2006); Schmidt *et al.* (2006). Despite having the same equilibrium CS, it has been noted that there is not always a direct relationship between equilibrium CS and transient response Raper *et al.* (2001). The Transient Climate Response (the temperature increase at the point of CO_2 doubling, where CO_2 concentrations are increased steadily at 1% per year, i.e. after year 70 of the simulation) of the **ncar-ccsm3**

[1] Climate sensitivity will be explained in more detail in Section 4.4.1

and **giss-h** are 1.5 and 1.6 degrees Celsius respectively Solomon *et al.* (2007a). It seems that it is not possible to explain the difference in late 20th Century GMST response by differences in the NCAR and GISS models' CS.

The lack of exchangeability of these NCAR and GISS GCMs can be seen quantitatively by counting the number of **ncar-ccsm3** simulations that are hotter than all **giss-e-h** simulations at each time point. It is expected that, because there are 8 NCAR simulations, and 5 GISS simulations that the NCAR model will, on average, contain at least one warmer simulation if the two models are drawing from the same distribution. The theoretical number of simulations from an ensemble exceeding the hottest simulation from another ensemble can be calculated under the assumptions of exchangeability (simulations from the two GCMs draw from the same distribution) and a lack of temporal correlation.

The hottest simulations from the 5 member ensemble is expected to be exceeded $\frac{1}{6}$th of the time by an ensemble drawing from the same distribution i.e. for an ensemble of size 8, an average of $\frac{4}{3}$ simulations each year. Thus, the probability of all 8 NCAR simulations being hotter than all 5 GISS simulations is $\sim 5.9 \times 10^{-7}$ if both models are drawing from the same distribution (this assumes a Binomial distribution of NCAR simulations hotter than GISS simulations with a probability of $\frac{1}{6}$ over 8 trials). There is less than a 1% chance for at least 5 of the 8 NCAR simulations being hotter than all 5 GISS simulations under the assumption of exchangeability. Figure 3.9 shows the number of NCAR CCSM simulations that are hotter than the hottest GISS simulation for each year over the 20th Century. The horizontal line shows the theoretical value of $\frac{4}{3}$ for the expected number of NCAR simulations warmer than the warmest GISS simulations for each year. From 1900 to 1940 there are typically 1 or 2 NCAR simulations warmer than the warmest GISS model simulations. Figure 3.9 then shows a significant trend from 1940 onwards, with 6 or more NCAR simulations being hotter than all GISS simulations by about 1970. This suggests that these models are not exchangeable and are responding in a different way to

the rising GHGs that drive mid-to-late 20th Century warming. Given the different dynamics of these models, it seems unreasonable to treat these GCMs' ensemble members as exchangeable. The theoretical statistics calculated here depend on a lack of temporal structure if the available data set is small, although the results are so significant that temporal correlations are unlikely to account for the differences seen. The temporal correlation of these GCMs is estimated in Section 3.6.1.

3.6.1 Estimating temporal correlation using order statistics

Many statistical method, including the tests for exchangeability presented in this Section, rely on the assumption that model output is temporally uncorrelated. A method is presented here that estimates temporal correlation within an ICE using order statistics. This method is applied as follows:

1. At time t, define the extremal simulations of an ICE of size n as $x_{min,t}$ and $x_{max,t}$. The extremal simulations are defined whenever $x_{min,t} \neq x_{min,t-1}$ or $x_{max,t} \neq x_{max,t-1}$ (this is done to avoid over–counting simulations which define the minimum or maximum in consecutive years). Henceforth the calculation is notated for the case of $x_{m,t}$, where m represents either the minimum or maximum.

2. Define the mixing time, τ, as the time taken for the time series of the simulation defined by $x_{m,t}$ to cross the median.

3. A simulation $x_{m,t}$ at time t is defined to cross the median where its rank is less (greater), if the calculation uses $x_{max,t}$ ($x_{min,t}$), than $\frac{n}{2}$.

4. $x_{m,t}$ crosses the median at time $t + \tau$, where τ denotes the mixing time.

The distribution of τ provides an estimate of the temporal correlation within a model. The above calculation was carried out for the 5 GCMs with 5 or more available simulations. These distributions are shown in Figure 3.10 in red. The

median mixing times for these distributions are 2, 2, 3, 3 and 4 years respectively. The maximum mixing time for these distributions are 10, 16, 17, 21 and 23 years respectively. The theoretical distributions are plotted in black in Figure 3.10. Based on the assumption that the mixing time of each model is 1 year, the theoretical distribution of estimated mixing times was calculated using a Geometric distribution. If the theoretical distribution is a good fit (a Chi–squared test could be used to test this formally) to the empirical distribution, it might be concluded that the model does not typically have significant temporal correlation. If the theoretical distribution is a poor fit, then the mixing times could be repeated using multi–annual means, increasing the temporal averaging until the theoretical distribution well approximates the empirical distribution. The theoretical distribution is a good approximation for the **mri-cgcm2-3-2a** and **miub-echo-g** models and a poorer fit for the other three GCMs, especially the **ncar-ccsm3** GCM. This suggests that the mixing times for some GCMs is longer than 1 year, although this is not always the case. It would be possible to estimate the mixing time of a GCM by repeating the above analysis based on a n year temporal mean, and increasing n until the theoretical distribution well approximates the empirical distributions.

3.6.2 Testing the exchangeability of GCM output

In order to show the results indicating a lack of exchangeability between the **ncar-ccsm3** and **giss-model-e-r** GCMs are robust when expanded to include all GCMs with at least 5 simulations a Kruskal-Wallis test is carried out here. The Kruskal-Wallis test is an extension of the Mann-Whitney rank sum test Mann & Whitney (1947) that tests the equality of medians across 3 or more distributions. Here, the 5 GCMs with at least 5 simulations are selected for analysis (5 members are generally required for the Kruskal-Wallis test statistics to be well-approximated by a Chi-squared distribution Kruskal & Wallis (1952)). Output from the **mri-cgcm-3-2a**, **miub-echo-g**, **giss-model-e-h**, **giss-model-e-r** and **ncar-ccsm3** GCMs is used

here, which have 5, 5, 5, 8 and 9 simulations respectively. The test was then carried out for each year of the 20th Century on the 32 simulations from these 5 GCMs, giving a time series of p-values. The model–mean adjusted anomaly time series are used – the significant differences in baseline GMST is not included in this test. This time series is shown is Figure 3.11. From 1901–1960 models there is little evidence that the models have different annual medians, although it should be noted that all models are forced to have zero mean over the 1901–1950 period by the use of anomalies. From 1960–2000 there is strong evidence that the models have different medians for every year. This test provides evidence that it can not be assumed, over the second part of the 20th Century that these 5 GCMs are sampling from the same distribution. The test carried out here assumes that there is no temporal correlation within these GCMs simulations. As shown in Section 3.6.1, the mixing time of these GCMs can be longer than 1 year. In order to show that these results are robust to temporal correlations, the Kruskal-Wallis test was repeated for multi-year mean time series. Temporal means are taken for 2, 5, 10 and 25 year means and the results are shown in Figure 3.12. In all cases, the test is generally non–significant during the first half of the 20th Century, but is always significant towards the end of the 20th Century. This indicates that these 5 GCMs are not exchangeable during the second half of the 20th Century, suggesting different dynamical responses to mid to late 20th Century forcings.

3.6.3 Discussion of Results

Since the GCMs analysed show different dynamical behaviour in response to 20th Century forcings, the relevance of certain multi–model statistics is dubious. An important difficulty follows when non–exchangeable models are used to form multi–model statistics based on an "ensemble of opportunity" (one can only analyse the simulations that have been provided). In this case, statistics will be biased depending on which models are included and, in some cases, the number of simulations that

are provided for each model. This is particularly a problem in climate model comparison where there is little co-ordination of experimental design. Similar difficulties arise when statistics are calculated across all available simulations disregarding the number of simulations produce by each model. In this case, models with more available simulations will be effectively weighted more heavily e.g. the two variants of the GISS model will receive, jointly, 14 times more weight than either the Russian **inmcm3-0** or the UK's **hadgem1** GCMs.

Further to the problems facing statistics complied from non–exchangeable ensembles of opportunity, the presentation of multi-model means from non–exchangeable models faces further difficulties. In particular, multi-model means are necessarily low in variability and can lead to a cancellation of temporal variability. This can be shown with a simple example. Suppose model A produces output with temporal variability similar to that of a *sine* function and model B similar to that of an offset *sine* function such that the sum of the two *sine* functions is 0. When taking the multi–model mean of Models A and B from two ensembles of equal size (or giving equal weight to each model in some other way), the effect will be a straight line with no temporal variability. The multi–model mean need not bear a strong resemblance in temporal variability to its constituent simulations.

It has been shown in this Section that GCM output can not be assumed to be exchangeable. This result has important consequences for the utility of multi–model mean statistics that treat models as being interchangeable.

3.7 Recommended Presentation of Model Output

An alternative way of presenting GCM output to the IPCC AR4 Figure 8.1 is suggested in this Section. Three amendments are suggested:

1. If anomalies are to be used, these anomalies should be taken with respect to

each model, not each simulation. This approach is consistent with the fact that IC members from the same GCM follow the same dynamics and have the same systematic dynamical biases.

2. The use of a linear offset to reduce systematic differences should be stated explicitly when presenting anomalies. It is important to give the magnitude of this adjustment especially when it is large. In addition to this it is important, in the accompanying text, to state explicitly the justification for using this offset and whether it is likely to hold out–of–sample.

3. The performance of individual models in capturing the observations should be central to the evaluation of multi–model ensembles, especially where models can not be assumed to be exchangeable. This could be achieved by running an ICE (perhaps 10+ members) and looking at how often the model captures the observations, such as in Weisheimer *et al.* (2004). When using GCMs to extrapolate, it is important that individual models capture the dynamics of change in the observations.

The first two suggestions above are important whether the intention is to show the fact that all models warm under anthropogenic forcings, as in Chapter 9, Figure 9.5 of the AR4, or to give evidence to trust models out–of–sample, as in Chapter 8, Figure FAQ 8.1 of the AR4. The third condition is aimed more at providing a clearer view of a model's likely skill out–of–sample.

3.8 Conclusion

The presentation of model output in Figure 3.2, as shown in IPCC AR4 does not acknowledge significant systematic biases in the model's base GMST and large discrepancies between individual model simulations and observations. The presentation of data in IPCC Figure 8.1 suggests that the combined effect of 20th Century

anthropogenic and natural forcings is sustained warming and that this result is robust across GCMs. The IPCC Figure is more suitable in the case of establishing the sign of the net effect of forcings on GMST, such as in Stott *et al.* (2006), although a table could equally provide this information. It does not, on the other hand, give sufficient evidence to support the statement that GCMs can accurately re–produce observed GMST in either its mean value or its dynamics. Individual GCMs must accurately capture the dynamics of observations in–sample in order to justify their extrapolations as reliable out–of–sample.

It has been shown here that the presentation of GCM output in–sample in the IPCC Figure is inappropriate and I could not find an explicit acknowledgement of the concerns raised here in the AR4. In cases of decision–support, it is critical to present uncertainties with such transparency that the users of climate predictions will not be surprised when comparing observations in 2010 or 2020 (or updated model output before that time) to current GCM output. To overstate the predictive skill of GCMs risks misleading decision makers and will likely weaken confidence in climate science. The in–sample differences shown in this Chapter can be seen as a "straw-man" test, similar to the in-sample test explained in Chapter 2. The differences between GCMs simulations and observations in-sample provides an upper bound on their likely accuracy out-of-sample. These differences can be significant e.g. residuals values of up to 0.5 degrees Celsius. The ability of GCMs to extrapolate future climate changes in this case does not seem well founded without an clear acknowledgement of **a)** the use a linear offset to produce anomalies and **b)** the structure in the residuals. Point **b)** is important since residual structure suggests that the some GCMs used in this analysis may not be capturing the dynamics of temperature change observed, an important condition for useful out–of–sample simulation.

The 47 simulations looked at in this Chapter are insufficient for a detailed sensitivity analysis of model response to GHGs. Since the range of behaviour across different structural models underestimates the full range of possible climate responses Stain-

forth *et al.* (2005) it is necessary to look at the diversity of model behaviour in a large set of simulations in order to make robust statements about future climate. The methods of analysing uncertainties used in the remainder of this Thesis were only applicable thanks to the availability of a large set of GCM simulations from the CPDN experiment. Chapter 4 introduces the details of the CPDN experiment and the data sets that will be analysed in subsequent Chapters.

New results presented in this Chapter are:

- There are significant differences between different GCMs' global mean temperature of up to 3 degrees Celsius. Such large differences call into question the physical basis of these models.

- The effect of taking different types of anomalies gives significantly different presentations of GCMs' in–sample fit. The use of anomalies in the IPCC AR4 distorts the variability of models both within Initial Condition Ensembles and across different structural models.

- Model output has been compared to observed global mean temperatures over the 20th Century. Residuals are compared on a model by model basis and it has been shown that **1)** There can be considerable structure in the residual time series and **2)** The magnitude of residuals can be large (up to 0.5 degrees Celsius) in comparison to observed 20th Century global warming ($\sim$0.74 degrees Celsius).

- CMIP3 GCMs are not exchangeable, calling into question the relevance of many methods of statistical analysis for climate model output. GCMs can not be assumed to be sampling from a common distribution.

- A new method for estimating the temporal correlation within GCM time series is proposed. The mixing time for some GCMs is not significantly different from 1 year (the higher frequency of data used here), but that it can be higher for other GCMs.

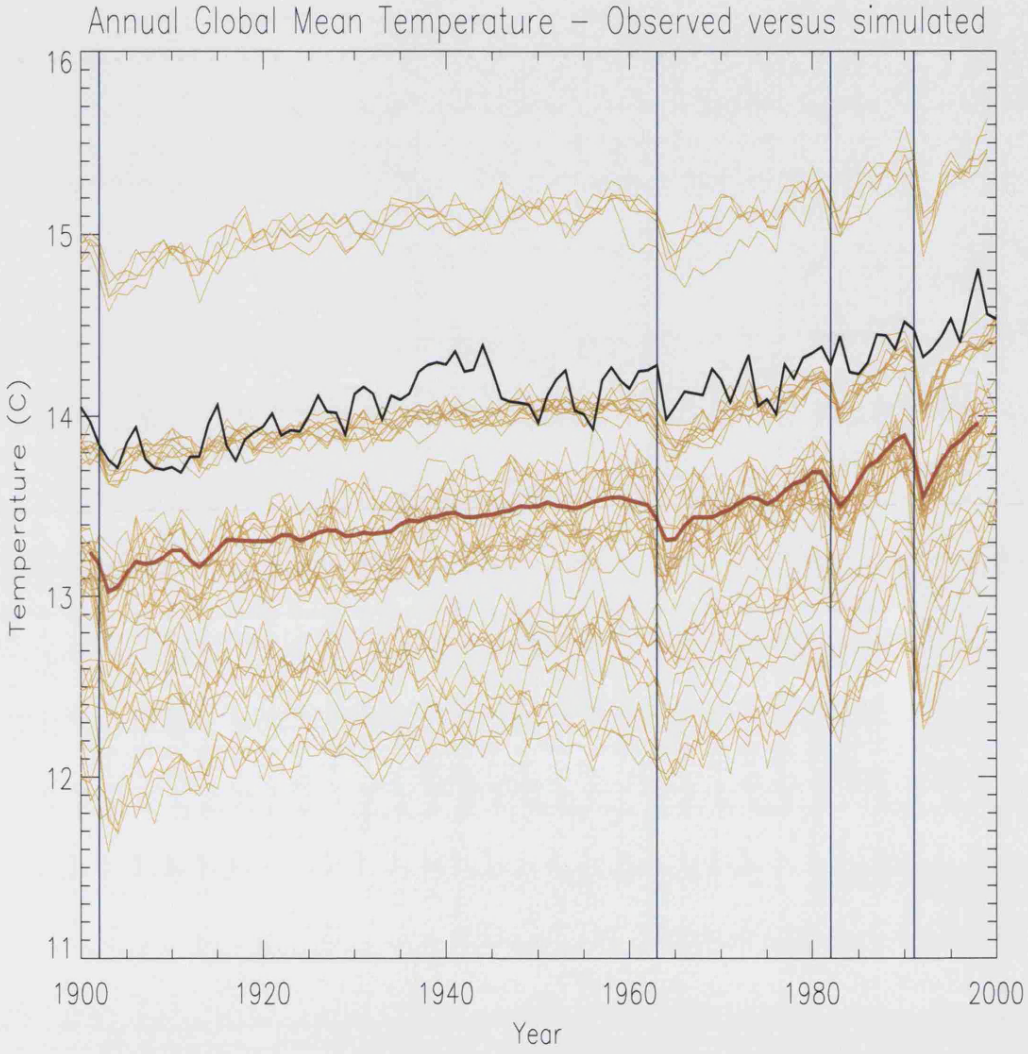


Figure 3.1: The absolute values of GMST from 47 simulations are plotted in yellow. The HadCRUT3 observations are plotted in black (the anomaly time series is offset using the 1961–1990 global mean (14.0 degrees Jones *et al.* (1999))). The multi–model mean is plotted in red. There is a difference of up to 3 degrees between simulations' GMST.

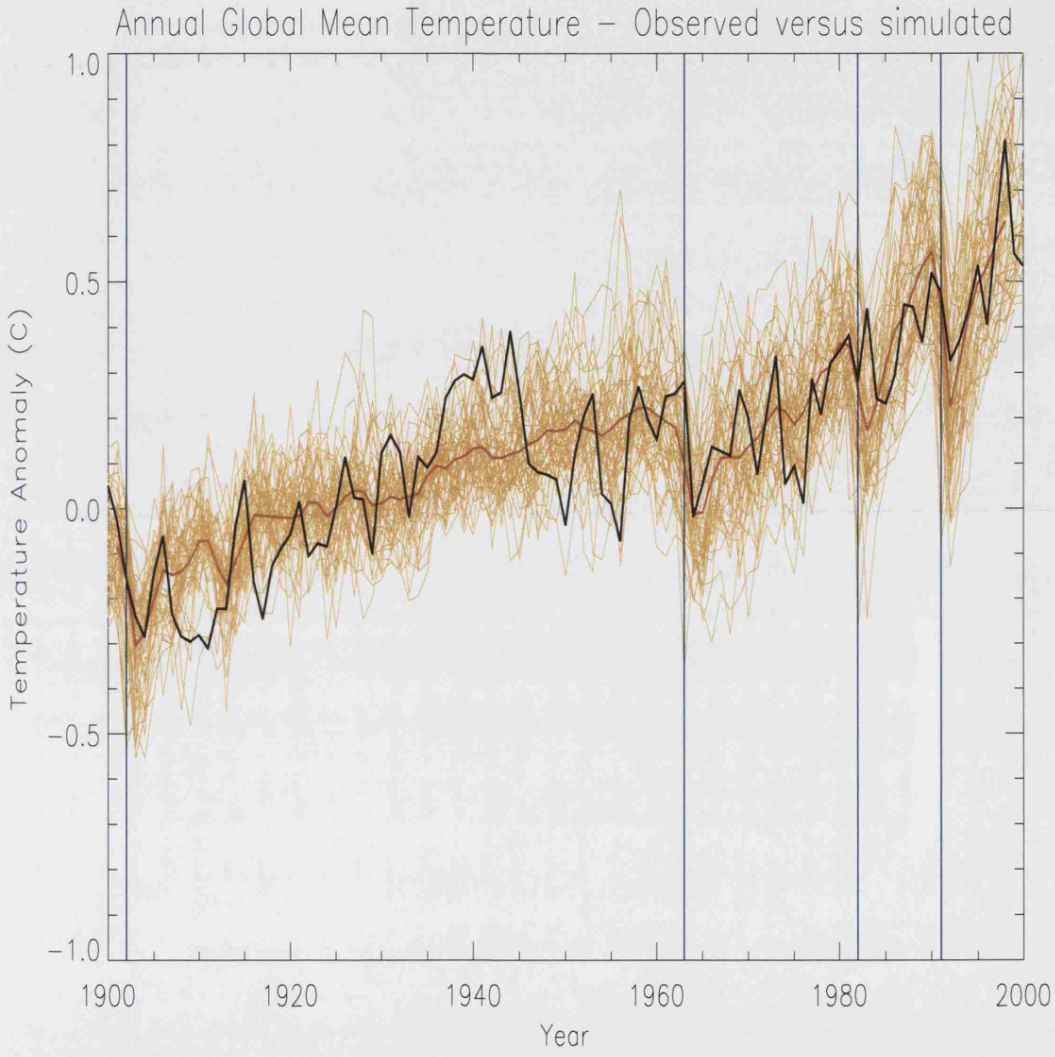


Figure 3.2: (Reproduction of IPCC Figure 8.1) Comparison of 47 simulations from 11 structurally distinct GCMs (yellow) used in the AR4 to HadCRUT3 observations (black). The multi–model mean is plotted in red. Each model simulation is "centred" by taking anomalies relative to 1901–1950. Blue lines show the timings of four major volcanic eruptions – Santa Maria, Agung, El Chichon and Pinatubo.

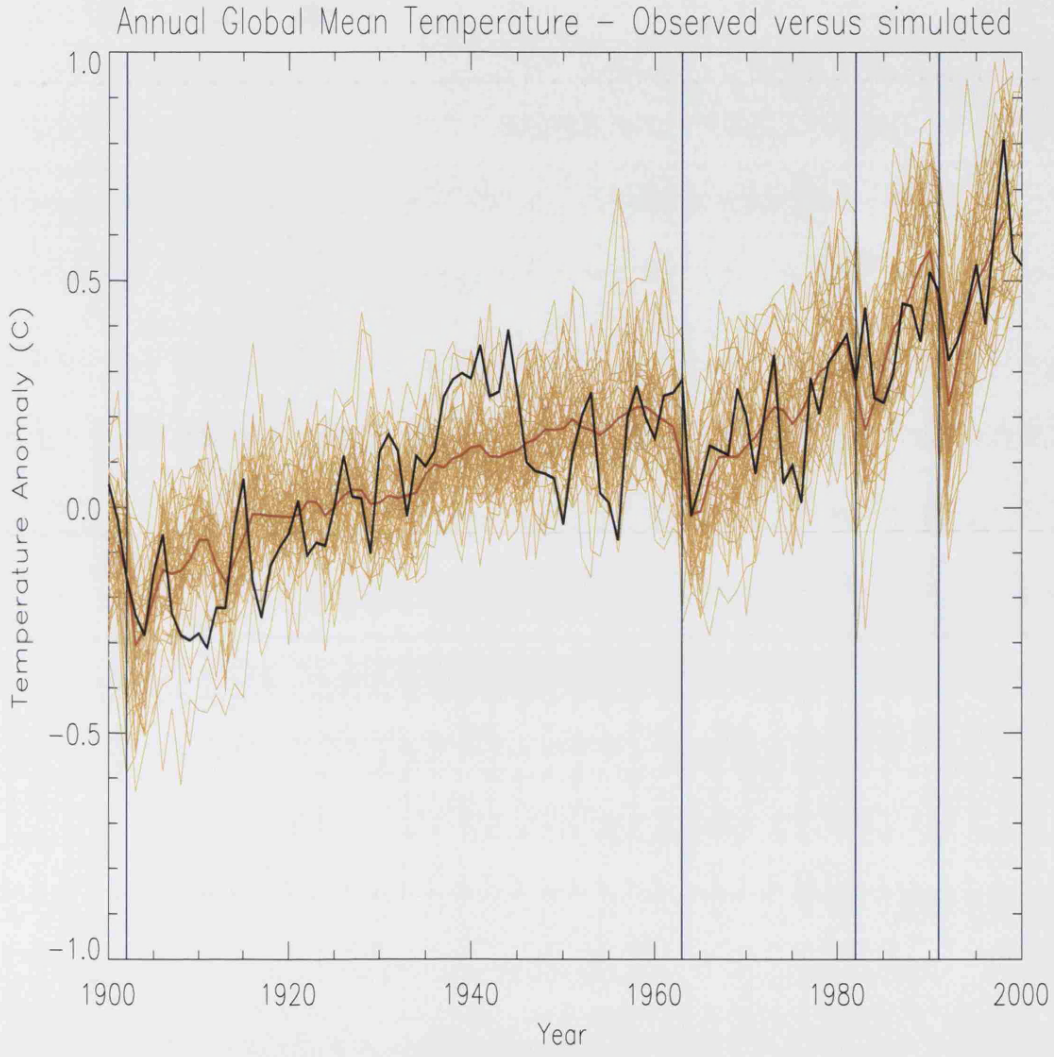


Figure 3.3: Comparison of 47 simulations from 11 structurally distinct GCMs (yellow) used in the AR4 to HadCRUT3 observations (black). The multi–model mean is shown in red. In this plot the model is centred using the mean 1901-1950 anomaly for each model (averaged over IC members). There is slightly more variance across model simulations, during the 1901–1950 period where anomalies are taken, in this plot than in Figure 3.2, as expected.

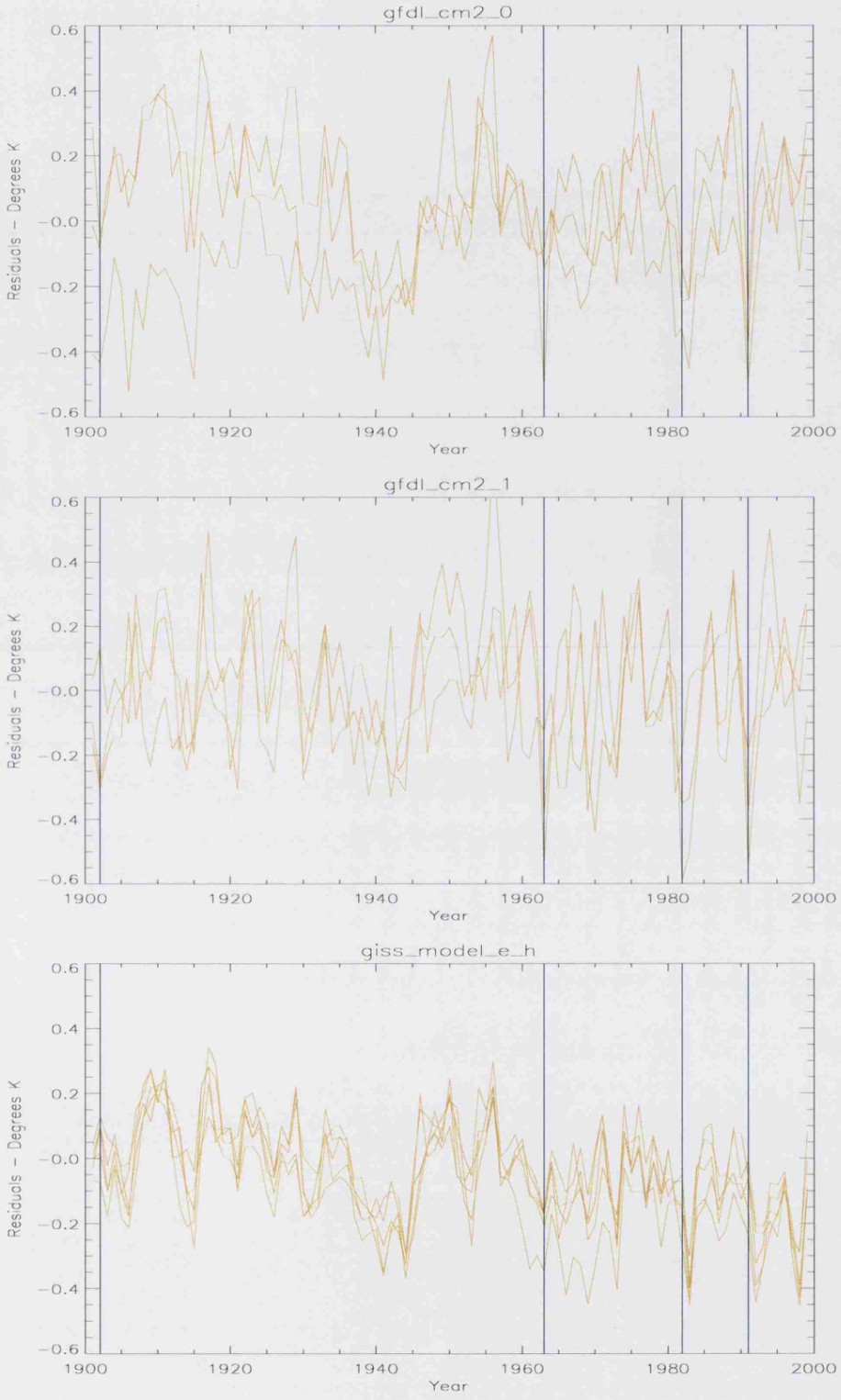


Figure 3.4: The residuals for 3 different GCMs are shown as a time series. Residuals for each simulation are found by subtracting the HadCRUT3 observations from each simulation (and adjusting for any differences in baseline 1901–1950 GMST).

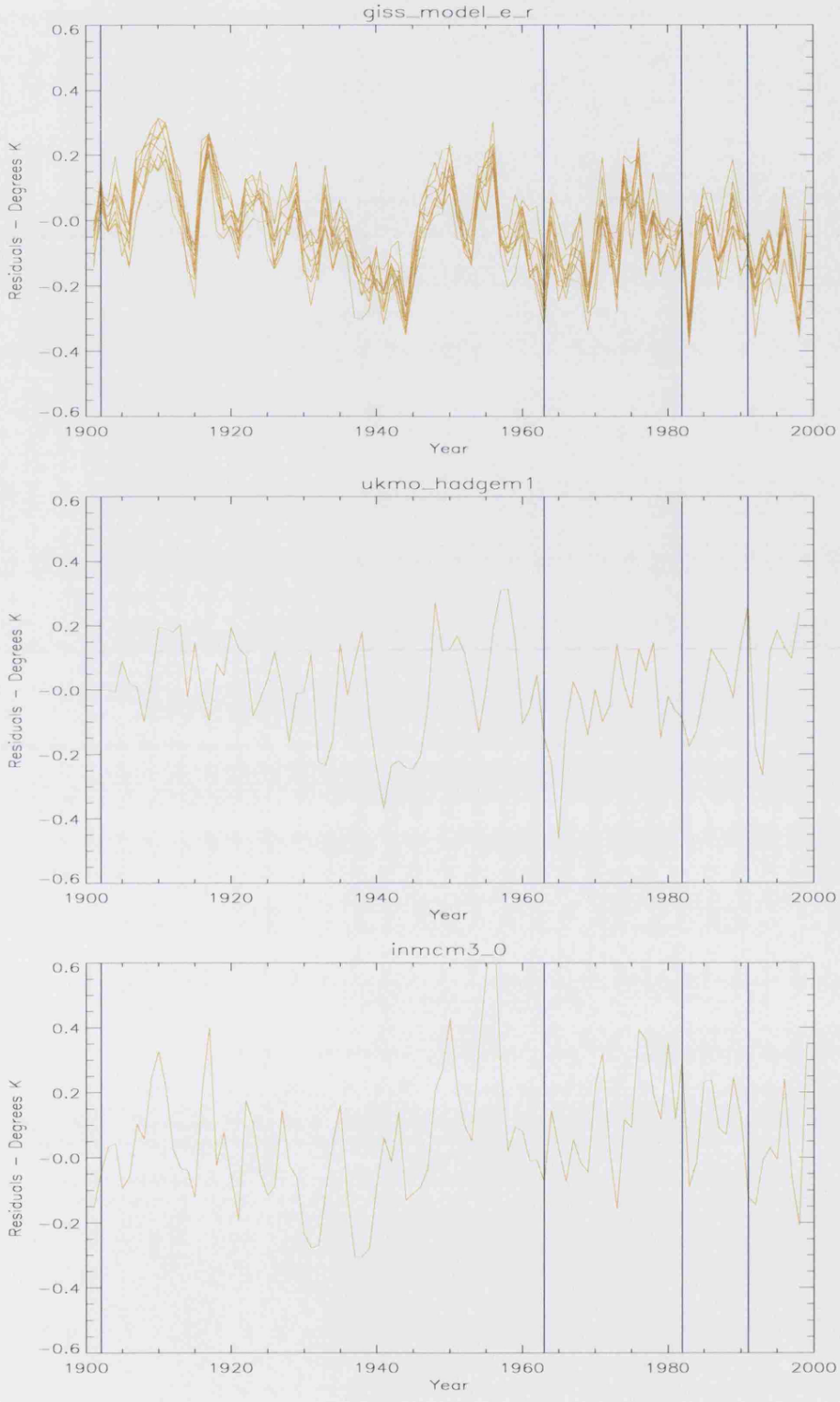


Figure 3.5: The residuals for 3 different GCMs are shown as a time series. Residuals for each simulation are found by subtracting the HadCRUT3 observations from each simulation (and adjusting for any differences in baseline 1901–1950 GMST).

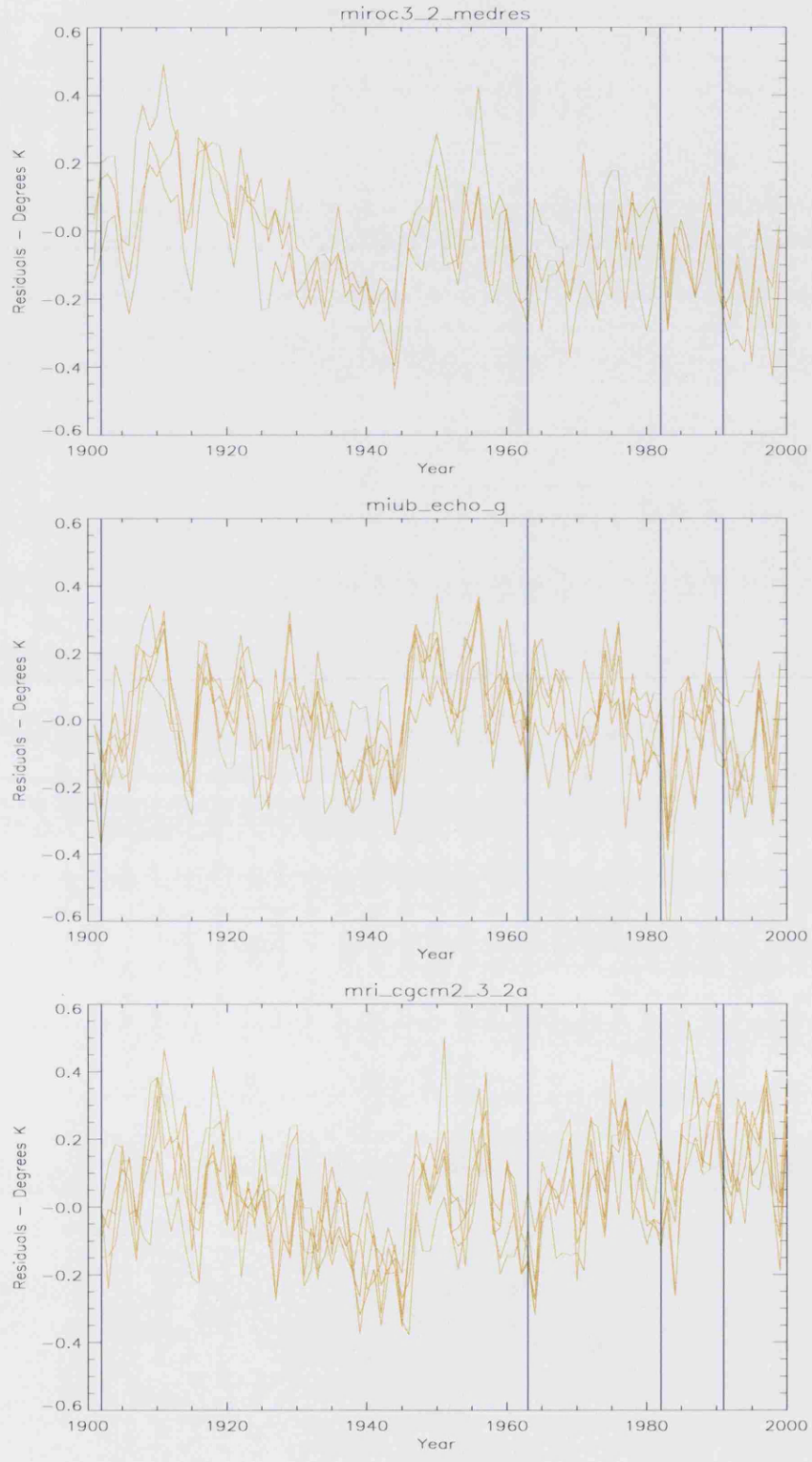


Figure 3.6: The residuals for 3 different GCMs are shown as a time series. Residuals for each simulation are found by subtracting the HadCRUT3 observations from each simulation (and adjusting for any differences in baseline 1901–1950 GMST).

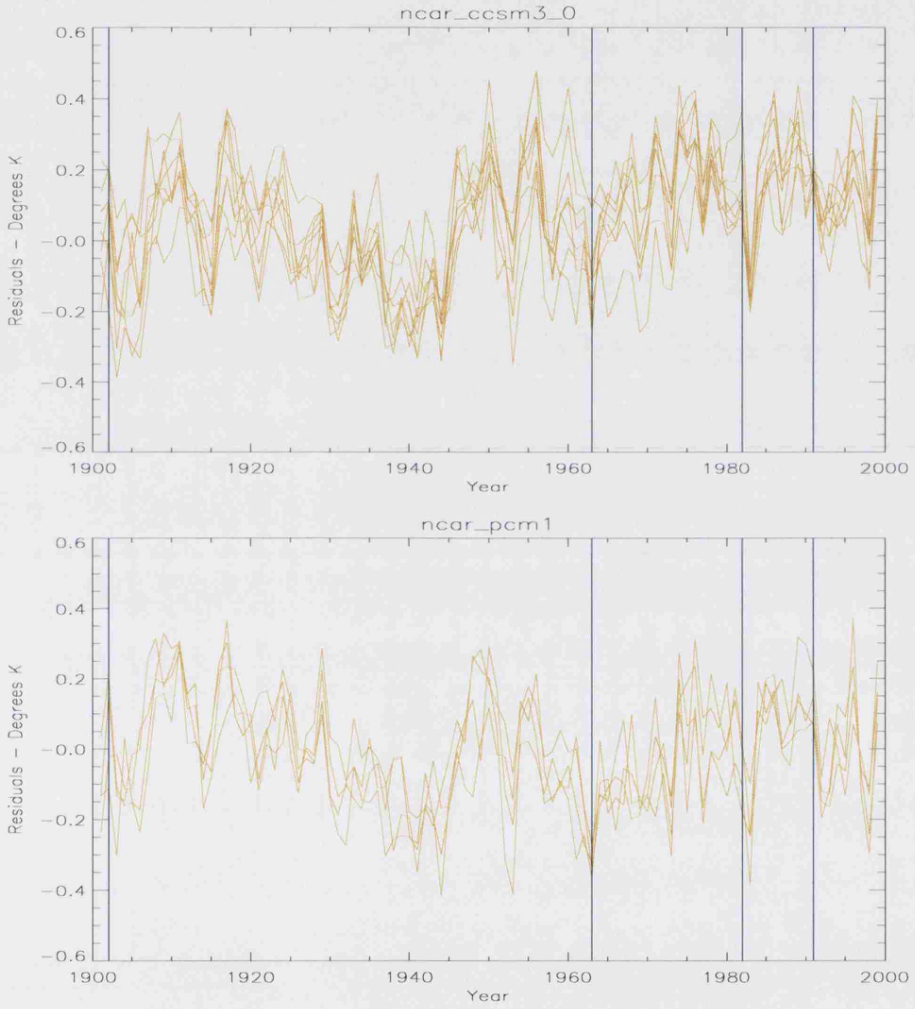


Figure 3.7: The residuals for 2 different GCMs are shown as a time series. Residuals for each simulation are found by subtracting the HadCRUT3 observations from each simulation (and adjusting for any differences in baseline 1901–1950 GMST).

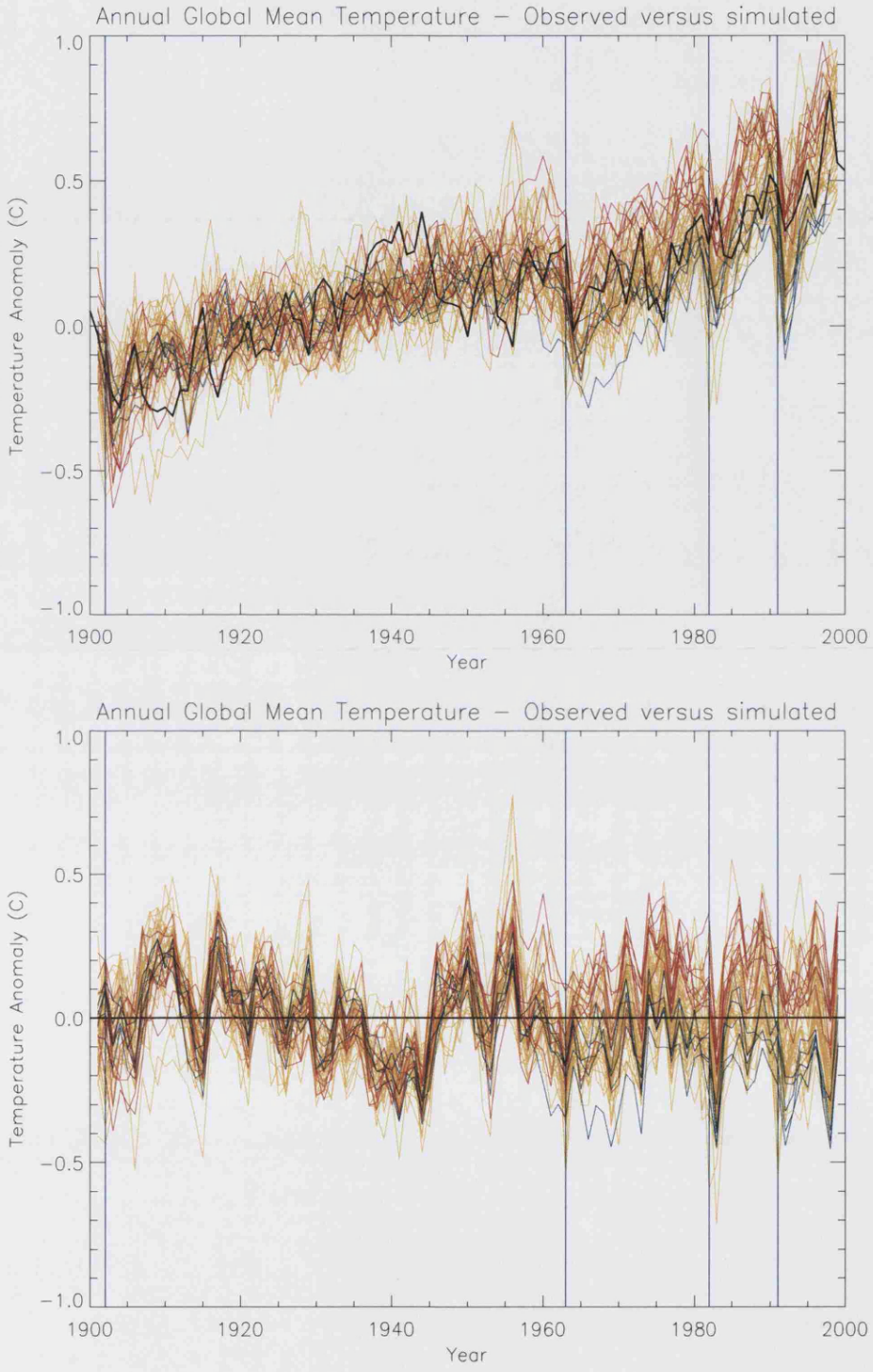


Figure 3.8: The time series of 47 GCM simulations is plotted against observations as 1901–1950 anomalies (top) and as residuals (bottom). The NCAR PCM1 GCM is highlighted in red and the GISS–h model in blue. The highlighted models overlap in the first half of the 20th Century but diverge from 1960 onwards.

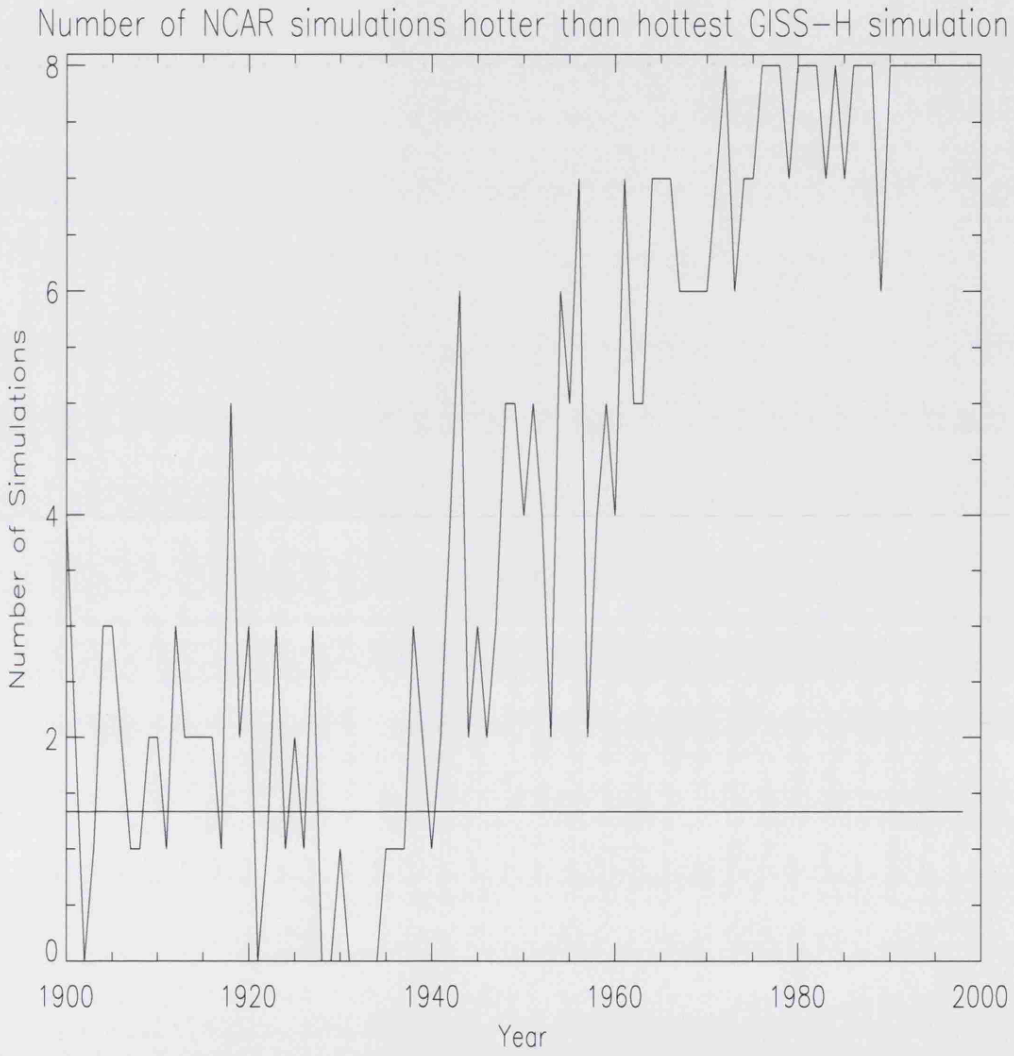


Figure 3.9: The number of NCAR PCM1 simulations that are hotter than the hottest GISS–h simulation over the 20th Century. The horizontal line shows the number of simulations we would expect to be hotter, on average, at each time point if the models were sampling from the same distribution. The horizontal line shows the 5% significance level used in this test.

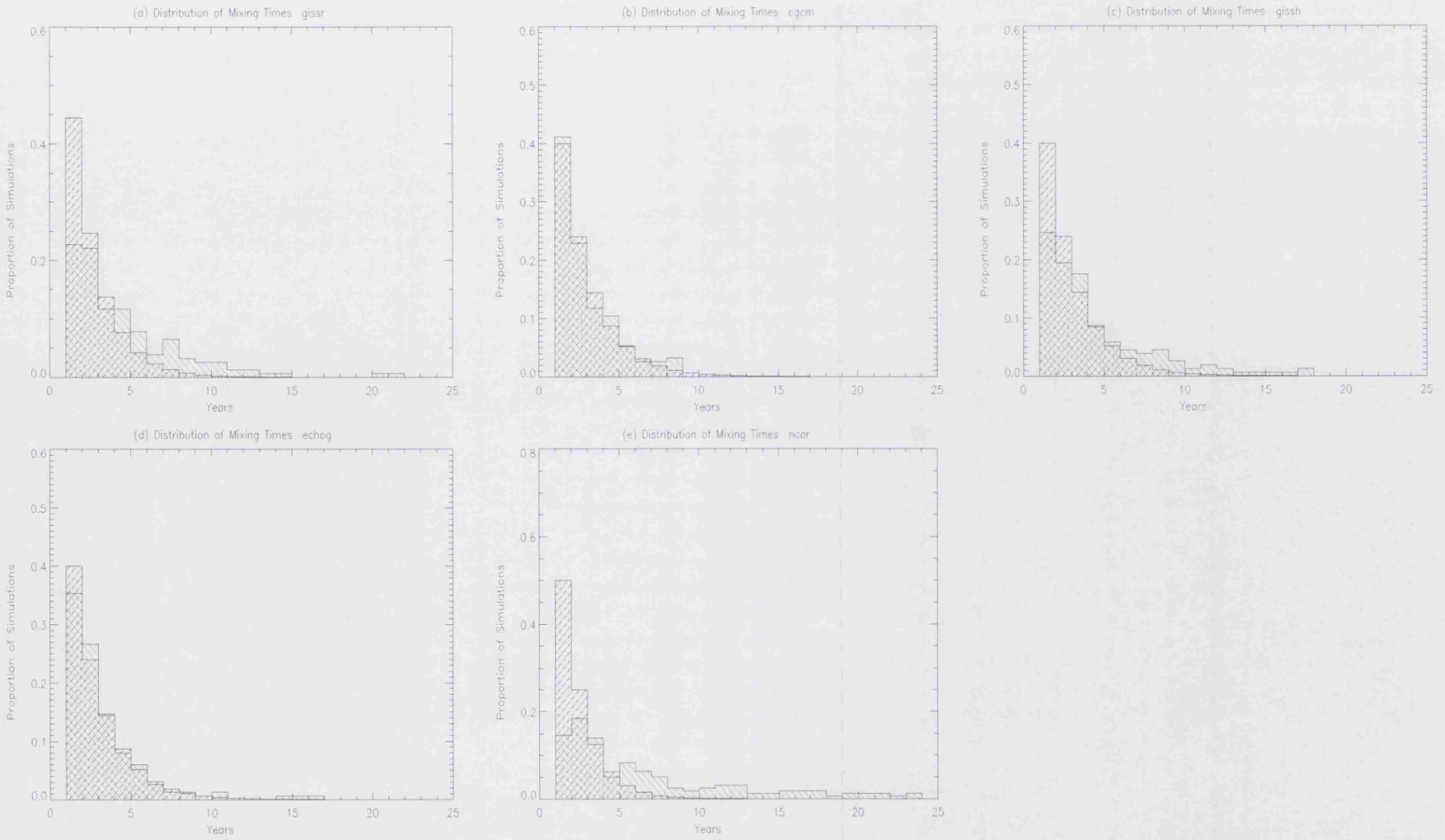


Figure 3.10: The distribution of mixing times for 5 GCMs - **mri-cgcm2-3-2a**, **miub-echo-g**, **giss-echo-e-h**, **giss-echo-e-r** and **ncar-ccsm3** with 5, 5, 5, 9 and 8 simulations respectively.

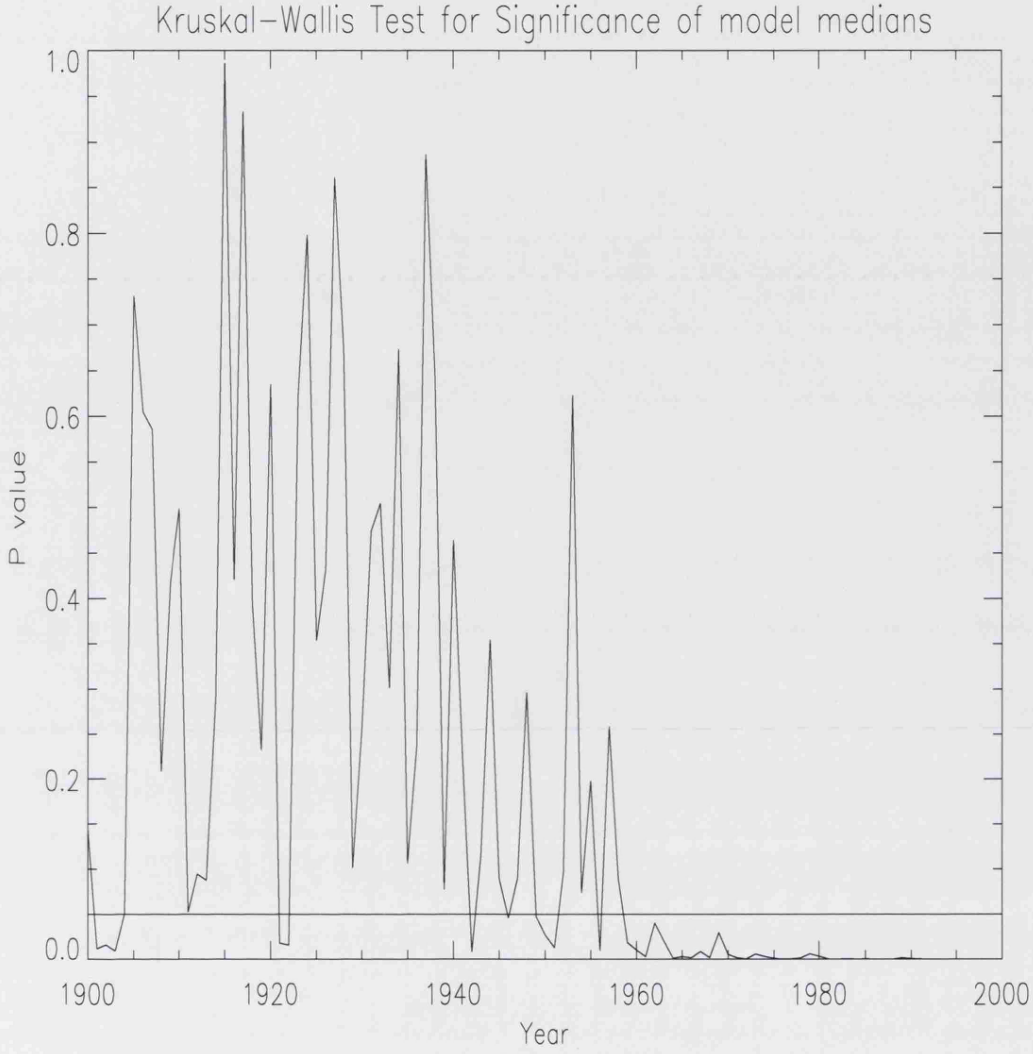


Figure 3.11: The p-values for the Kruskal-Wallis test are shown for the 20th Century. Low values suggest evidence against the null hypothesis that all five models have the same median.

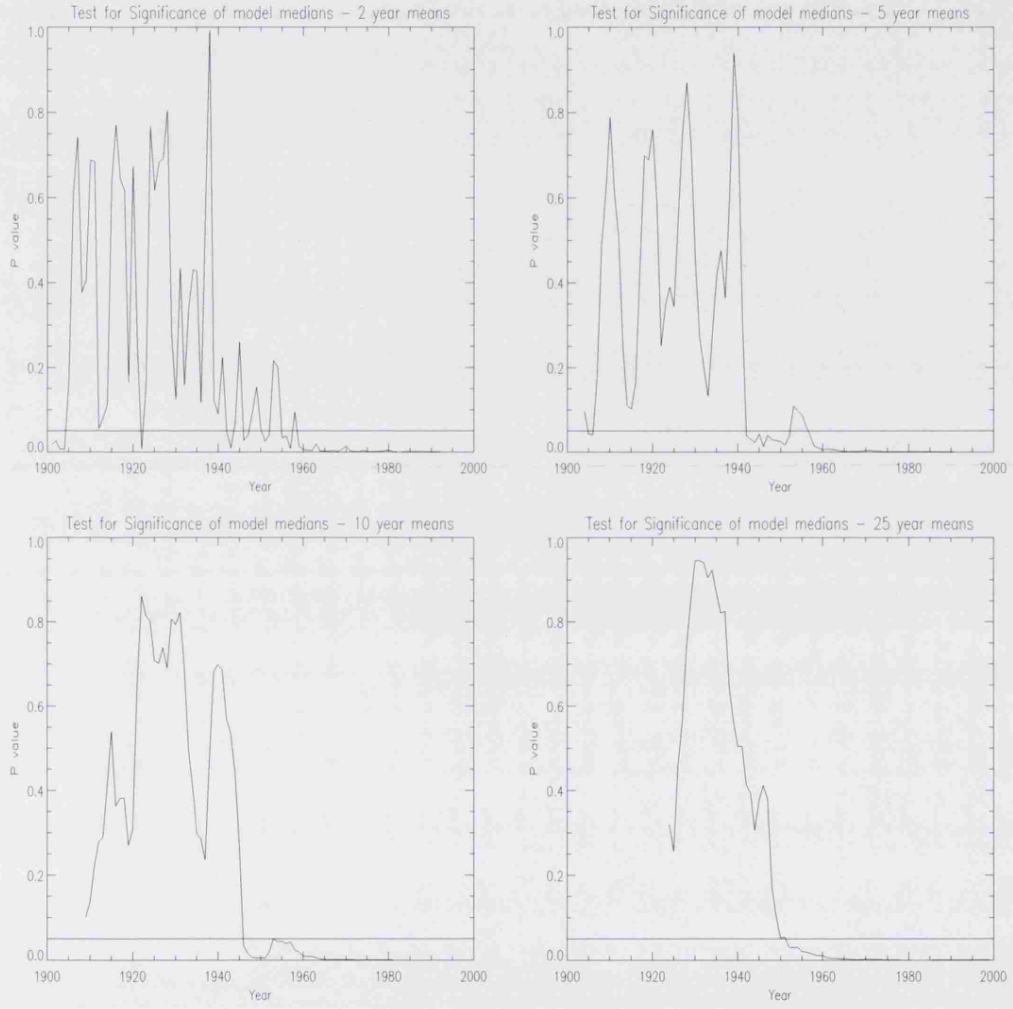


Figure 3.12: The p-values for the Kruskal-Wallis test are shown for the 20th Century for multi-year running medians of 2, 5, 10 and 25 year means respectively. In all cases, the test is non-significant during the first half of the 20th Century, the becomes significant towards the end.

Chapter 4

Introduction to the *climateprediction.net* experiment

4.1 Introduction

This Chapter gives an overview of climate models and details of the climate prediction.net experiment. The scientific communities' response to demands for decision–relevant information on future climate has largely focused on the use of complex climate models, such as GCMs. These models are explained in Section 4.2.
Section 4.3 introduces the CPDN experiment – the largest climate modelling project ever undertaken. The climate model used in this experiment, developed at the Hadley Centre, is briefly presented as well as the experimental design of the first CPDN experiment. The CPDN project provides an opportunity to evaluate some of the uncertainties in climate modelling using a multi–thousand member ensemble of simulations. Section 4.4 looks at the type of data produced by CPDN from 45,644 model simulations. Methods for estimating key climate statistics and the process of data quality control are also presented in this section. A new method of quality control is introduced that corrects for problems identified in previous methods Stainforth *et al.* (2005).

4.2 Climate Models

Climate models have been used to aid our understanding of the Earth's climate system for a long time Manabe (1975); Manabe & Bryan (1969). A variety of models have been developed, ranging from simple Energy Balance Equations Budyko (1958); Sellers (1969) to highly complex GCMs Gordon *et al.* (2000); Johns *et al.* (2006); Pope *et al.* (2000). All the climate models looked at in this Thesis are numerical, computer–implemented, representations of scientists' understanding of the Earth's climate system[1]. Climate models are primarily physically based, although they often contain empirical and statistical components.

4.2.1 On statistical methods of climate prediction

SOTA climate models are based on physics and other natural sciences. Statistical models are not widely used, despite their relative simplicity, in climate prediction. A particular reason for this is that the climate system is changing in a way that has never been previously observed. Data–driven methods that do not encode some fundamental aspect of the way the system works are unlikely to prove reliable out–of–sample. It is thought that the physical properties of the climate system will be more robust under extrapolation than those derived from a purely statistical approach. This is an important rationale for the use of complex physical models in climate modelling.

The problem of modelling the future climate is primarily one of extrapolation and thus a statistical model that assumes stationarity, or a constant relationship between variables may not be appropriate. Despite this, statistical models could be usefully applied in the field of climate prediction. Statistical models can be relevant in Numerical Weather Prediction (NWP) and seasonal climate prediction, especially when used as a "straw–man" Binter *et al.* (2009). These straw–men are often simple,

[1]It is interesting to note that climate models are not restricted to the Earth; modelling of Mars' atmosphere is also an ongoing area of research Lewis (2003).

computationally cheap, statistical models that set a minimum standard for a more complex, physical, model to outperform. The process of model evaluation in NWP and seasonal prediction depends largely on the availability of out–of–sample data; in the case of climate prediction such out–of–sample *verifications* are few due to the long time scales involved. Nevertheless, the use of simple statistical models as sanity checks for more complex climate models is a potentially fruitful area of future research.

4.2.2 Energy Balance

Perhaps the simplest way to understand the physics of the climate system is using an Energy Balance Model (EBM). A simple EBM considers the Earth's global mean temperature as a uniform sphere. Incoming energy (or radiation) from the Sun is balanced with outgoing energy from the Earth. Annually averaged incoming energy from the Sun is approximately 1370 W/m^2 (every second per metre squared perpendicular to the Sun's rays) at the top of the atmosphere. Denote this constant as S (radiation from the Sun is in fact not constant and does have an effect on Earth's climate). Some energy is reflected back into space (either from clouds or the Earth's surface) and the remaining energy warms the Earth. The reflectivity of the Earth is called *albedo*. Denote albedo as α, where α can take values between 0 (no energy is reflected) and 1 (all the energy is reflected). The net amount of solar radiation reaching the surface of the Earth is:

$$S(1-\alpha) \tag{4.1}$$

The Earth absorbs energy like a disk (from the Sun's perspective), but loses energy like a sphere (from the Earth's perspective). Energy gain for a disk of radius R is:

$$\pi R^2 S(1-\alpha) \tag{4.2}$$

Energy lost from a sphere obeys Stefan–Boltzmann's Law Boltzmann (1884); Stefan (1879) (this treats the Earth like a black body), given by:

$$4\pi R^2 \epsilon \sigma T^4 \tag{4.3}$$

where ϵ is the emissivity of the Earth (a constant close to 1) and σ is the Stefan–Boltzmann constant. T denotes the temperature in Kelvin. In equilibrium the radiation balance is zero. Thus,

$$\pi R^2 S(1-\alpha) = 4\pi R^2 \epsilon \sigma T^4 \tag{4.4}$$

Which reduces to:

$$S(1-\alpha) = KT^4 \tag{4.5}$$

where K is a constant, or

$$T \propto S^{1/4} \tag{4.6}$$

Solving this equation results in a global mean temperature of roughly -16 degrees Celsius. The actual GMST of the Earth is roughly 14 degrees Celsius New *et al.* (1999) due to the "Greenhouse effect" Fourier (1824) [1]. GHGs allow visible light from the Sun to pass through the atmosphere but absorbs a proportion of the infrared radiation emitted back from the Earth's surface at a lower frequency. Important GHGs are water vapour (H_2O), Carbon Dioxide (CO_2), Nitrous Oxide (N_2O), and Chlorofluorocarbons ($CFCs$). Changing the concentrations of GHGs may lead to changes in the greenhouse effect. Of particular interest is the rising concentration of CO_2 in the atmosphere since this change is significant and largely anthropogenic Solomon *et al.* (2007a).

An EBM is a representation of the Earth as a black body, with a uniform surface that responds thermodynamically to changes in radiative forcing. This model allows an insight to important aspects of the Earth's energy budget but is limited due to its simplicity and does not allow an understanding of possible climate feedback mechanisms (see Section 4.2.3). Although energy balance is a fundamental

[1] Actually this is a misnomer since a greenhouse works in a different way, but the phrase has stuck.

principle underlying all climate models only a basic understanding of the climate system can be gleaned from the energy balance equation. More sophisticated models that include feedbacks can be used in order to gain a deeper understanding of the climate system.

4.2.3 Feedbacks

Of key importance to the study of climate change is the idea of feedbacks. A feedback is defined here as a process in which the output in turn affects the input. A positive feedback acts to exacerbate the initial input's effect and a negative feedback dampens its effect. In the context of the climate system, an example of a feedback would be if rising levels of CO_2 were to lead to ice melting, which led to more heat being absorbed at the surface (partially due to lower albedo) which results in further surface heating. Feedback processes are non–linear and require models much more complex than an EBM to be understood. There is strong evidence that feedbacks resulting from increasing levels of CO_2 will be positive Solomon *et al.* (2007a), exacerbating the initial warming effect of GHGs. The magnitude and speed of feedbacks is still very uncertain; understanding the nature and extent of feedbacks is a key aim of climate modelling. In order to study feedbacks, complex physical models (GCMs) have been developed.

4.2.4 GCMs and Grid Boxes

GCMs are discrete, 3–dimensional representations of the Earth's climate, that numerically solve fundamental equations describing the conservation of mass, energy, momentum etc. of fluid motion. The model is configured as a set of *grid boxes* – a set of discrete points over the Earth's surface and in the vertical direction. Some models can be adapted to run at different grid resolutions but this can require a re–working of the representation of some physical processes. It might be thought that the finer the resolution of models, the better their representation of the physics (and

less parameterisation is required - see Section 4.2.5) and therefore an improvement in their ability to predict climatic changes. The tendency is for new models to work on finer resolutions, requiring more computing resources. Increasing resolution in a GCM requires an exponential increase in the amount of computational resources e.g. ceteris paribus a model with 10 times increased resolution in three dimensions requires 1000 more time to run. There is a trade–off between the complexity of the model, the number and length of simulations we are able to run.

4.2.5 Parameterisation

Climate models operate on a system of grid boxes with a resolution typically of order $\sim$10,000km^2 e.g. the HadAM3 atmosphere model operates on a resolution of 2.5 degrees latitude by 3.75 degrees longitude Pope *et al.* (2000) (this resolution is equivalent to 416km by 278km at the equator, reducing to 293km by 278km at 45 degrees latitude). At this resolution, it is not possible to capture all physical processes of interest e.g. clouds. Where modellers include such sub–grid scale processes *parameterisations* [1] can be used, based on observational studies and statistical models. Some parameterisations are well–understood and have been evaluated using observations, such as in Phillips *et al.* (2004); other parameterisations are more uncertain or may simply ignore processes altogether McGuffie & Henderson-Sellers (2006).

4.2.6 Parameter values

In contrast to the parameterisation schemes used to represent sub–grid scale processes, a number of parameter values are defined in a climate model. An example of parameters used in HadSM3 is the speed at which a convective cloud mixes into surrounding clear air (for a full list of parameters used in the CPDN experiment see Section 4.4.3). It should be noted that the role a parameter plays in the model may

[1] A parameterisation is a representation of processes that operate on length scales smaller than a grid–box in the model or that are omitted from explicit representation in the model

be different to its "real world" namesake. Model parameters are firmly rooted in a world of grid boxes and need not relate to empirical measurements of variables of the same name. There are two reasons why model parameters need not correspond directly to empirical counterparts include:

1. Measurements can mean different things when looked at on different spatial scales. For example a variety of different spatial patterns of precipitation could result in the same constant drizzle when averaged over a model grid box.

2. Parameter values in the model can be artificial e.g. expressing the speed at which cloud droplets form rain in a number may not represent anything directly empirical in the real world. In some cases, model parameters hold a tenuous relationship to anything we can measure.

It is not always clear how parameter values should be chosen in a GCM, given their partial detachment from empirical phenomena. They could be chosen to match as closely as possible observations or chosen such that they produce a more realistic looking model. It has been noted that there are O(100) uncertain parameter values in the HadAM3 atmospheric component of the HadCM3 and HadSM3 climate models Palmer *et al.* (2005). In the case of CPDN, a range of values for each of 29 selected parameters (obtained by expert elicitation, as in Murphy *et al.* (2004)) is explored. Having explored some of the uncertain parameters at pre–assigned levels in the CPDN experiment it has been shown in Sanderson *et al.* (2008) how future experimental design might be more efficient by selecting parameter values based on their likely impact on model behaviour such as CS.

4.2.7 Time steps

For the set of differential equations that make up a GCM to be computed they are first transformed into a discrete spatio–temporal set of differential equations.

These equations are then solved on discrete time steps – 30 minutes is used in the HadSM3 model in the CPDN experiment. Some parts of the model are integrated over longer time steps e.g. incoming radiation operates on longer time scales than atmospheric dynamics. The model integration scheme also takes into account the interaction between grid boxes so that the larger scale dynamics of the climate can be represented.

4.2.8 HadSM3

This subsection presents the GCM used in the CPDN experiment - HadSM3 Williams *et al.* (2001). Developed at the Hadley Centre, HadSM3 is a GCM operating on a 2.5 degree Latitude by 3.75 degree Longitude grid, with 19 vertical layers, giving roughly 140,000 distinct grid boxes. Including the 100+ physical variables used at each grid box, the dimensionality of HadSM3 runs into order 10^7. HadSM3 consists of the atmospheric model HadAM3 Gordon *et al.* (2000); Pope *et al.* (2000) coupled to a 50m deep slab ocean and sea–ice model Williams *et al.* (2001). In the vertical direction, there are 19 layers over land. These layers are not evenly distributed, in either distance or pressure. The vertical layers are narrower near the surface, where more complex physical processes occur Pope *et al.* (2000).

HadSM3 consists of about 1 million lines of Fortran code (~40Mb) and takes roughly 2–3 weeks to run one simulation (45 years of "model time") on a Pentium4 3.2GHz home PC using distributed computing software Christensen *et al.* (2005). Since the model does not include a deep ocean, and the atmosphere responds to forcings comparatively quickly, the response time of HadSM3 is faster than its deep ocean–coupled counterpart, HadCM3 Cox *et al.* (2000); Gordon *et al.* (2000). The lack of a deep ocean component allows experimental phases to be shorter thus saving computational resources, although the HadSM3 model requires flux adjustments. Two important aspects of the HadSM3 model, the slab ocean and the heat flux adjustment, are discussed in the remainder of this Section.

The Slab Ocean

The HadSM3 model uses a slab ocean, a single–layer ocean of constant depth of 50 metres throughout the globe. Many GCMs now have dynamic oceans with vertical layers extending below the ocean surface. Whilst the use of a slab ocean is a significant simplification, the model can be useful for understanding the atmospheric response to changing GHGs. The ocean operates on much longer time scales and can take 100s of years to reach equilibrium. The atmosphere, and land areas, react much quicker and it is over land and in the atmosphere that the effects of climate change will be most significant. Models with slab oceans are not suitable for studying transient climate response since the lack of ocean dynamics will result in an unphysically rapid response to forcings. In order to prevent unphysical behaviour in the slab ocean, a *heat flux* adjustment is used, described in the following subsection.

Heat Flux Adjustments

Heat Flux Adjustments (HFA) are seasonally varying artificial fluxes of heat, between the ocean and atmosphere, applied to maintain *Sea Surface Temperatures* (SSTs) close to climatological values Hewitt & Mitchell (1997); Williams *et al.* (2001). HFA is used to prevent unphysical model drifts that can occur in models using slab oceans Stainforth *et al.* (2005). This model drift is undesirable since even a model simulation with no external forcing factors (e.g. GHGs or solar variation) may display climate change.

The HadSM3 model includes HFA, calculated in the spin–up phase, as explained in Section 4.3.1. The use of HFA in the HadSM3 model, its benefits and potential problems are discussed in detail in Chapter 5. The experimental design of the CPDN experiment is explained in more detail in the next Section.

4.3 The CPDN experiment

CPDN is a publically distributed computing experiment Allen (1999); Allen & Stainforth (2002); Christensen *et al.* (2005) that harnesses the computational resources of members of the public. This allows a very large amount of experimental resources to be used to run a number of different climate modelling experiments. Similar projects include the SETI (Search for Extra-terrestrial Intelligence) project Korpela *et al.* (2001).

This Section explains the details of the first CPDN experiment. This experiment has produced the largest ensemble of climate simulations to date. Over 300,000 members of the public have contributed to producing in excess of 200,000 simulations. CPDN has run several experiments including a comparison of pre–industrial to doubled CO_2 climate, a transient forcing experiment, and others including a sulphur cycle and an experiment to look at possible shutdown of the thermohaline circulation (see *www.climateprediction.net* for more details on these experiments). It is the first of these that is analysed in this Thesis, which was also the experiment with the most available simulations at the time of analysis. One of the key aims of this experiment is to better understand ICU and the effect of parameter perturbation to explore model uncertainties.

The particulars of the experimental design, as well as some issues that arise in the analysis of the data set are presented in this section.

4.3.1 CPDN Experimental Design

The first CPDN experiment has created a *grand ensemble*, a set of ICEs, run under parametrically perturbed versions of the HadSM3 model. Parameters are perturbed from a standard set to form *model versions*. The parametrically unperturbed model containing the standard set of parameters is referred to in this Thesis as the standard HadSM3 model. Each model version is run with multiple ICs plus some duplicate simulations that are used to verify the experimental design (especially whether the

use of publically distributed computing is a reliable one for climate modelling). For the standard HadSM3 model more ICE simulations are sent out than for perturbed parameter model versions. This larger ICE allows a more detailed analysis of the models' internal variability and is explored in Chapter 6.

Calibration Phase

Each simulation in the first CPDN experiment consists of three distinct 15–year phases. The first 15 year phase of the CPDN experiment, the *calibration* phase, is run in order to calibrate the HFA field. The HFA is required to account for anomalies in the coupling between the atmospheric and ocean components of the HadSM3 model and to produce a stable control climate. Different model versions can require significantly different adjustments. In the CPDN experiment, the HFA is calibrated for each simulation. Thus, each simulation uses a different HFA. In a different experimental set–up, the same HFA should be used for every simulation within a model version since simulations sharing the same dynamics should require the same HFA (assuming the initial condition has no significant effect on the HFA). This could be achieved by averaging the HFA over each ICE, providing a more robust estimate of the common HFA, or by only calibrating the HFA for one simulation, saving computational resources.

The HFA field is calibrated such that the model's SST matches a 1961-1990 *re–analysis* of observations New *et al.* (1999). The HFA field used in subsequent phases is defined for each grid box and varies as a seasonal climatology Piani *et al.* (2005), but remains fixed from year to year. See Chapter 5 for a detailed discussion on this phase and characteristics of the HFA field in reducing model drift, its convergence and its relationship with CS.

Control Phase

The second 15 year phase, the *control* phase, is run under pre–industrial CO_2 concentrations, as in the calibration phase. The HFA, derived from the calibration phase, is held fixed from year to year but varies seasonally during this phase. The control phase can also be used as a reference climatology for the third phase, where pre–industrial CO_2 concentrations are doubled. The control phase also provides a measure of the model's internal variability and is used in quality control to pick up unstable or drifting model simulations. The methods of quality control applied are discussed in section 4.4.2.

Doubled CO_2 phase

At the start of the third 15 year phase (the doubled CO_2 phase), CO_2 concentrations are instantaneously doubled. The use of an instantaneous forcing means that this phase can be shorter than would be required in a transient forcing set–up (gradually increasing CO_2 over many years). Since the CO_2 forcing applied is highly artificial, there is no direct comparability of the experiment's time series to the Earth's climate; the effects of doubling CO_2 are isolated in the context of estimating the equilibrium response of the model. Both the control phase and the doubled CO_2 phase are initiated from the end point of the calibration phase. The model response is then analysed, especially the GMST response. For some models, the GMST stabilises by the end of the third phase, whereas others are still warming Stainforth *et al.* (2005). Since not all the models have reached equilibrium, the CS must then be estimated using the different methods explained in Section 4.4.1. Section 4.4 gives information on the data produced by the CPDN simulations.

4.4 Data

Climate models produce a huge amount of data due to their high dimensionality and large number of variables. HadSM3 operates in an order 10^7 dimensional space and a 45 year simulation consists of almost a million time steps. Recording all the model output would require a large amount of disk space. In fact, only a sub–set of meta–data is retained from each simulation – it is this sub–set of summary data that is explained in this Section. Despite saving only a fraction of the entire model output, the CPDN experiment has produced quantities of data on the order of terabytes (~15Mb per simulation).

This Section describes three specific sets of data – the ICE of the Standard HadSM3 model, the initial release of data (2578 model simulations) and a larger set of 45644 simulations.

Section 4.4.1 explains an important statistic in climate modelling, climate sensitivity (CS).

4.4.1 Climate Sensitivity

A number of different definitions of CS exist. In this Thesis, CS refers to equilibrium climate sensitivity, defined as the equilibrium global mean temperature response to a doubling of CO_2 concentrations. CS can be understood more intuitively as the eventual amount of global warming we expect to occur if CO_2 concentrations were doubled and then maintained. CS is an important component of our understanding of climate change and has attracted a wide range of interest in the scientific community Annan & Hargreaves (2006); Arrhenius (1896); Manabe (1975); Roe & Baker (2007); Solomon *et al.* (2007a); Stainforth *et al.* (2005). Whilst the output of an entire model simulation can be reduced to the single scalar statistic of CS, much important information is lost as shown in Chapter 8

The CS of a simulation can be estimated by taking the average GMST over a long simulation with pre–industrial CO_2 concentrations and subtracting this from the

average GMST over a long simulation with doubled pre–industrial CO_2. Each phase is 15 years long. Where there is still transient warming and it can not be assumed that the simulation has reached a new equilibrium CS must be estimated from the time series of GMST.

Three different methods for estimating CS are explained here. Two of these methods are based on the imposed radiative flux due to doubling CO_2 concentrations and the heat capacity of the oceans[1].

The third method is based on an exponential fit to the temperature change under a doubling of CO_2 concentrations. This statistical method uses fewer physical quantities as the other two methods, but gives comparable results (shown in Section 4.4.1).

Heat Capacity

Two of the physical variables used to calculate CS are the climate system's heat capacity and the top of atmosphere radiative flux imbalance (TOA flux). The heat capacity is an expression of the rate at which the climate system warms. Figure 4.1 shows the variability in estimated heat capacity over 1460 quality controlled simulations. Most of the simulations' heat capacities lie in the region of 3.4–4.4 W/m^2. The different heat capacities of simulations means that they warm at different rates. This has an impact on the calculation of CS.

Radiative forcing

Radiative forcing relates to an imbalance between the incoming and outgoing radiation from the model. This can be understood as an imbalance in the Energy Balance equation in Section 4.2.2. The equation shows how temperature (or albedo) can change when the balance of energy changes. When a change in radiation occurs the model's temperature adjusts, over a period of time, to a new level. Once this period

[1] Heat capacity measures how effectively a substance stores heat such that: $\delta(H) = C.\delta(T)$, where H is heat, C is the heat capacity and T is the temperature.

of transition, which can take 10s or 100s of years, is complete, a new equilibrium is reached. Both the speed and nature of the period of transient change and the equilibrium value are of interest to climate scientists and decision–makers.

The TOA flux has been estimated at 4.37 W/m^2 in the IPCC Second Assessment report Houghton *et al.* (1995), revised to 3.7 W/m^2 in the Third Assessment report (Section 6.3.1) Houghton *et al.* (2001) and held constant for the Fourth Assessment Report Solomon *et al.* (2007a). Taking into account the importance of the uncertainty in the estimation of the ocean heat capacity and the radiative forcing due to doubling CO_2 could considerably change estimates of CS. In particular, estimates of CS could change by up to 20% depending on whether the best guess of the radiative forcing resulting from doubled CO_2 is taken from the IPCC's Second or Third report.

Method 1: Physics–based Exponential fit

This method is based on the GMST response to a doubling of CO_2, allowing for uncertainty in the heat capacity of the model climate, as used in Stainforth *et al.* (2005). Four physical quantities are used to estimate the CS parameter, λ. The last 8 years (of the 15 year phase) of the pre–industrial phase are taken as the control GMST. The TOA radiation imbalance, F, is calculated as the incoming shortwave radiation minus the outgoing longwave and outgoing shortwave radiation (incoming longwave radiation is assumed to be negligible). The heat capacity of the climate hc, is calculated from the change in temperature and change in radiative forcing between the first five years of the doubled CO_2 phase and the control phase.

Finally, the time series of GMST change between the doubled CO_2 phase and the control temperature is used, from phase year 8 onwards, as an annual mean. The following exponential fit is then carried out using a gradient–expansion algorithm to compute a non–linear least squares fit Press (1992).

$$\frac{dT}{dt} = \frac{F_{x2co_2}}{hc} exp(\frac{\lambda}{hc}) \tag{4.7}$$

Where $\frac{dT}{dt}$ is the change in GMST, T, with respect to time, t, in years. CS is estimated by $\frac{F_{x2co_2}}{\lambda}$.

Method 2: Gregory Method

This method is based on Gregory's equation introduced in Gregory *et al.* (2002) and uses a linear fit between the doubled CO_2 phase GMST difference (the time series of GMST change seen under a doubling of CO_2) and the TOA atmospheric radiation imbalance, as defined in Method 1. The linear fit is accomplished using minimum least squares (see *linfit.pro* in the IDL documentation for more details). The CS is then estimated by $\frac{F}{\lambda}$, where λ is the negative of the gradient of the linear fit, and F is the intercept.

Model 3: Statistical Exponential Fit

This method uses a statistical fit to the doubled CO_2 phase temperature to estimate the equilibrium state.

$$\frac{dT}{dt} = A(1 - exp(B))$$

Where $\frac{dT}{dt}$ is the change in GMST, T, with respect to time, t, in years. CS is estimated by the parameter A. Unlike the other two methods, this method does not use any physical information from the model, other than GMST. In principle this method might be extended to estimate the sensitivity to increased CO_2 for other variables or length scales.

Other possible methods for calculating CS might be to fit a single exponential curve to all the members of an ICE simultaneously, or the method used in Knight *et al.* (2007) where the heat capacity is taken as fixed.

Does it matter which method is used?

This subsection looks at the relationship between the three methods for estimating CS presented and whether it makes any important difference which is used. Figure 4.2 shows the three methods plotted against each other for 1460 simulations from the initial release of data presented in Stainforth *et al.* (2005). The diagonal panels show a perfect correlation where each method is plotted against itself. There is a very strong linear relationship between the three methods, with small deviations from the line $x=y$. Figure 4.3 shows the range of values estimated for each run (maximum estimated CS of the three methods minus the minimum) verses the mean. The range is typically on the order of 0.1 degrees Celsius, but can be as high as 1 degree. This range of values estimated is typically small, with a median of 0.1572 degrees Celsius (tenth and ninetieth percentiles are 0.0487 and 0.401 degrees respectively), far less than most quoted ranges of CS, e.g. 2–4.5 degrees in Solomon *et al.* (2007a) and 1.9–11.5 in Stainforth *et al.* (2005). These results suggest that it does not make a significant difference which method is used for the purposes of estimating CS.

Knight *et al.* (2007) takes the radiative imbalanced due to doubling CO_2 to be 3.74 W/m^2 in calculating CS for a similar CPDN data set. Figure 4.4 shows the estimated TOA Flux over 1460 perturbed–physics simulations of the HadSM3 model using Method 1. Both values (4.74 and 3.7 W/m^2 in the Second and Third Assessment Reports, respectively) adopted by the IPCC are contained within the range of values seen. Method 2 gives different results, seen in Figure 4.5. Whilst the peak of the histogram is still around 3–4 W/m^2, more simulations have higher estimated TOA fluxes. Using Method 1, most simulations have less than 4.5 W/m^2, whereas using Method 2 there are simulations with TOA fluxes up to 6 W/m^2 and above. This discrepancy is likely due to the method of estimation and explains some of the differences in estimates of CS produced by the different methods for estimating CS outlined above.

Unless explicitly stated otherwise, Method 1 is used to calculate CS in this thesis. This method is chosen since it uses more physical information from each simulation than the other Methods presented. Results are not thought to be sensitive to the Method chosen.

High Climate Sensitivity Simulations

A key feature of the CPDN grand ensemble is that some simulations show a large magnitude of warming in response to doubling CO_2. Results presented in Chapter 7 show that there are 350 simulations ($\sim$ 1.5% of simulations) with an estimated CS of over 10 degrees Celsius.

In the case of these very high CS simulations (defined in this Section as greater than 10 degrees Celsius), the 15 year doubled CO_2 phase does not allow sufficient time for an equilibrium to be reached. The time series for the 350 simulations with CS over 10 degrees (using Method 1) are plotted in Figure 4.6. Also shown are 822 simulations with a CS close to 2 degrees Celsius (simulations with a CS between 1.9 and 2.1 degrees Celsius, taken from the larger set of 45644 simulations) for comparison. The 2 degree simulations seem to have reached an equilibrium by the end of the phase, whereas the simulations with CS greater than 10 degrees have not. Whilst it appears that simulated warming in the final phase follows an approximately exponential pattern in many simulations, it may not be possible to accurately estimate very high values of CS. Despite this, the range of estimates for CS across the 3 Methods is still less than 1 degree Celsius. It is relevant to note that when the HadSM3 model was developed such high levels of CS were likely not considered possible. As such, it might be more prudent to judge any simulation with an estimated value of CS much above 10 degrees Celsius simply as an extremely severe case of global warming, without assigning much importance to the details that distinguish between such simulations.

4.4.2 Quality Control

This Section presents a new method of quality control for the analysis of data from CPDN and other similar experiments.

Not all of the simulations produced by the CPDN experiment are suitable for analysis for a number of reasons. Two such reasons are that **1)** some simulations have missing data and **2)** some simulations show unphysical behaviour to a degree that renders them useless for understanding CO_2–forced climate changes. Such simulations should be neglected from analysis on physical grounds. Four stages of quality control are presented here. These stages do not aim to be exhaustive but rather aim to purge simulations with gross internal inconsistencies.

Quality control methods were developed based on the initial release of 2578 CPDN simulations, then applied to the set of 45644 simulations. The four stages of quality control are **1)** Rejection of simulations with insufficient data, **2)** Rejection of simulations with significant global temperature drift during the control phase, **3)** Rejection of simulations with evidence of a specific, unphysical, regional cooling feedback in the East Pacific and **4)** Simulations with extremely rapid shifts in climate on seasonal timescales. These four stages are now explained in more detail:

1. Any simulations that contain insufficient information to calculate CS are rejected. A similar criteria of ruling our simulations with missing data was applied in Sanderson *et al.* (2008). It is in the nature of a distributed computing experiment Stainforth *et al.* (2002) that some simulations will be returned without the full complement of information. These simulations are not included in the subsequent analysis of quality controlled data. It is assumed that there are no systematic biases in results due to eliminating simulations with missing data. The errors are believed to arise to due numerical errors or participants dropping out and not any relevant properties of the simulations' physics. Of the full set of 45644 simulations, 31793 pass this first stage of quality control.

2. Any simulations that show a significant GMST drift in the control phase are ruled out as unstable. A critical level of 0.02 degrees Kelvin per year is considered unstable, as in Knight *et al.* (2007); Stainforth *et al.* (2005). This degree of drift is very close to the maximum magnitude of drift seen in the standard HadSM3 model (~0.019 degrees Kelvin per year). Where simulations show significant climate change during a period without any forcing effects they are dismissed. In the presence of significant GMST drift, it would be difficult to tell whether the climatic changes seen under a doubling of CO_2 occur as a result of the increased GHG forcing, or due to some other mechanism. Of the 31793 simulations that pass the first stage of quality control, 23050 pass this second stage.

3. The main cause of unphysical GMST drift in both the control phase and the doubled CO_2 phase is an unphysical negative feedback mechanism that occurs in the East Pacific. This problem arises as a result of the slab ocean's inability to re–distribute heat through ocean currents. Cool SSTs in a specific area of the East Pacific (known henceforth as "Area 51") and low–lying clouds cause a run–away feedback that affects SSTs first locally, then on the global scale. The effect is small at first, then becomes significant, leading to a drop in GMST of up to 27 degrees Celsius Stainforth *et al.* (2005).

   An "Area 51" statistic is presented here in order to rule out simulations with an East Pacific negative feedback. This statistic is defined as the difference between an identified problem Pacific grid box (located at longitude 78.75 West and latitude 2.5 North) in the doubled CO_2 phase and the same grid box in the calibration phase, adjusted by an Atlantic grid box of the same latitude (called Area 52 here), averaged over the last 8 years of each phase. Area 52 is defined by the grid box at longitude 48.75 West and 2.5 North. The Area 51 statistic, A, is given by:

$$A = (area51_3 - area52_3)) - (area51_1 - area52_1) \tag{4.8}$$

Where $area51_i$ represents the value of the Area 51 grid box in phase i and $area52_i$ the value of the Area 52 grid box in phase i. The statistic A detects simulations with a significant local cooling that might be missed when looking at global mean temperature drift, as applied in the second stage of quality control. Figure 4.7 shows the distribution of values for the Area 51 statistic after stage 1 of quality control. This distribution is bi–modal; the larger peak is about 0 and the smaller about -27. The first peak about 0 represents simulations that have not yet developed a significant feedback in this area. The second peak represents simulations that have drifted significantly already. Applying the second stage of quality control i.e. looking at complete simulations with non–significant GMST drift, this distribution changes to that shown in Figure 4.8.

Having eliminated simulations with significant GMST drift, the second peak (around -27 degrees) disappears, but there is still a long tail to the distribution of A. Whilst the distribution now appears uni-modal, there are still a number of problematic simulations with values of A as low as -30 degrees. The statistic A is used to overcome the problem seen in Stainforth *et al.* (2005), where simulations are admitted for analysis that show localised unphysical cooling. Any simulations for which this anomaly is less than -15 degrees Kelvin are ruled out as having an unphysical regional feedback. The value of -15 degrees is chosen to eliminate simulations with a very strong local cooling. The choice of this value is a trade–off between rejecting simulations with a significant, unphysical, feedback and not rejecting simulations without this problem.

It is not clear which level should be chosen as critical in the presence of a

continuous tail. Here, a value of -15 is chosen so as to be sure of not ruling out simulations that have not drifted. Choosing a higher value, say -5 degrees, would not rule out very many more simulations (22329 simulations would pass all stages of quality control using -5 degrees, compared to 22723 using -15 degrees). It would be important not to set the critical value too high, for fear of ruling out simulations that show a physically consistent global cooling (this would be an important result). 23050 simulations pass the first two stages, and if no Area 51 quality control is applied 22871 pass the fourth stage of quality control. The exact value of the Area 51 statistic is not critical to the results of the analysis carried out in this Thesis – cut–off levels for A of -5, -10, -15 and -20 allow 22329, 22643, 22723, 22775 simulations respectively.

4. Simulations that show extremely rapid seasonal changes in GMST are ruled out (of more than 20 degrees Celsius in any 3 month period). These jumps could be due to a numerical error (or lost data) for a short period of time that show up in the available time series as a more moderate jump in seasonally averaged temperature. It is judged here that a jump of more than 20 degrees in a single season is unphysical and thus those simulations should be dismissed from analysis. It is not expected that all simulations with unphysical jumps are detected by quality control, only the worst offenders. This last stage of quality control eliminates simulations with numerical errors that might distort the models' results.

The four stages of quality control and the number of simulations remaining after each stage are shown in Table 4.1. Approximately half (49.8%) of all simulations pass all four stages of quality control and are then available for further analysis. The effect of this quality control process can be seen in Figure 4.9. The initial release of 2578 simulations, presented in Stainforth *et al.* (2005), are shown in their original form (panel (a), with 2578 simulations), after two stages of quality control (panel (b), with 1460 simulations) and all four stages of quality control (panel (c),

Stage	Description	Critical Value	Simulations
0	None	NA	45644 (**100%**)
1	Insufficient data	NA	31793 (**69.7%**)
2	Control phase GMST drift	0.02 degrees/year	23050 (**50.5%**)
3	East Pacific regional cooling	-15 degrees	22899 (**50.1%**)
4	Unstable time series	20 degrees	22723 (**49.8%**)

Table 4.1: The four stages of quality control are given with critical values, where relevant, and the number of simulations remaining at each stage (also shown as a percentage of the initial number of simulations in brackets).

with 1447 simulations). Applying the first two stages of quality control means ruling out simulations without a full complement of data or simulations with a significant GMST drift (as applied in Stainforth *et al.* (2005)). Applying the full quality control as described in this Section also rules out simulations that cool unphysically in the doubled CO_2 (all these simulations cool due to the Area 51 problem). Also shown in Figure 4.9 (panel (d)) is the result of applying full quality control to a 45644 member ensemble, leaving 22723 simulations. Whilst no simulations show sustained cooling in the doubled CO_2, one simulation in particular shows unstable behaviour; this can be seen in panel (d) of Figure 4.9 in years 38–40 where one simulation shows a significant, but temporary dip. By the end of phase year (year 45 in the plot), this simulation has warmed by more than any other simulation (about 10 degrees above its mean control phase temperature). This simulation was not rejected by the quality control procedure, but it is likely that further investigation might reveal unphysical behaviour in this model.

It is interesting to note that parameter perturbation in the CPDN ensemble can result in differences of up to 1.5 degrees Celsius in absolute GMST during both the calibration and control phase of the experiment, approximately half the difference between structural models shown in Chapter 3. This difference can be seen in Figure 4.9. This suggests that parameter perturbation can have, in some cases, less effect on the physical properties of a climate model than structural differences.

All the simulations with an unphysical cooling in the set of 2578 simulations are deemed unphysical using the quality control processes proposed in this Section. It

is important to note that the process of quality control does not assume that simulations should warm and rule out cooling simulations correspondingly. Rather, any simulations that are unstable in the control phase are excluded, whether this drift is a warming or cooling. Furthermore, the critical level of cooling used to detect an Area 51 anomaly is chosen to be low such that no simulations that cool due to another reason are mis–diagnosed and rejected accordingly.

In general, the quality control aims to be conservative in ruling out simulations so that only clearly unphysical simulations are dismissed. Quality control should be careful not to ignore the potentially most important simulations of all e.g. simulations showing global cooling or rapid fluctuations in climate.

4.4.3 Data Format

For each simulation, data is available for each of the three 15 year experimental phases. For the calibration phase, additional data is recorded on the HFA (recorded as an 8 year mean field for each month). Each simulation produces data on a range of meteorological variables such as 1.5 metre surface temperature, precipitation rate, cloud cover, wind speed etc. on global, regional and grid box scales. Only a small sample of the model output is stored due to restrictions on storage and processing time. There are two general types of data available:

1. Global (or regional) mean data as a monthly time series. 22 regional averages are defined as in Giorgi & Mearns (2000) over land areas as well as areas denoting the tropics, Northern and Southern extra–tropics and the North and South hemispheres.

In order to calculate an area–weighted statistic, T_g using the following method is used:

$$T_g = 2\pi r^2 \frac{(sin(lat_j) - sin(lat_{j-1}))}{nlons} \tag{4.9}$$

Where r = the radius of the Earth (~6371km), $nlons$=number of longitude grid boxes (96 for the HadSM3 model). lat_j=the mid–point of the j—th grid box, normalised to the range (-1,1); radians are calculated as the latitude in degrees times π, divided by 180.

2. Seasonal field data, recorded for all grid boxes, as 3 month averages ([December January February], [March, April, May], [June, July, August], [September, October, November]). These seasonal averages are henceforth referred to as DJF, MAM, JJA and SON respectively. These fields are found from the last 8 year mean of each phase; there are 12 fields available for each variable – one for each phase and season.

Initial Condition Ensembles

For each model version (set of parameter values) between 1 and 10 ICE members are available. For the parametrically unperturbed model (the standard HadSM3 model) 64 simulations are available. ICEs were obtained by using the same parameter values and perturbing ICs. ICE perturbation was achieved by perturbing a single ocean grid box by a small amount (for most ICE members by between -0.1 and +0.1 degrees Celsius, from -3 to +3 for the standard HadSM3 ICE). This allows the same deterministic HadSM3 model version to be run a number of times, allowing of quantification of internal variability. The data set of 45644 simulations can be viewed as a set of 13535 ICEs, containing an average of just over 3 members in each. The standard HadSM3 model ICE is investigated in detail in Chapter 6.

Initial release of metadata

The original release of data from the CPDN experiment has been analysed in Knutti *et al.* (2006); Sanderson *et al.* (2007); Stainforth *et al.* (2005). Analysis of this data set showed a then unprecedented range of behaviour in GCMs, with values of CS ranging from below 2 to over 11 degrees Celsius. Such a wide range for CS has attracted an increase in efforts to constrain the range of values of CS Annan *et al.* (2006); Hegerl (2006); Knutti *et al.* (2006). The question of constraining ensembles of climate projections is discussed in Chapter 7. This data set consists of 484 model versions, making up 2578 model simulations in total.

Grand Ensemble of 45644 model simulations

This Section explains the simulations forming the CPDN *grand ensemble* Stainforth *et al.* (2005). Up to 21 parameters values in the HadSM3 model are perturbed in this experiment, each with between 2 and 4 levels determined by expert elicitation, following Murphy *et al.* (2004). A sparse (a sample of 45644 from 3^{27} possible parameter combinations is very sparse) Latin hypercube sampling strategy is adopted. The 21 parameters perturbed in the CPDN experiment were chosen as being potentially important and are varied within expert-elicited ranges as in Knight *et al.* (2007); Murphy *et al.* (2004); Stainforth *et al.* (2005). Perturbing atmospheric and sea–ice parameters allows a sampling of uncertainties in climate feedbacks Collins *et al.* (2006). The parameters perturbed in the CPDN experiment are re-produced below, as from
http://www.climateprediction.net/science/parameters.php):

- *vf1 Ice fall speed through clouds - important for the development of clouds and determining type (rain, sleet, hail, snow) and amount of precipitation*
- *ct This relates how quickly cloud droplets convert to rain.*
- *rhcrit "critical relative humidity" relates the grid box scale atmospheric hu-*

midity to the amount of cloud in that grid box

- $cw-land, cw-sea$ *This relates how much water there is in a cloud to when it starts raining, which is dependent on the condensation nuclei concentration - the more condensation nuclei there are (bits of dust, salt etc. in the atmosphere on which raindrops can form) the smaller the raindrops.*

- *entcoef This parameter determines how rapidly a convective cloud (imagine a plume rising over a power station, or a bit thunder cloud) mixes in clear air from around it.*

- *eacf Empirically adjusted cloud fraction This calculates how much cloud cover there will be when the air is saturated.*

- *ice − size This gives an effective radius for ice crystals in clouds i.e. what radius would they have if they were perfectly spherical. It is important in the radiation scheme, to calculate how much incoming or outgoing radiation is reflected etc.*

- $i-st-ice-sw, i-cnv-ice-sw, i-st-ice-lw, i-cnv-ice-lw$ *These parameters all allow for non-spherical ice particles in the radiation scheme.*

- *asym − lambda This has to do with how rapidly air mixes by turbulence in the boundary layer (the layer of the atmosphere closest to the Earth).*

- *G0 This has to do with the fact that the ability of turbulence to mix air varies with how stable the air is - the more stable the air, the less turbulent mixing you get.*

- *z0fsea This parameter governs the transfer of momentum and energy between tropical oceans and the air (wind) above them.*

- *charnock This parameter governs the transfer of momentum and energy between seas and the air (wind) above them.*

- *$r-layers$ This is related to the number and size of plant roots in the soil - and, consequently, to how water is taken up from the soil and into the atmosphere by plant transpiration.*

- *$eddydiff$ This parameter governs the diffusion of heat from the slab ocean to ice, where there is sea-ice present in the model.*

- *$start-level-gwdrag$ Gravity waves are waves in the atmosphere for which gravity is the restoring force - think of air passing over a mountain, it is forced upwards over the mountain, and then gravity will pull it back down, resulting in an oscillation (you often see clouds form downstream of mountains as a result). The air particles oscillating in these waves tend to lose energy because of friction (drag), and this energy manifests itself as heat. This parameter determines the lowest model level on which gravity wave drag is applied*

- *$kay-gwave, kay-lee-gwdrag$ These parameters govern the way that gravity waves are formed as air interacts with surface features, such as mountains.*

- *$Alpham, dtice$ These have to do with the fact that the albedo (reflectivity) of sea ice varies with temperature.*

- *$diff-coeff, diff-exp$ Diffusion coefficients and exponents govern how quickly something spreads through the material it is in - so, for example, if you put a drop of oil dyed purple into a beaker of un-dyed oil, how rapidly the dyed oil mixes with the oil around it until all the beaker has the same colour. Diffusion refers to mixing due to the random motion of particles, rather than turbulent mixing which happens when there are actual vortexes mixing things (which would happen if you stirred the beaker with a spoon). In the case of the atmosphere, the horizontal diffusion coefficient and exponent determines the rate of diffusion of heat from a warm air mass to a cold one.*

- *$diff-coeff-q, diff-exp-q$ These diffusion parameters determine the*

rate at which water vapour diffuses from a very humid air mass to a relatively dry one.

Duplicate Simulations

A number of duplicate simulations are sent out in order to check that distributed computing is a reliable method for generating ensembles of GCM simulations. In particular, it is important that any differences between duplicate simulations, arising from flaws in the experimental design, are small in comparison to the model behaviour being investigated. One source of differences between duplicate simulations is the use of different computing architecture and processors. The differences in computing architecture has been analysed by Knight *et al.* (2007) and differences shown to be small in comparison to the effect of parameter perturbation. The magnitude of the differences, in terms of CS is of the same order of magnitude as the impact of different initial conditions. The presence of differences even between duplicate simulations does have important consequences for the interpretation of distributed computing experiments. The results of a distributed computing experiment are therefore not always strictly reproducible. Duplicate simulations are not looked at in this Thesis.

4.5 Conclusion

The motivation behind climate modelling and the basics of state–of–the–art climate models has been introduced in this Chapter. The rationale behind the use of physical models has been discussed and key concepts in climate modelling such as the use of parameterisations, flux adjustments, energy balance and feedbacks are introduced. The details of the HadSM3 model, used in the CPDN experiment have been explained in this Chapter.

The experimental design of the CPDN data analysed in this Thesis has been explained, as well as some issues that arise in the analysis of the data set. In particular,

the methods of estimating CS and the process of quality control have been discussed. The different data sets that are used in this Thesis are explained; **1)** the initial release of a grand ensemble of 2578 simulations, **2)** a 64 member ICE of the standard HadSM3 model and **3)** a grand ensemble of 45644 simulations. Subsequent analysis is restricted by the availability of data; the accommodation of a multi-thousand grand ensemble means that limited model output can be stored from each simulation. Chapter 5 will look at data from data sets **1)** and **2)**. Chapter 6 focuses on set **2)** and Chapters 7 and 8 looks primarily at set **3)**.

Original work presented in this Chapter are:

1. A new method of quality control, correcting for problems identified in previous quality control methods Stainforth *et al.* (2005). An unphysical local feedback in the East Pacific is detected using a local anomaly statistic. This method was shown to eliminate simulations with significant global cooling which fail to be detected when using global mean statistics.

2. Features of the CPDN experiment have been documented for the first time e.g. the availability and description of data, experimental design and issues in data analysis. Such documentation is important for other studies based on CPDN data sets.

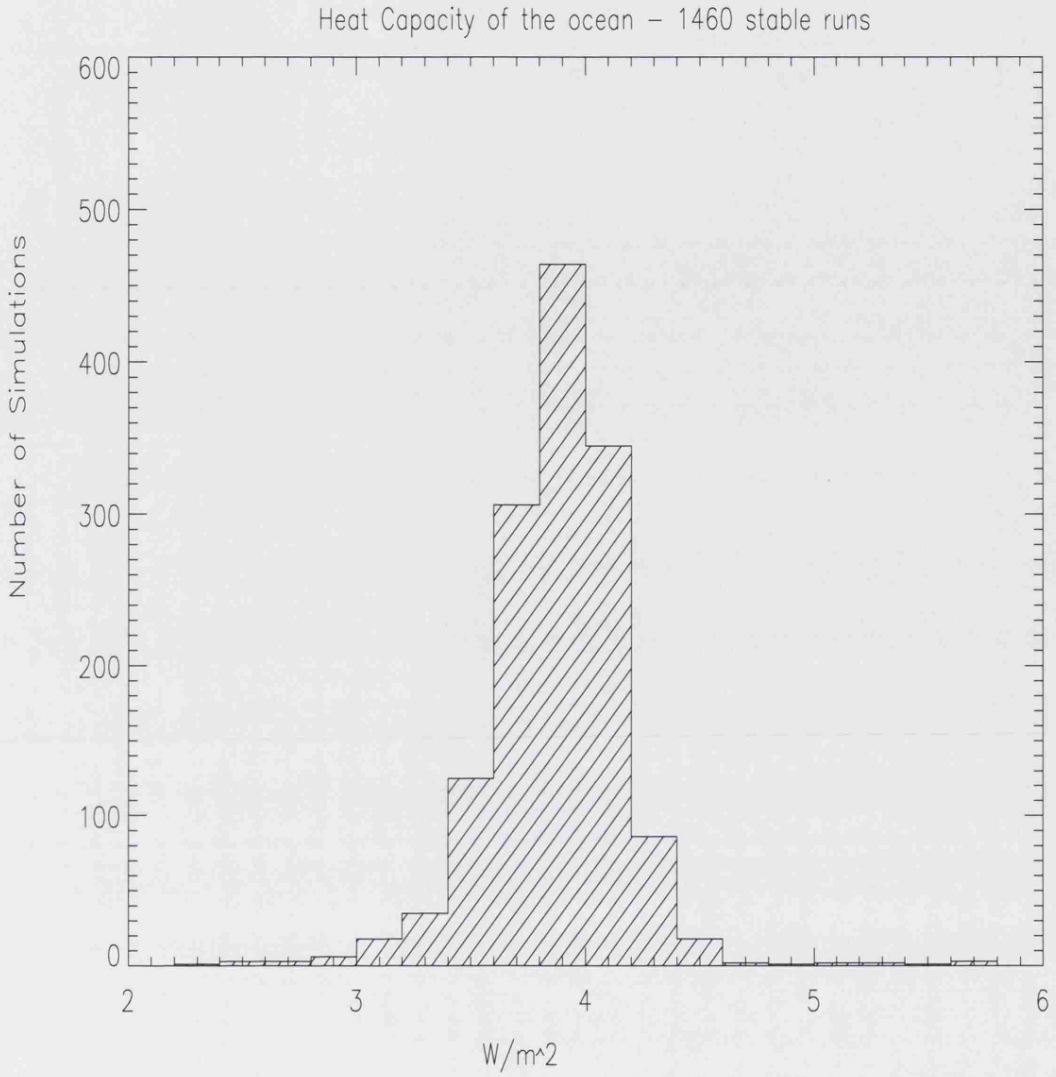


Figure 4.1: The global mean heat capacity in the doubled CO_2 phase is plotted over 1460 quality controlled simulations.

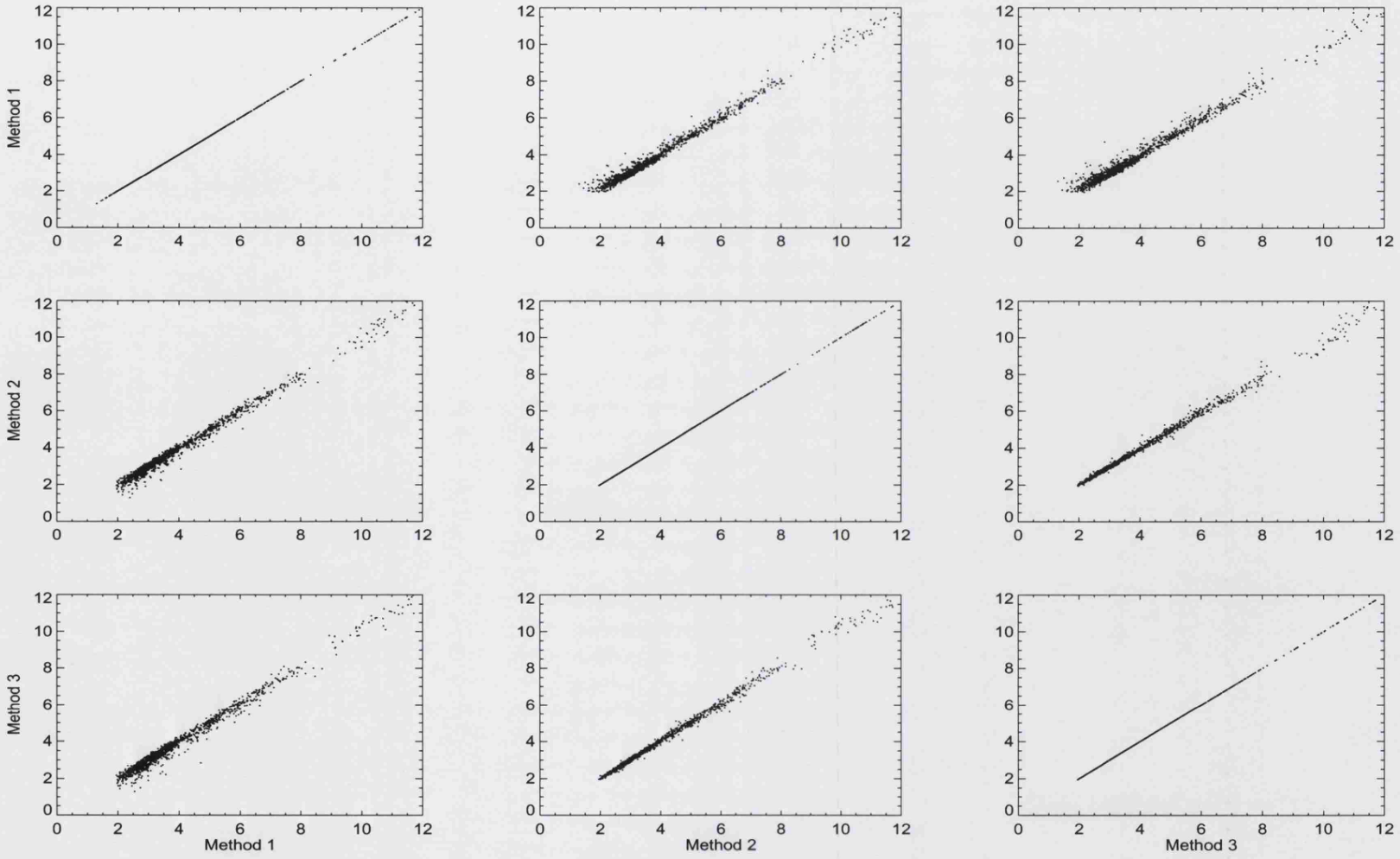


Figure 4.2: Estimates of CS are plotted against each other for three different methods over 1460 quality controlled simulations. There is a strong linear relationship between each of the methods.

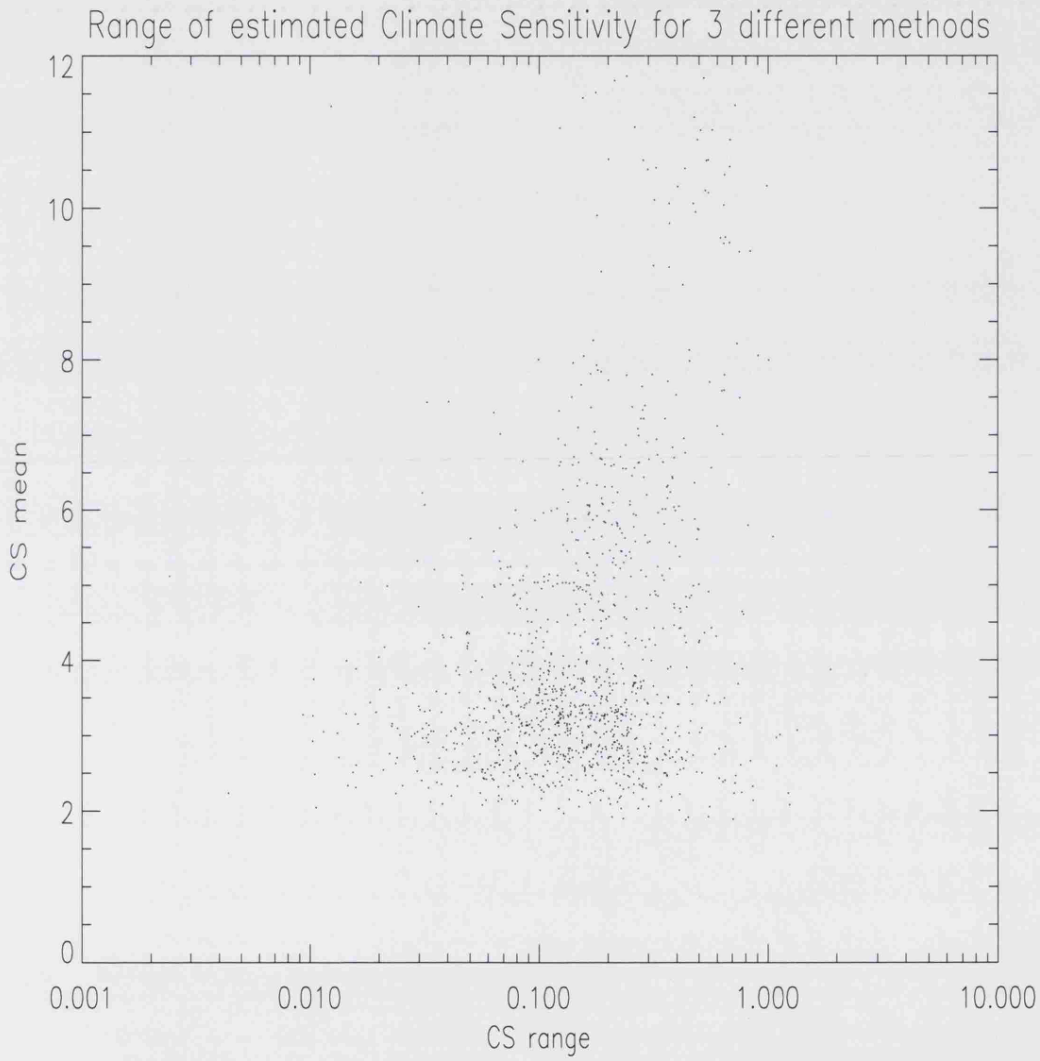


Figure 4.3: The range in estimated CS for three different methods is plotted against the mean for 1460 quality controlled simulations.

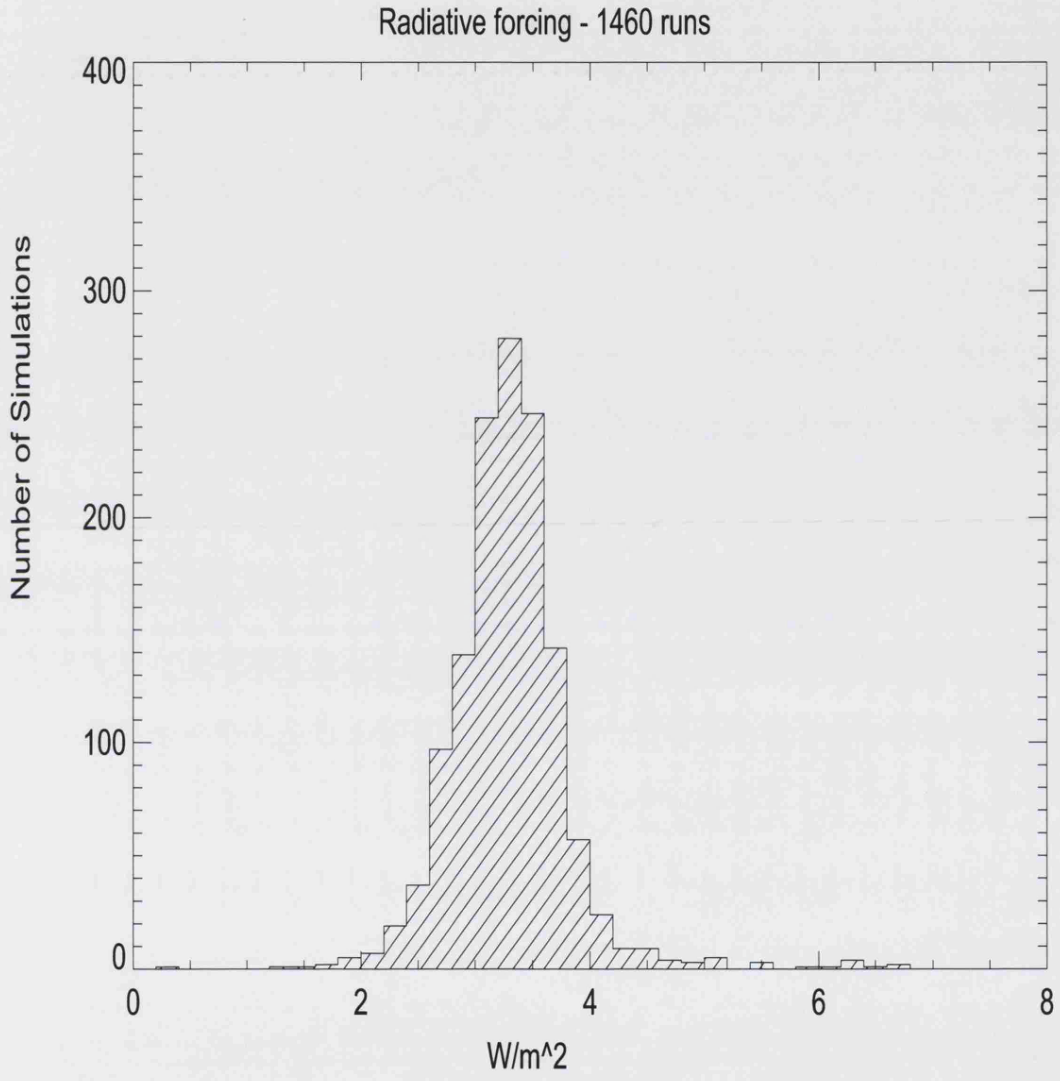


Figure 4.4: The distribution in top of atmosphere radiative flux imbalance in the doubled CO_2 phase is shown for 1460 quality controlled simulations. The method used to estimate this flux imbalance is the exponential fit of temperature change. There is a wide range of estimates for heat capacity, ranging from below 2 W/m^2 to over 6 W/m^2.

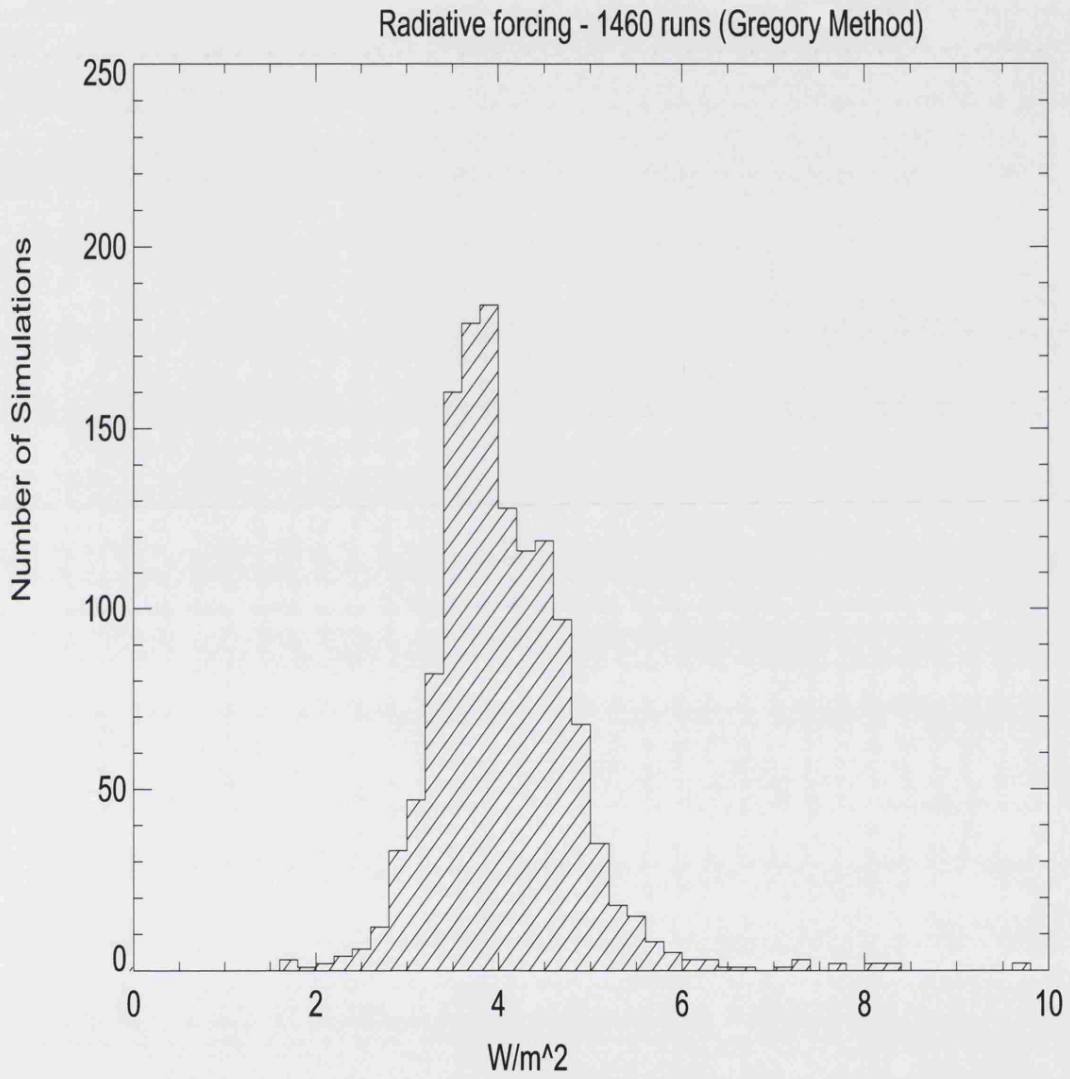


Figure 4.5: The top of atmosphere radiative flux imbalance in the doubled CO_2 phase is plotted over 1460 quality controlled simulations. The method used to estimate this flux imbalance is the Gregory plot method. There is a wide range of estimates for heat capacity, ranging from below 2 W/m^2 to over 6 W/m^2

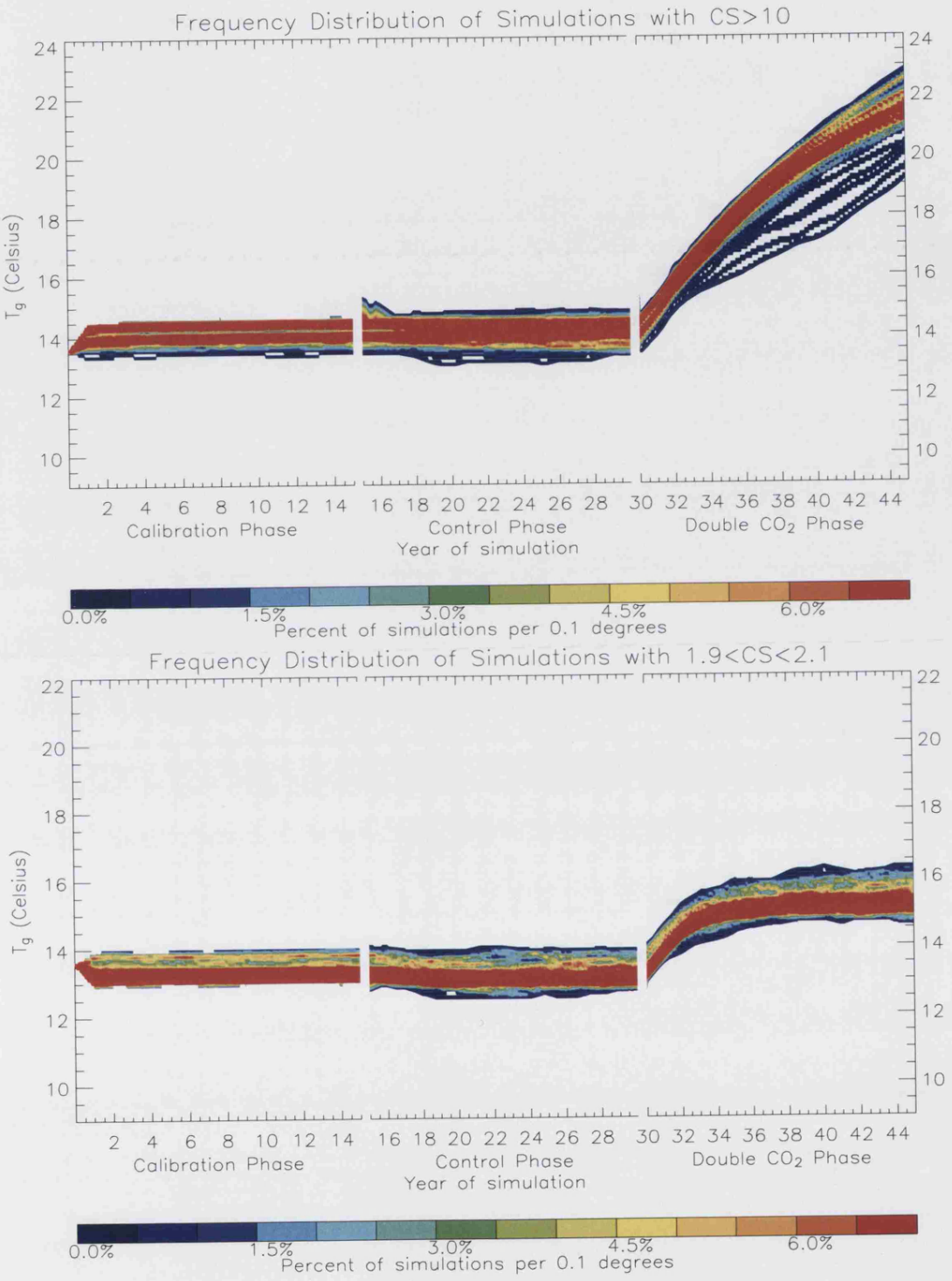


Figure 4.6: The GMST time series over the 3 experimental phases of the CPDN experiment for 350 simulations with estimated CS over 10 degrees Celsius (top). Also shown is the time series for 822 simulations with 2 degrees CS (bottom). Whereas the 2 degree simulations seem to have reached an equilibrium by the end of phase 3, for the simulations with a CS greater than 10 degrees the simulated warming is so extreme that 15 years is not enough time for an equilibrium to be reached.

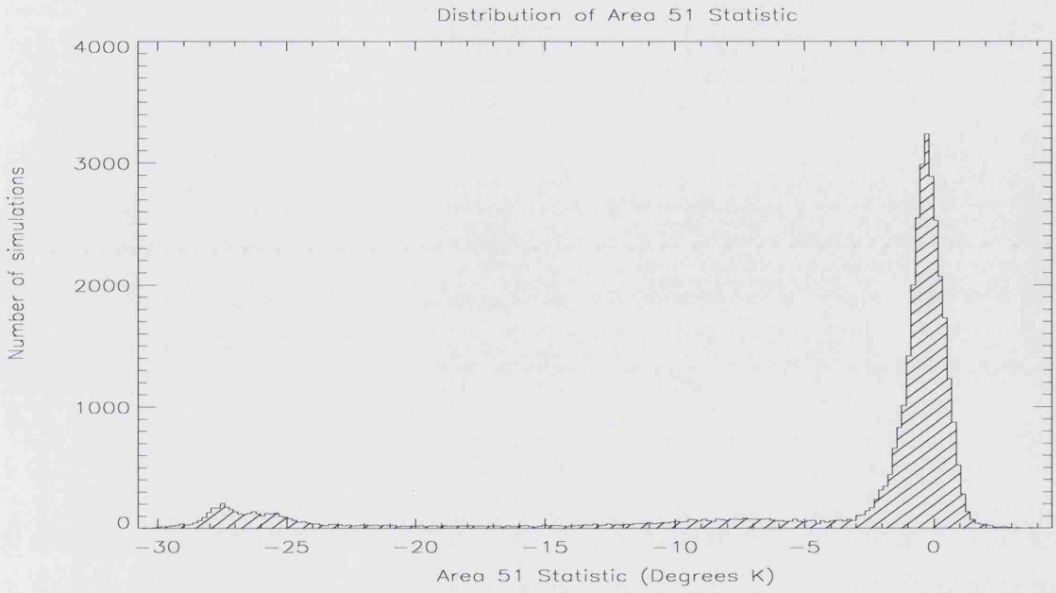


Figure 4.7: The distribution of values of the Area 51 anomaly for 45644 simulations before applying any quality control. The distribution is clearly bi–modal, representing simulations that do not exhibit a negative feedback (peak around 0), those that have (peak around -27) and a smaller number of intermediate simulations that are drifting (between -5 and -20).

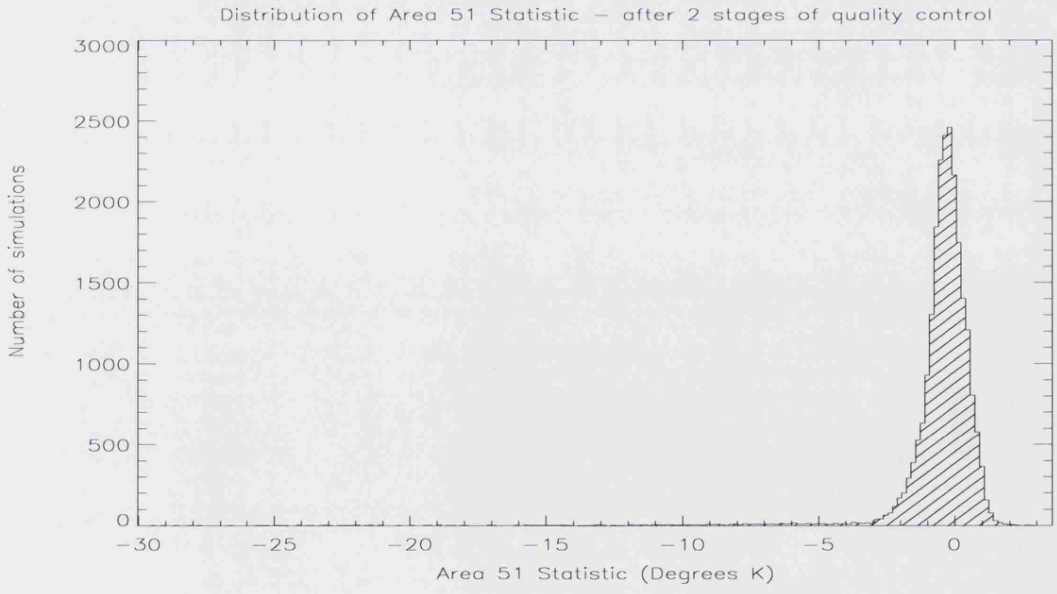


Figure 4.8: The distribution of values of the Area 51 anomaly for 23050 complete simulations with a non–significant GMST drift. The distribution is uni–modal about 0, with a long tail in the negative values.

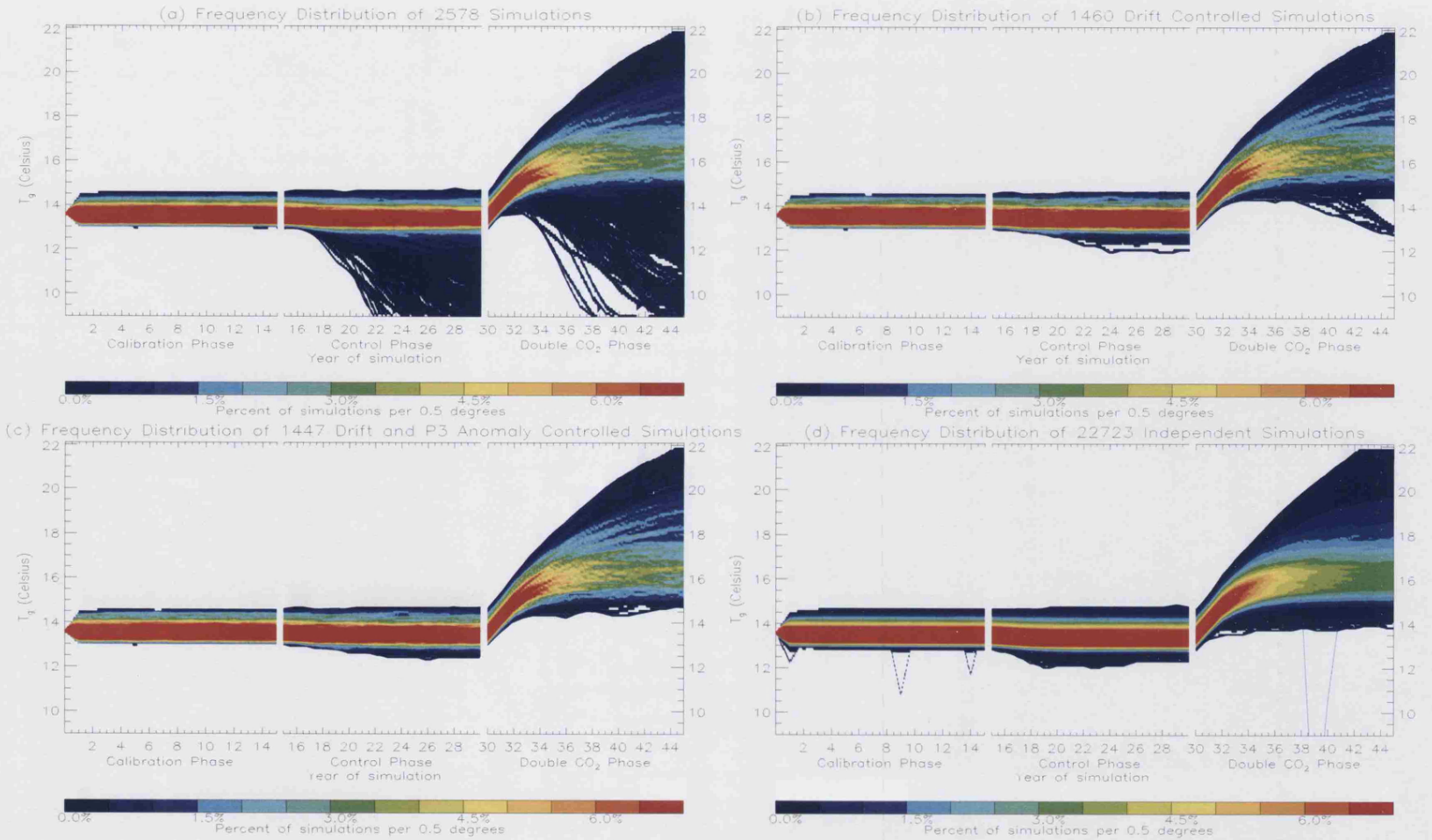


Figure 4.9: Panel (a) shows the time series of 2578 simulations with no quality control applied. Panel (b) shows the same time series, applying only the first two stages of quality control, leaving 1460 simulations. Panel (c) shows the time series with full quality control applied, leaving 1447 simulations. Panel (d) shows the time series of 22723 simulations after full quality control was applied to 45644 simulations.

Chapter 5

Investigating variations in heat flux adjustment in the CPDN ensemble

5.1 Introduction

Heat flux adjustments (HFA) are often used to account for the flow of energy between the atmosphere and ocean in climate models that do not achieve realistic balance naturally Collins *et al.* (2006); Knutson (2008); Murphy (1995); Murphy *et al.* (2004); Sausen *et al.* (1988); Shackley *et al.* (1999); Stainforth *et al.* (2005); Williams (1999). Flux adjustments play a number of different roles in climate modelling. These roles are discussed in this Chapter and relationships between flux adjustments and the HadSM3 model's dynamics are investigated. Details of the HFA used in the CPDN experiment are documented in this Chapter for the first time and it is shown that the HFA used in the HadSM3 model contains information on CS – the equilibrium GMST response to a doubling of CO_2.

In this Thesis, HFA is defined to be the anomalous amount of heat transferred from the ocean to the atmosphere such that the SSTs match the 1961-1990 climato-

logical observations presented in New *et al.* (1999). This is not to be confused with sensible or latent heat fluxes (sensible heat flux relates to the energy flux in heating a surface without evaporation and the latent heat flux the flux of heat associated with the phase change of liquid through evaporation) that arise through physical processes in the model.

Investigating the potential relationships between the HFA and certain aspects of model behaviour (such as CS) is important for understanding the range of simulated behaviour for the purposes of decision support, in particular if the HFA produces a detectable bias. The HFA field may hold important clues for improvement of the model and for the design of future experiments, as well as the analysis of the current data set.

It is valuable to look at the dynamics of the HFA itself and its effect on the HadSM3 model's response to rising CO_2 concentrations in order to better design and interpret flux adjusted modelling experiments. The use of HFA has been justified in a number of ways: to maintain a realistic climate during the spin–up process Johns *et al.* (1997), to prevent significant model drifts in the control climate Murphy (1995); Santer *et al.* (1994) and to represent missing ocean heat transports that are not reproduced by a slab ocean Sanderson *et al.* (2007). It has also been noted that models not using flux adjustments produce SSTs biases that might affect the reliability of regional simulations Collins *et al.* (2006). Furthermore, the use of a slab ocean with flux adjustments can enable a GCM to be run faster that would be possible with a dynamic ocean. For climate models with slab oceans HFA is used to simulate SSTs and produce stable model simulations. In the case of the CPDN experiment, HadSM3 is quicker that its fully coupled counterpart, HadCM3, allowing more simulations to be run in a given amount of time. For a detailed discussion of the use of flux adjustments and some scientists' views on them see Shackley *et al.* (1999).

In the recent IPCC Assessment Reports Houghton *et al.* (2001); Solomon *et al.*

(2007a) there has been a decrease in the use of flux adjustments due to their apparently unphysical nature. Criticisms of the artificial nature of flux adjusted models are relevant since flux adjustments may not be able to mimic important feedbacks in response to rising GHGs. A solution recently adopted Houghton *et al.* (2001); Solomon *et al.* (2007a) of subtracting the control climate from the transient simulation (giving the "model anomaly") in an attempt to eliminate model drift has not yet been shown to be better than applying flux adjustments in the case where unphysical model drift occurs in the control phase. It remains an open question how best to design climate modelling experiments to deal with systematic model biases.

This Chapter examines the HFA used in the CPDN experiment and is structured as follows. Section 5.2 looks at the variation in the HFA field across ICs and different sets of parameter values. The variability due to IC perturbation is shown to be small in comparison to the variability present across parameter sets. Parametric perturbation is shown to have a significant effect on HFA fields. The question of HFA stabilisation in the calibration phase is discussed. If the HFA has not stabilised by the end of the calibration phase, this would mean that simulations within an ICE would contain a systematic source of variability. It might not then be valid to consider the individual simulations of an ICE as drawing from the same distribution if the HFA is introducing a further, systematic source of variability. In Section 5.3, the seasonal variability of the HFA on global and regional scales is examined. In Section 5.4 a relationship is shown to exist between the HFA field and CS. Potential causes and consequences of this relationship are discussed. Section 5.5 looks at the HFA field as a possible cause of climate drift or a trigger for an unphysical negative feedback. No evidence is found that HFA is a significant cause of model drift. Conclusions and original work in this Thesis are given in Section 5.6. Each Section considers an aspect of the HFA and its relevance to the design and interpretation of flux adjusted climate modelling experiments. The remainder of the introduction

explains the use of HFA in the CPDN experiment.

5.1.1 HFA in the CPDN experiment

The HadSM3 model used in the CPDN experiment requires HFA Williams (1999). The CPDN grand ensemble includes a large number of different *model versions*, each employing a different set of parameter values. Since different parameter values can result in significantly different dynamics, the HFA is calibrated for each model simulation. Arguably, one should either calibrate the HFA once for each model version, thus saving computational resources or to average the HFA over ICE members to obtain a better estimate for the HFA. In either case it is recommended that, in future, the same HFA should be used for all simulations within an ICE. There are at least two challenges to achieving this in the CPDN experiment;

1. Distributed computing does not easily enable users to obtain data from other simulations. This would require additional coding and the downloading of data that would likely reduce the speed at which simulations are completed.

2. It is useful to have ICE data from the calibration phase to use as a comparison to subsequent experimental phases and to analyse the difference between simulations sharing a single set of dynamics. Nevertheless, this data could be stored and a single HFA field applied for each member of an ICE.

During the first phase (the calibration phase) of the CPDN experiment the required HFA is calculated such that the model's ocean matches 1961–1990 observed SSTs New *et al.* (1999). The HFA is calculated as the addition or subtraction of heat required for HadSM3 to produce observed SSTs. Henceforth, positive values of the HFA denote a flux of heat into the ocean and negative values out of the ocean. Having then defined the HFA (the last 8–year mean of the 15 year phase is taken for each month), the same field of ocean HFA is then used each year in the control phase and the doubled CO_2 phase Piani *et al.* (2005).

The second (control) phase is run using pre–industrial concentrations of CO_2, as in the calibration phase. The only difference between the calibration and control phase is that the HFA is held constant from year to year in the control phase, thus there is no requirement for the model to re–produce observed SSTs. Models which display any significant and persistent drift during the control phase are disregarded as unphysical, as described in 4.4.2. Possible relationships between the HFA and model drift are looked at in Section 5.5.

The same HFA is applied in the doubled CO_2 phase as in the control phase. The HFA varies spatially and seasonally but is fixed annually and is not altered during the control or doubled CO_2 phases. There is no HFA over land.

The HFA fields over the 15 year calibration phase are examined in Section 5.2 in order to better understand the effect of parameter perturbation on the model versions' dynamics. Analysis of the HFA can help to understand the results of slab–model experiments and might help improve the design of future flux adjusted modelling experiments. The convergence of HFA fields over a 15 year simulation in the CPDN experiment is looked at in Section 5.2 in order to better understand the effect of parametric perturbation.

5.1.2 HFA Data Sets

Four different subsets of CPDN data are analysed in this Chapter, both produced by the CPDN experiment. The data sets used are:

1. A 48 member ICE of the standard HadSM3 model, taken from a grand ensemble of 45644 simulations, is used to look at ICE variability of the HFA more closely. Note that the the number of simulations available from the standard HadSM3 model is different to the 64 Standard HadSM3 simulations referred to in Chapter 4. This is because a full complement of data on HFA was available for only 48 of the 64 simulations. This data set will be referred to as the Standard ICE in the remainder of this Chapter.

2. A 6 member ICE of the standard HadSM3 model, taken form the grand ensemble of 2578 simulations presented in Stainforth *et al.* (2005). This set will be referred to as the Initial ICE henceforth and is used to assess the internal variability of the Standard HadSM3 model. The Initial ICE is used in preference to the Standard ICE in one case only for ease of presentation.

3. A grand ensemble of 484 model versions with between one and nine distinct[1] ICs simulation under each, giving 2578 simulations in total. This data will be referred to as PPE_{2578} in the remainder of this Chapter. This grand ensemble was used instead of the lager grand CPDN grand ensemble of 45644 simulations to investigate relationships between the HFA and model behaviour, as were the methods of quality control presented in Chapter 4. This was done so that any results found in the analysis of the HFA could be used as a part of quality control. For example, this could be the case if the HFA can be attributed as a key contributor of model drift, an inconsistency that is looked for in the process of quality control. Since the results do not suggest that the HFA can be linked to any important internal inconsistencies in simulations, HFA is not used in quality control. Despite this, the relationship between HFA and CS is examined as a method for constraining simulations, explored in Section 5.4 and Chapter 7.

4. After applying quality control, PPE_{2578} contains 1460 simulations. This new set is referred to as $PPE_{quality}$

The Standard ICE, PPE_{2578} and $PPE_{quality}$ data sets are used as appropriate in the analysis, the Standard ICE being used to investigate the internal variability of the standard HadSM3 model and PPE_{2578} and $PPE_{quality}$ are used to analyse the HFA across parameter perturbation.

[1]Some duplicate simulations, with identical ICs were also sent out in order to verify the experimental design. These duplicate simulations are not considered here as distinct simulations unless they produce different output.

5.1.3 Examples of HFA fields

In order to motivate ideas and to better understand the HFA used in the HadSM3 model, Figure 5.1 shows three HFA fields randomly selected from $PPE_{quality}$. There are regional patterns such as a blue band (representing heat being removed from of the oceans) about the equator and more heat being put in around Japan and the East Coast of the USA. These spatial patterns are discussed in more detail in Section 5.2. Figure 5.1 shows that these randomly selected HFA fields show similar regional characteristics although the magnitude of the HFA varies. Where the HFA is close to zero, this indicates that the model requires no significant flux of heat to produce observed SSTs. Where the HFA if far from zero, the model requires a large flux of heat to match observed SSTs. Different models require different adjustments according to various missing ocean dynamics and model biases. Areas showing a significant flux of heat, e.g. around the equator, indicate the atmospheric and slab ocean components of the HadSM3 model are not in balance.

The variability seen in the $PPE_{quality}$ is now analysed in Section 5.2.

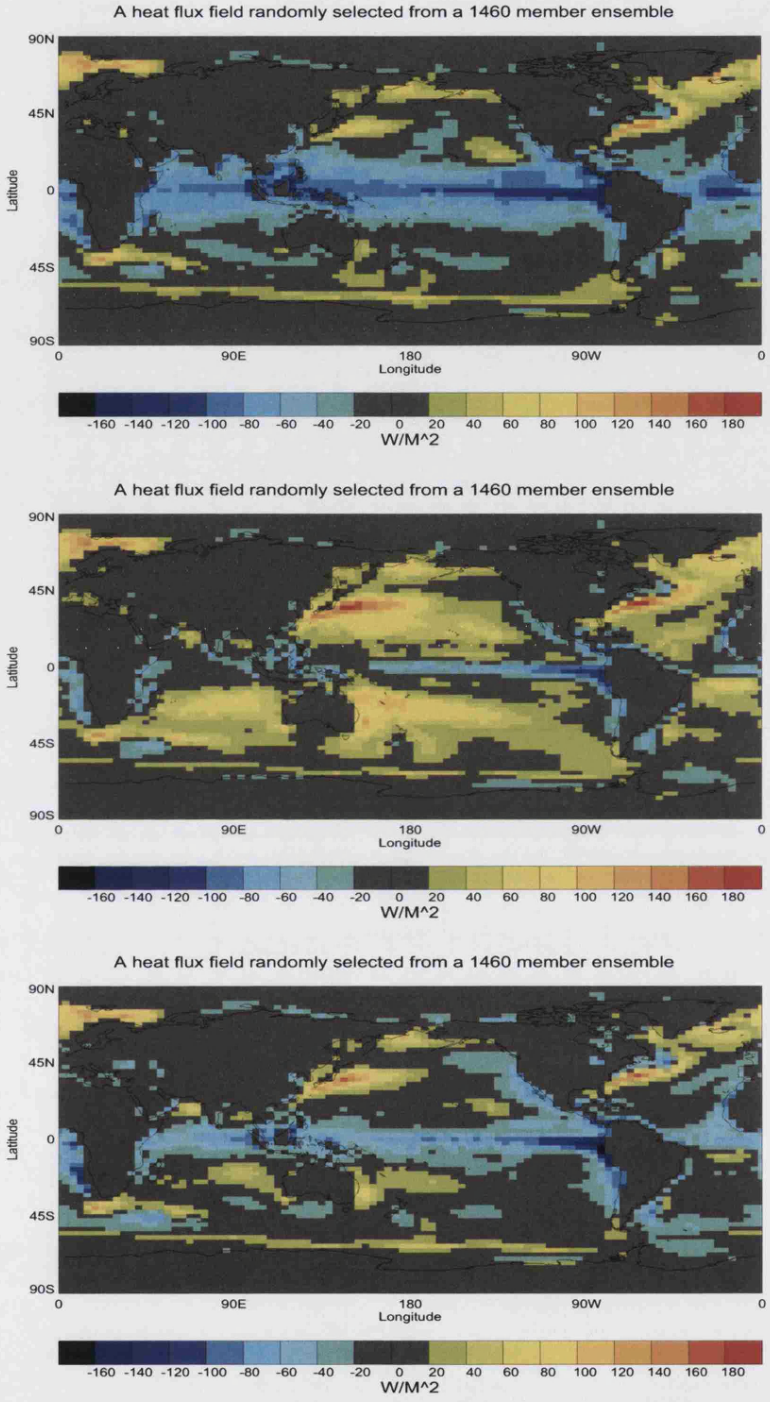


Figure 5.1: Three randomly selected HFA fields from $PPE_{quality}$. There are significant differences in HFA by region. Some areas require a reduction of heat in the ocean by more than $150W/m^2$ whereas others require more than $200W/m^2$ to be added. A HFA of $200W/m^2$ is approximately the same effect as an increasing the solar constant by 50%.

5.2 HFA variability

This Section looks at variability in the HFA within the Standard ICE and $PPE_{quality}$. Studying the variability of the HFA can help inform the interpretation of flux adjusted experiments and aid future experimental design. Variability is looked at within the Standard ICE from the standard HadSM3 model initially, then across different model versions.

Large variations are shown across $PPE_{quality}$ in Section 5.2.1. Section 5.2.2 shows there are spatial variations in the HFA within an ICE of the standard HadSM3 model, indicating that some variability is due to IC perturbation and not parameter perturbation. It is important to consider whether perturbing model parameters has a significant effect on the HFA. This is looked at in Section 5.2.3. The stabilisation of the HFA fields is discussed in Section 5.2.4 with reference to the global mean HFA time series over the calibration phase. Stabilisation of the HFA is defined here as the limiting state of the HFA to within ICE variability. After the HFA has converged, further calibration will have no significant effect.

Stabilisation is analysed in this Section on regional scales, although no regional HFA times series were available in this data set. The variability across 8 year mean fields is used to look for evidence that suggests the HFA has not converged. Since ICE members should produce similar HFA fields if the calibration phase is run to equilibrium, the relative variability in HFA fields within and across model versions can be used to test the hypothesis of stabilisation. This test is carried out in Section 5.2.3 and it is shown that perturbing parameters has a greater impact on the HFA than perturbing ICs.

It is shown that, whilst the stabilisation of the HFA can not be established conclusively with the available data, the model output is consistent with the hypothesis that the HFA has reached an approximate stabilisation.

5.2.1 The HFA bounding box

Spatial variability in the HFA field can be expressed using a bounding box Judd *et al.* (2007); Weisheimer *et al.* (2004). A bounding box comprises two extremal HFA fields (the maximum and the minimum) where each grid–box in the bounding box is defined by the most extreme member of the ensemble. A very wide range in the bounding box indicates a wide set of behaviour across simulations whereas a small bounding box shows that all members produce similar HFA fields (see Judd *et al.* (2007); Weisheimer *et al.* (2004) for a discussion on the use of bounding boxes as a means of ensemble evaluation). Figure 5.2 shows the maximum, minimum and range (maximum minus minimum) of $PPE_{quality}$. The minimum and maximum fields shown in panels (a) and (b) respectively show that, in some areas, HFAs of over $260W/m^2$ and below -$120W/m^2$ are used. These are significant fluxes of heat, of the same order of magnitude as the total radiation from the sun ($\sim 340W/m^2$). The physical validity of using HFA in cases where such large corrections are required is dubious. In order for models with large values of HFA to be considered physically relevant, it must be shown that the HFA does not introduce any biases that significantly affect model results. It is an open question whether it is better to apply large artificial adjustments to correct model error or to use an un–flux adjusted model that has significant systematic errors Collins *et al.* (2006).

In the bottom panel of Figure 5.2, showing the range of the bounding box, gives an idea of the spread of HFA in $PPE_{quality}$. Dark blue areas represent a tight ensemble whereas a yellow, or red, grid–box implies more variation between ensemble members. This gives an idea of how the HFA adjustment varies by region. There is relatively little variability in the high latitudes, and a tongue–like area off the coast of Peru. This East Pacific region is looked at in more detail in Section 5.5 when possible relationships between the HFA and model drift are examined. The largest range occurs in the West Pacific, just North and South of the equator. In some of these areas, there can be a difference of over $200W/m^2$ between model simula-

tions. This shows that the HFA can be very different between model simulations on regional scales.

5.2.2 Variability with Initial Condition

Attention is now restricted to the Standard ICE. The Standard ICE can be used to assess stabilisation of the HFA in the calibration phase through a quantification of a model versions' internal variability.

For the purposes of experimental design it is important to know whether the HFA field has stabilised by the end of the 15 year calibration phase. Stabilisation is defined here as the HFA reaching a equilibrium to within the range of ICE variability. An early stabilisation could allow for a shorter calibration phase, whereas a lack of stabilisation by the end of the phase might suggests a longer period of calibration is required. The range of values between members of the Standard ICE gives an idea of the internal variability of the model, and hence a means of assessing stabilisation. It is only possible to assess with confidence whether the HFA has stabilised to an equilibrium value by the end of the calibration phase using a time series. A time series is necessary to assess whether simulations are stabilising in time towards a common distribution. Without a time series, the question of stabilisation can not be answered conclusively but it is possible to look for information in the data available that is consistent with this hypothesis. For the HFA, only global means were available as a time series. The available spatial fields represent the 8 year monthly mean of years 8–15 of the calibration phase. The analysis carried out in this Chapter is forced to work within this weakness of the experimental design.

An ICE, as a collection of spatial fields, can help to assess the internal variability present in the HFA and so place a lower limit on the degree of stabilisation attained by the end of the phase. If the HFA has converged, it would be expected that the difference between IC members would be small compared to the effect of perturbing parameter values. This is shown to be the case in Section 5.2.3. Furthermore, it

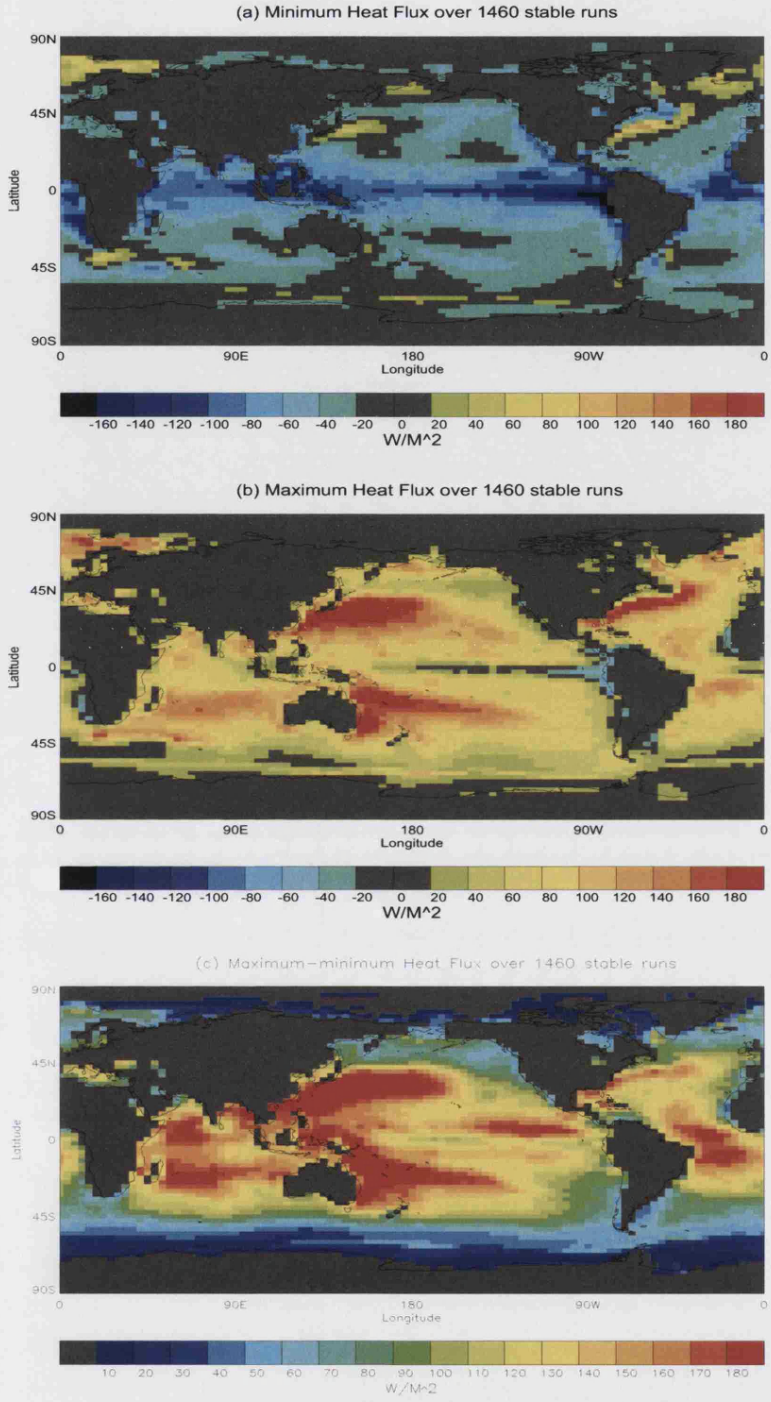


Figure 5.2: The three panels show the (a) minimum, (b) maximum and (c) range of the HFA field for $PPE_{quality}$. The bounding box relates to the spread of the ensemble at each grid–box. Positive values denotes heat intro the ocean. The range of values shown in panel (c) can be as large as $200W/m^2$.

might be expected that when looking at the 8 year mean spatial field across an ICE that each members' anomalies (from the ICE mean) would differ in a random fashion, without any systematic biases. The presence of systematic differences between ICE members would suggest that ICE members are not being drawn from the same distribution, indicating a lack of stabilisation.

If the fluctuation between IC members were random and not due to any systematic differences, the bounding box would be defined at a roughly equal number of grid–boxes by each member and in a fairly random fashion, with a certain amount of spatial correlation relating to the length scales on which the HFA operates. If the HFA is mimicking ocean dynamics, these length scales might be expected to be typically greater than a grid–box. This spatial correlation can be seen in Figure 5.3.

Figure 5.3 shows the grid–boxes at which each member defines the top or bottom of the bounding box (equivalent to the maximum and minimum of the ensemble, respectively) for a six member standard HadSM3 ensemble (the 6 members of the standard HadSM3 ICE available from the 2578 ensemble). The Initial ICE is used here instead of the Standard ICE for presentation purposes, to avoid over–crowding the Figure. The "Technicolour raincoat" effect shows that members do tend to define small spatially–correlated regions but that their distribution about the globe appears random. The same member rarely defines both the top and the bottom of the bounding box in Figure 5.3. The bounding box is not defined by one or two members but all members contribute at different locations in a random fashion. This lends support to the hypothesis that members of the Standard ICE are being drawn from the same underlying distribution.

The bounding box of the Standard ICE is used in order to understand the magnitude of HadSM3's internal variability. Figure 5.4 shows the bounding box for the Standard ICE. The minimum, maximum and the range are shown. Over most of the planet there is little variability between members, in particular around the equator,

as shown in panel (c). Some areas have much more variability between members. Around the East coast of North America there is variation of over $30W/m^2$, perhaps due to the absence of an important ocean process, such as the Western Boundary Current, in the slab model. Large variations in the HFA indicates that the same model version (with different ICs) requires differing amounts of heat for the model to produce the observed SSTs. This internal variability could be a result of a lack of stabilisation in the HFA or an irreducible feature of variability in the HFA. The internal variability of the model is useful in assessing the effect of parameter perturbation. The difference in inter–model and intra–model HFA variability in the calibration phase is examined in Section 5.2.2 where it is shown that the inter–model differences dominate the variability within ICEs.

Spatial Correlation

The spatial correlation of the Standard ICE is examined in this Section. Spatial correlation is defined here as the autocorrelation between pairs of data points a specified distance apart on a spatial grid. The spatial correlation of ICE members can help understand the degree of internal variability present in the HFA and give insight to the way the HFA mimics missing ocean processes. Each member of the Standard ICE is considered as an anomaly from the ICE mean. It would be expected that if all ICE members are being drawn at random from the same distribution, that the anomalies would not show any systematic differences. In order to motivate ideas the spatial correlation present in the HFA can be looked at in a sub–set of simulations; Figure 5.5 shows examples of the HFA for 8 simulations (the ICE mean has been subtracted at each grid–box). Simulations were selected at random from the 48 available in the Standard ICE. There is visible spatial correlation within the anomalies of local HFA, suggesting that the noise is not independent on a grid–box level. This spatial correlation is quantified in the remainder of this Section.
Figure 5.6 shows the HFA anomalies for a different set of 8 simulations. Instead

of plotting the anomalies themselves, each anomaly is assigned a rank (1 being the lowest anomaly, 48 being the highest anomaly). There is spatial structure in the rank ordered anomalies. Especially in the rank order anomaly fields, there seems to be small scale positive correlation and larger scale negative correlation (whilst there are patches of red and blue, red areas often sit next to blue areas). These results show that the ocean processes mimicked by the HFA operate on length scales larger than an individual grid–box. The presence of spatial correlation across the anomaly and rank–ordered anomaly fields indicates that this result is believed to be robust.

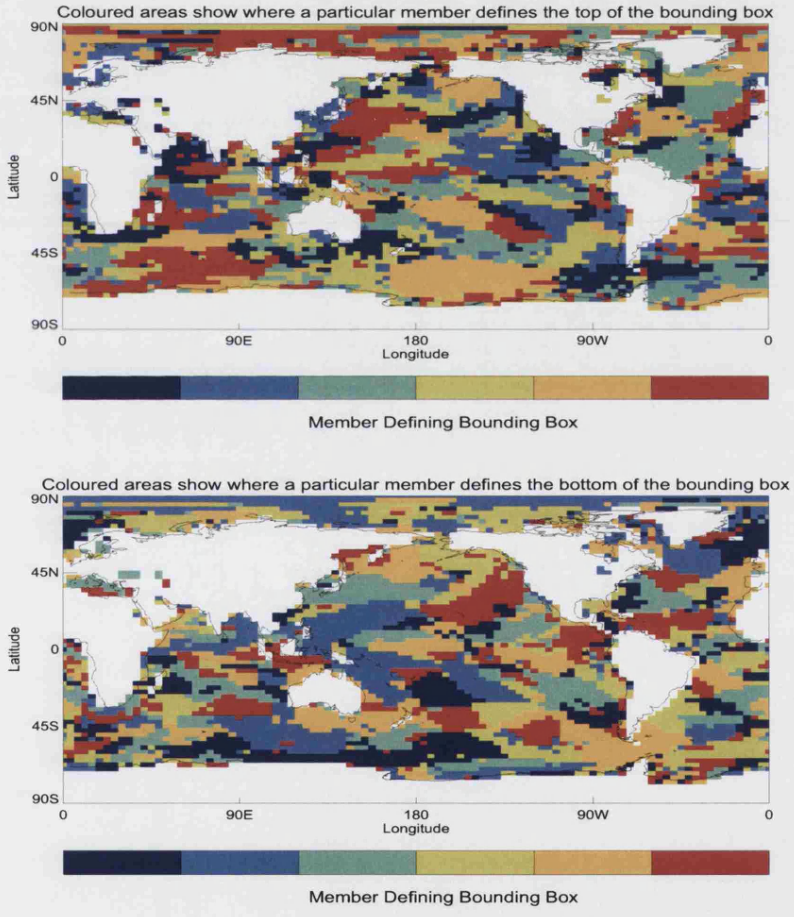


Figure 5.3: A colour is plotted for each of 6 members of the Initial ICE where it defines the bounding box. The top picture shows the top of the bounding box, and the bottom the bottom of the bounding box. The roughly even distribution of colours indicates that all members contribute to defining the bounding box and the patches of colour that there is some spatial correlation in the HFA.

A more formal test of spatial correlation is now carried out, as indicated above.

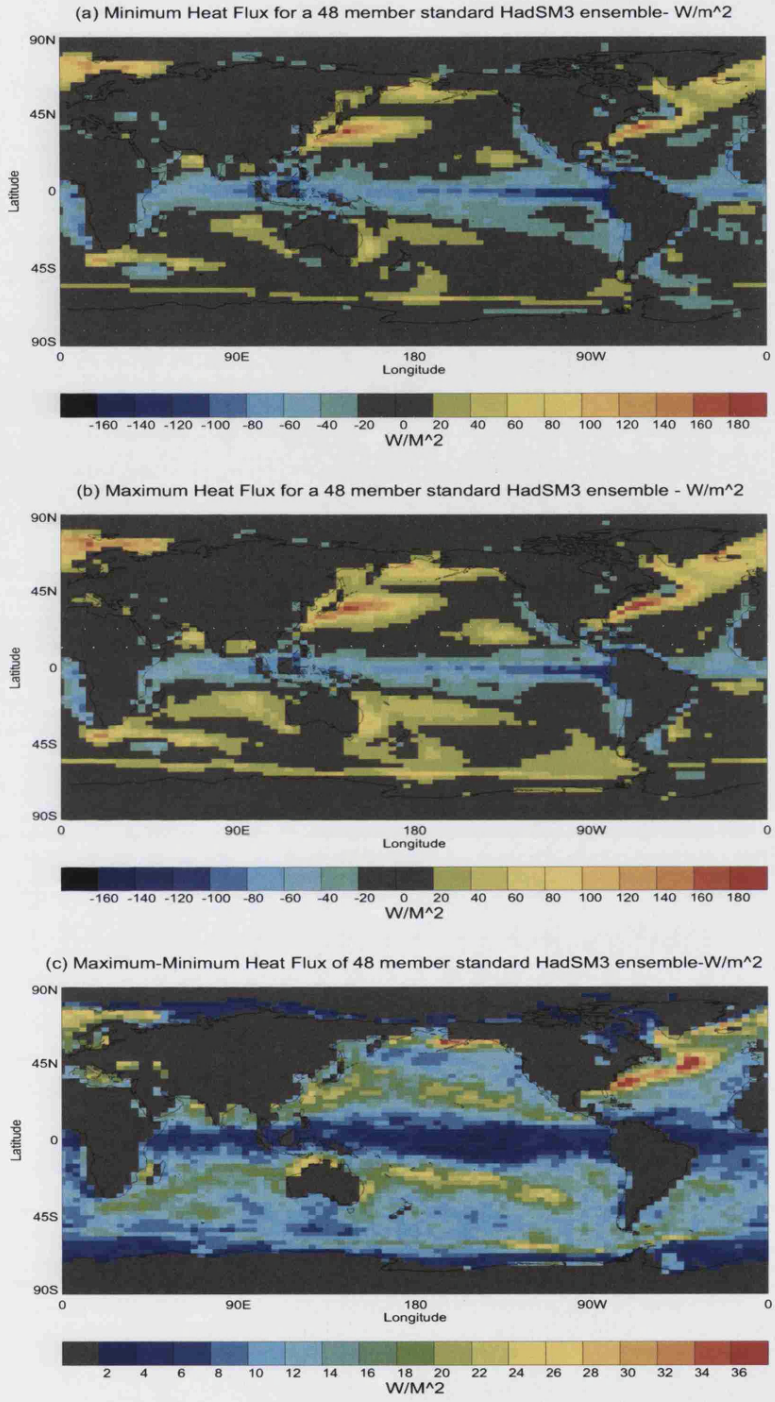


Figure 5.4: Panels show (a) minimum, (b) maximum and (c) range of HFA for the Standard ICE. Panel (c) shows that there are regions for which members of the Standard ICE require HFAs differing by less than $4W/m^2$ but other areas where the differences can be as large as $40W/m^2$

In order to quantify the extent of spatial correlation in the HFA fields, an adapted version of Moran's I statistic Moran (1950) is used to assess the length scales on which spatial correlation is present. This statistic measures the spatial correlation between a grid–box and its neighbours at a distance d in the horizontal and vertical directions only (thus four points are considered for each grid–box). By calculating Moran's I statistic for various length scales, it is possible to assess the scales on which spatial correlation is typically present. The standard Moran's I statistic is adapted in this Section in order to suit the purpose at hand. Firstly, since HFA can be considered as "missing" over land, correlation is not calculated for or with land grid boxes. Only spatial correlation between ocean grid boxes is considered here (spatial correlation is evaluated even where ocean grid boxes are separated by grid boxes containing land). Secondly, due to heteroscedasticity in the data, standardised HFA anomalies are used; each the Standard ICE mean is subtracted at each grid–box from all simulations and is scaled by the ICE standard deviation at each grid–box. Morgan's I statistic is defined as:

$$I = \frac{\sum_{i=1}^{m}\sum_{j}^{n-d} w_{i,j}z_{i,j}z_{i,j+d} + \sum_{i=1}^{m-d}\sum_{j}^{n} w_{i,j}z_{i,j}z_{i+d,j}}{\sum_{i,j} w_{i,j}z_{i,j}^2} \tag{5.1}$$

where i, j represents the grid–box i, j for dimension size n (number of longitude points) by m (number of latitude points), where n equals 96 and m equals 73. The dummy variable $w_{i,j}$ equals 0 for land grid–boxes and 1 for ocean grid–boxes and $z_{i,j}$ equals the HFA at grid–box i, j minus the global mean of the anomalies. d is the distance between grid–boxes for which spatial correlation is calculated. The statistic I represents an estimate of the spatial correlation between ocean grid–boxes. I is calculated over the Standard ICE for distances (values of d) ranging from 1 to 20. The values of I for each ensemble member is plotted against distance in Figure 5.7. In order to assess whether there is significant spatial correlation in the anomalies, the values of I are also calculated for an array of independent, standard Normal

variables where it is expected that there will be no significant spatial correlation present. Using this independent data set as a guideline for the presence of spatial correlation, it is possible to judge the average distance at which ICE members' anomalies are correlated from Figure 5.7. The values of I for the standard Normal data are used to define zero correlation for all distances, indicating no significant spatial correlation – these values are centred about 0 with a range of (-3.55, 3.40). Spatial correlation in the Standard ICE is defined to be significant where at least 90% (i.e. 39 of the 48) of simulations exceed the maximum value of the Normal data set. The HFA anomalies produce value of I are significantly greater than 0 for distances up to 7 grid–boxes. Above distances of 7, there seems to be no significant spatial correlation. This demonstrates that the HFA fields are typically spatially correlated on scales up to about 7 grid–boxes.

There are at least two possible reasons for the presence of spatial correlation in the HFA field:

1. The missing ocean processes the HFA is accounting for operate on length scales greater than an individual grid–box.

2. The HFA has not converged by the end of the calibration phase. Despite close agreement between the 48 members of the Standard ICE on the global mean level, it is possible that there are still regional differences that are yet to converge.

It is likely that **1)** is true on physical grounds since the ocean dynamics, missing in HadSM3, are expected to operate on length scales greater than a grid–box. The question of stabilisation, raised in **2)** is looked at further in Section 5.2.3 and 5.2.4 where stabilisation of HFA within each model version is analysed using $PPE_{quality}$.

5.2.3 Perturbed Physics Ensembles

It has been demonstrated that there are local variations of up to $30W/m^2$ resulting from IC perturbation and that there are spatial correlations within the HFA fields

of the standard HadSM3 model. In order to understand the significance of these variations and the effect of perturbing parameters on the HFA, attention is now turned to the $PPE_{quality}$ data set. Two questions are looked at here:

1. Does the perturbation of parameters lead to significant differences in HFA? If this question can be answered in the negative, it may not be necessary to run a calibration phase for each simulation, saving up to a third of the computational resources. It is shown here that this can *not* be assumed and that parameter perturbation does have a significant effect on the HFA.

2. Has the HFA converged by the end of the calibration phase? This question can only be fully answered using a spatial time series. Since a time series is only available for global mean HFA, it is only possible to assess whether stabilisation has occurred on a global level. Despite the lack of spatial time series, it is possible to use the 8 year mean HFA field to look for evidence against the hypothesis of stabilisation. Since model versions can have different dynamics and IC members can not, if the HFA had not converged at all, it would be expected that the HFA field would not be distinguishable between model versions. It is shown that the HFA from different model versions is significantly different on a spatial scale, indicating that some level of stabilisation has been reached by the final 8 years of the calibration phase.

If parameter perturbation has a significant effect on the equilibrium HFA, then it might be expected that the HFA from one ICE should be distinguishable from another. This is shown to be the case in this Section.
If the HFA has stabilised by the end of the calibration phase, ICE members that share the same parameter values should be similar on both global and regional scales (in particular, be more similar than randomly selected simulations). This similarity is based on the assumption that ICE members of the same model version will display the same dynamics and provide a lower limit on the variability that can

be seen across a PPE. Figure 5.8 shows the range of global mean HFA in panel (a) and CS in panel (b) for each set of parameter values in $PPE_{quality}$. Whilst there can be quite a large range across ICs for CS (panel (b)), the global mean HFA is extremely close, so much so that the bars (showing the minimum and maximum within that ensemble) often appear as one in panel (a). The maximum difference in global mean HFA within a model version is $0.419W/m^2$. This suggests that the HFA fields within each ICE might have stabilised by the end of the calibration phase on a global scale.

It is important to check whether this similarity is present on regional length scales as well as globally. This requires an analysis of how much regional variability in the HFA field is present amongst members with the same parameter values. This can be tested against the hypothesis that the HFA has not stabilised by the end of the calibration phase and the same level of similarity is present for HFA fields randomly selected across different model versions. The spatial patterns within the HFA fields of $PPE_{quality}$ are analysed using Singular Value Decomposition (SVD) Press (1992). The leading pattern of the HFA field within ICEs is compared to the leading pattern of randomly selected groups of simulations. This comparison allows the hypothesis to be tested that simulations sharing parameter values share similar HFA fields. An experiment is designed to test this as follows:

For each model version, the last 8 years of the control phase, annual mean, HFA fields for quality–controlled simulations were used to form a 96×73 matrix (the HadSM3 model used in this experiment is run on a 3.75 degree longitude by 2.5 degree latitude resolution, giving a 96×73 grid), equivalent to a 7008 vector in this case. Data used are the average over the last 8 years of the calibration phase. The matrix of HFAs for each model version is then an n×7008 matrix, where n is the number of IC members available for that model version.

The global mean HFA is subtracted from each simulation. SVD was then used to form n new, orthogonal, directions and the associated singular values. The

leading Singular Value is denoted LSV. This process is carried out for 440 model versions that have more than 2 quality controlled members (duplicate simulations are not included in the analysis). This gives 440 LSVs for the grand ensemble. For ensembles with different numbers of IC members, the LSV might be different and this needs to be reflected when testing the distribution of the 440 FSVs against randomly selected simulations. As each model version is analysed, the same number of HFA fields are randomly selected and an SVD applied to the matrix formed by their HFA fields. This process gives two sets of 440 LSVs that can be tested for difference.

A statistical test is carried out to compare these distributions since the distribution of LSVs is unknown (and may be difficult to estimate reliably) a non-parametric test is appropriate. Where the ensembles come from the same model version, the mean LSV obtained was 0.913 (the full range is from 0.707 to 0.971) compared with 0.638 (full range 0.439 to 0.957) for the randomly selected fields. For comparison, the 6 member unperturbed ensemble gave an LSV of 0.865. It was judged that the test of difference could be restricted to the LSV alone, since the LSV typically accounts for over 90% of variation of HFA within ICEs. A Non-Parametric Rank Sum test for equality of the means within model versions and across model versions (a Mann-Whitney test Mann & Whitney (1947) was used here) yields a p–value of less than 0.00001. Overall, there is strong evidence that the HFA fields are more similar where they share the same parameter values although some similarity exists between randomly selected HFA fields (the dot product of LSVs is always at least 0.439). Perturbing model parameters leads to a non–trivial change in the required HFA on both regional and global scales. This provides evidence that parametric perturbation significantly changes the pattern of HFA, which therefore must be calibrated for each set of parameter values. Furthermore, the result that parametric perturbation has more effect on the HFA than IC perturbation, spatially as well as globally is consistent with the hypothesis of stabilisation by the last eight years of

the calibration phase. Based on these results, it seems it is necessary to calibrate the HFA for each model version.

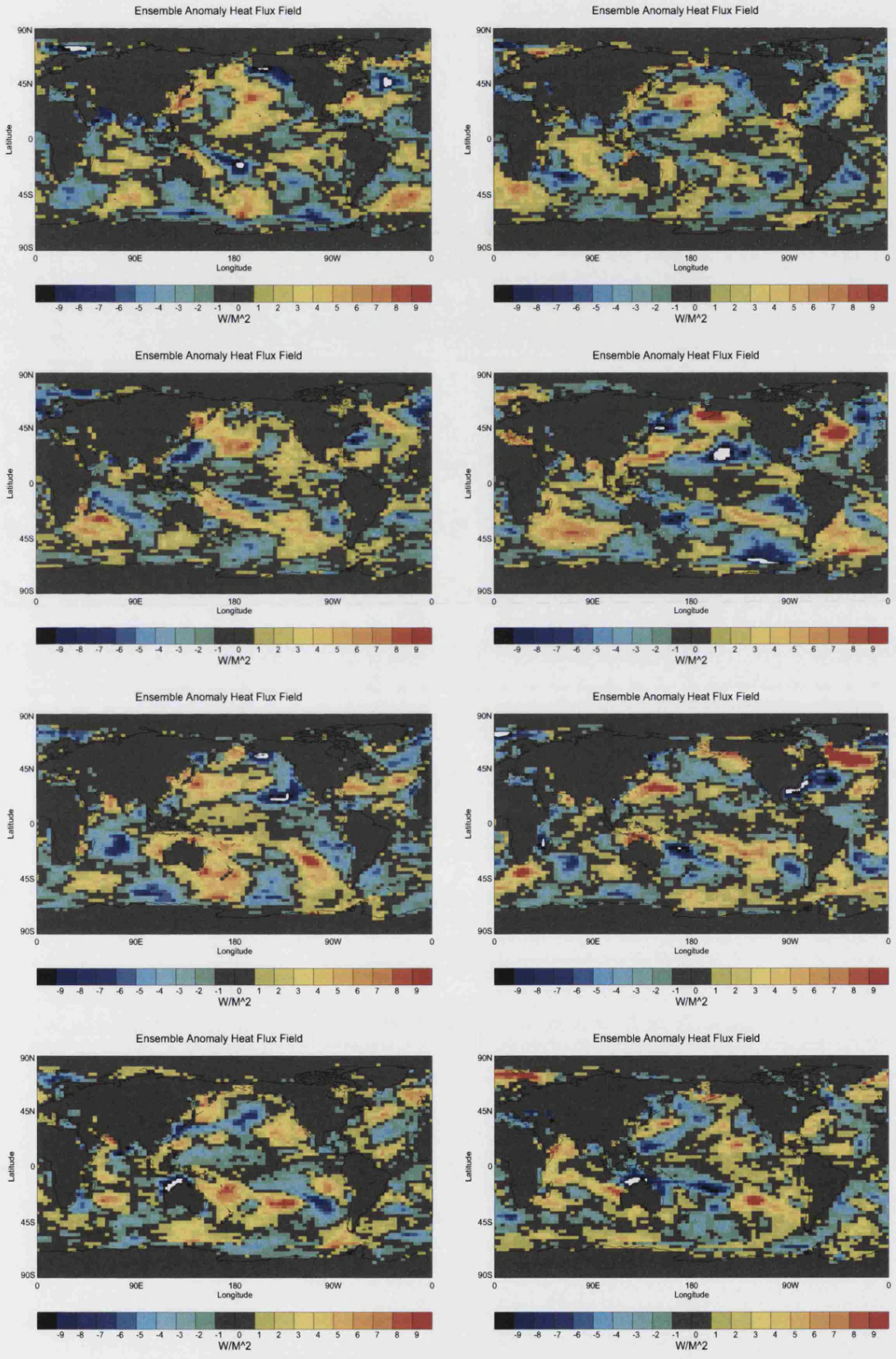


Figure 5.5: The HFA for 8 simulations randomly selected from the Standard ICE. The ensemble mean is subtracted from each simulation, giving an anomaly field.

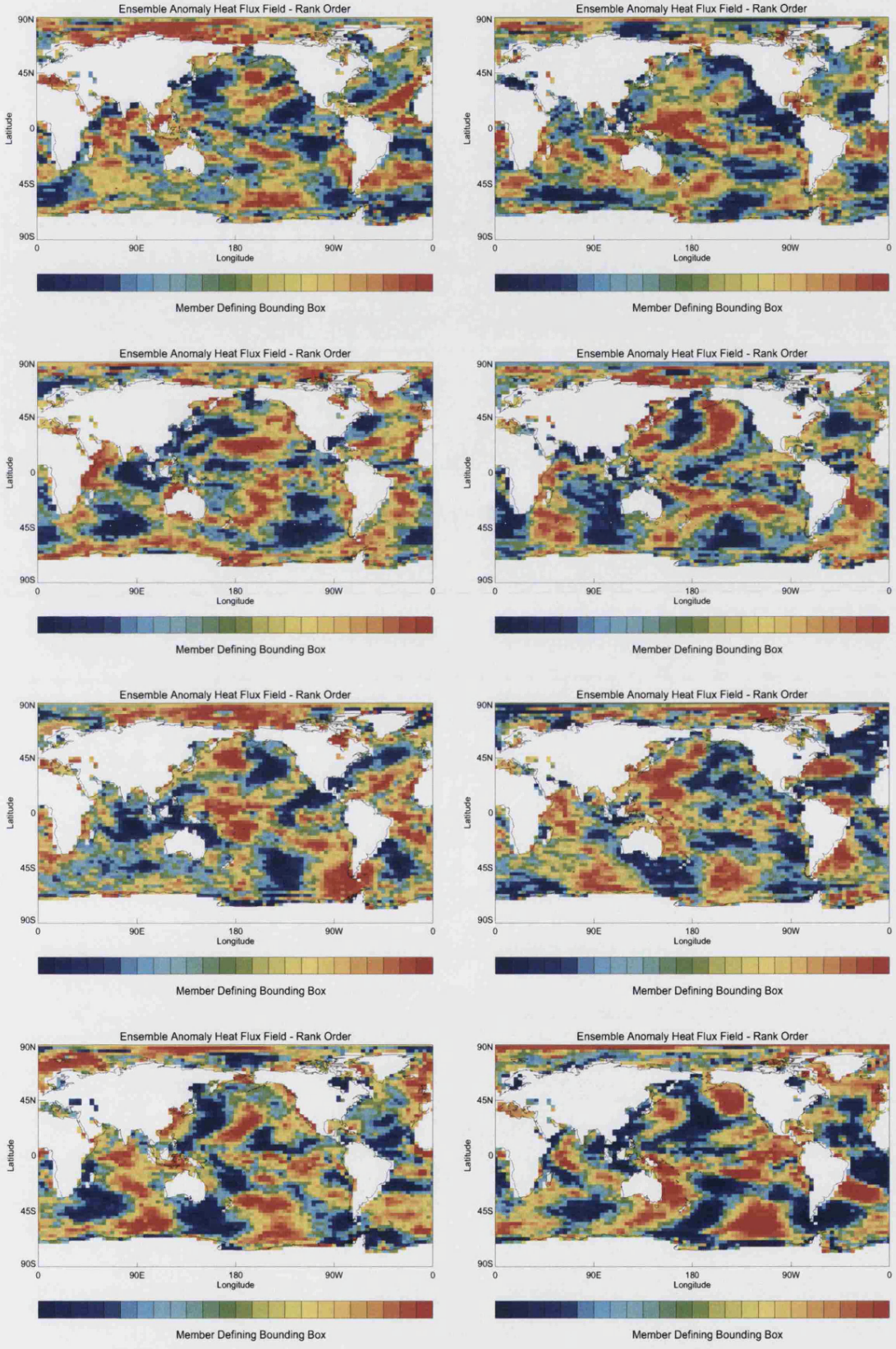


Figure 5.6: The HFA for 8 simulations randomly selected from the Standard ICE. The ensemble mean is subtracted from each simulation, giving an anomaly field, expressed in rank order.

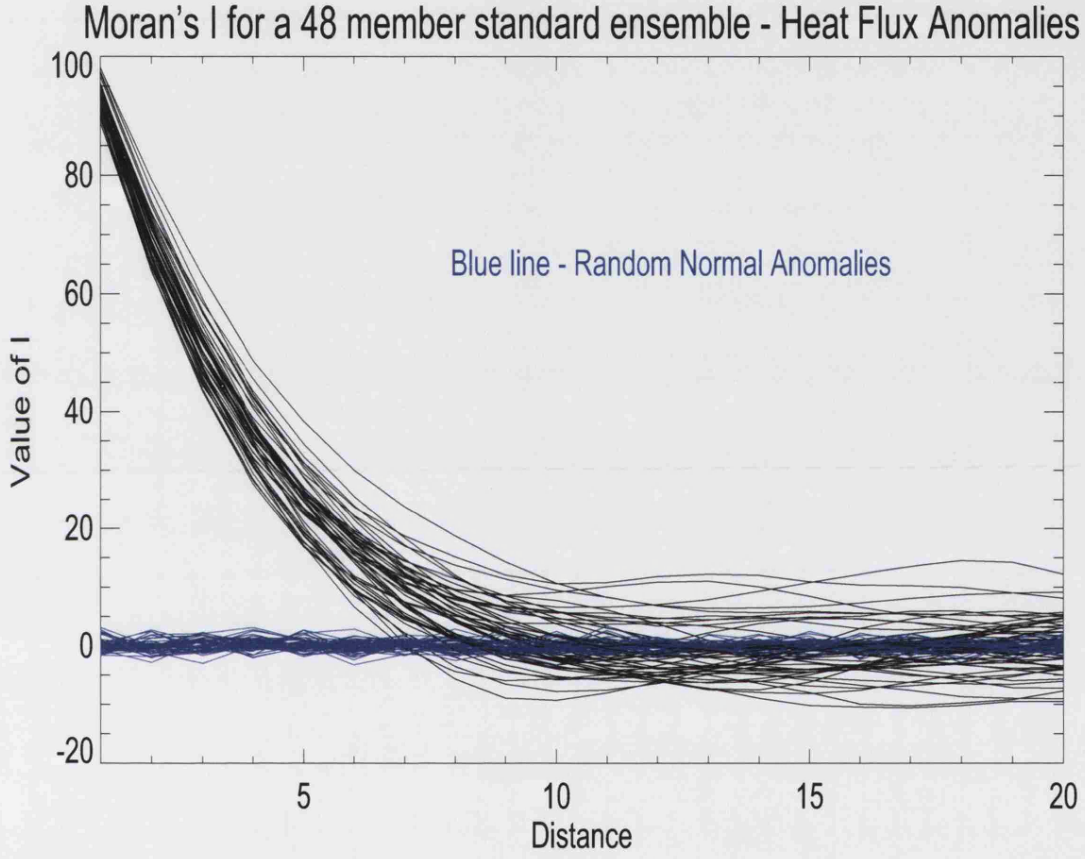


Figure 5.7: Moran's I statistic is plotted for the Standard ICE against distance in white. In blue, the statistic is calculated on a set of randomly generated data for comparison. Positive values of I indicate a positive correlation. For distances less than 7, grid–boxes show a positive correlation. Where the distance is greater than 10, there seems to be no significant correlation across the ensemble.

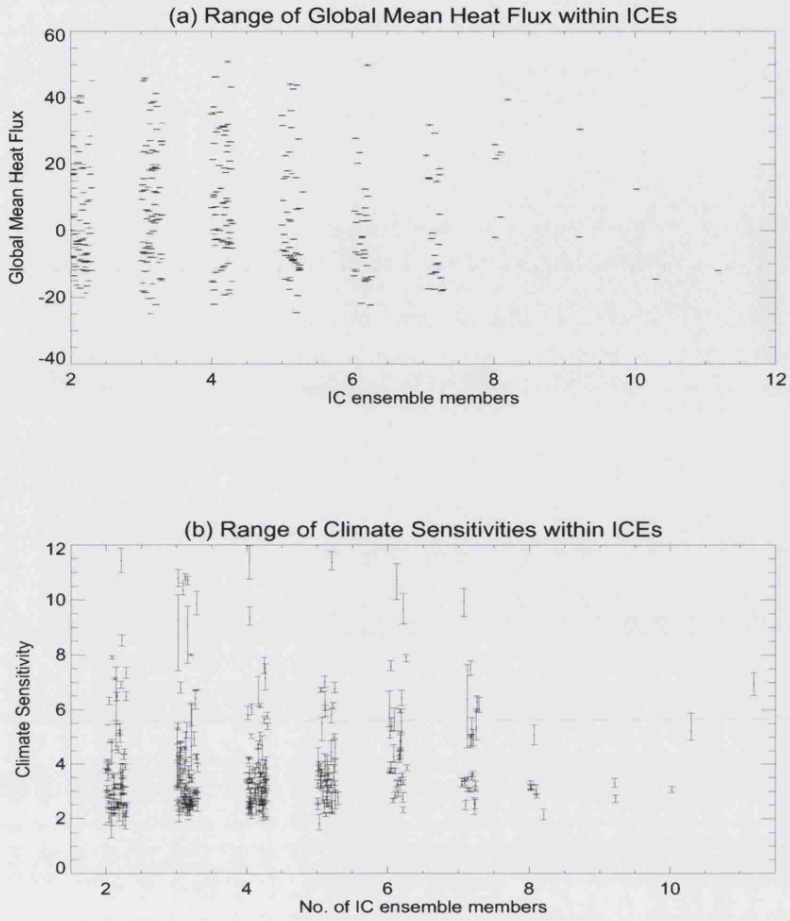


Figure 5.8: These graphs show the range of values for global mean HFA (top) and CS (bottom) as a function of ICE size. The range of values for global mean HFA is often so small that the minimum and maximum bars are almost indistinguishable. The number of ensemble members is "jittered" by adding a small amount of white noise so that each of the ranges is discernible. The top panel shows that whilst the global mean HFA can differ by over $70W/m^2$ between model versions, the range within each model version is very small - at most $0.419W/m^2$. The comparative range of values of CS within model versions is large in comparison to global mean HFA.

5.2.4 Stabilisation of Global Mean HFA

Figure 5.8 makes it clear that that there is very little variability in the global mean HFA within model ICEs in comparison to the differences between model versions. The CPDN experiment provides data to suggest that the HFA has stabilised by the

end of the calibration phase on the global level. But how long should the calibration phase be given a constrained amount of computational resources? This question is now addressed for the first time by examining the stabilisation, in time, of the HFA field over the calibration phase. The global mean HFA is recorded at each month of the calibration phase, giving a time series of 180 values.

Time series of global mean HFA is shown in Figure 5.9 which shows that the members within each model version are fairly close throughout the phase. There seems to be little discernible trend in the mean or variability of the HFA although the seasonal characteristics appear highly regular and similar across model versions, although the seasonal cycle is greater for higher CS simulations, shown in panel (c). The variability within model versions is small from very early on in the calibration phase. If the differences between ICE members are small, and the HFA has stabilised by the end of the calibration phase, it is reasonable to consider that the HFA does not introduce systematic differences to the individual members of an ICE. Under these conditions it is possible to treat the ICE members as drawing from the same distribution.

A shorter calibration phase would save experimental resources. Figure 5.10 shows the difference between the first 8 year ensemble mean and the last 8 year ensemble mean. If the difference between these 8 year means were insignificant the calibration phase might be truncated to 8 years. There is little difference between the two—often less than 0.1 W/m^2, although there is a tendency for simulations that have the lowest HFA fields to take increasingly more heat out of the oceans during the calibration phase, indicating the HFA has not stabilised buy the end of the first 8 years. This reduction in HFA is shown in Figure 5.11. As seen in Figure 5.9 the HFA field stabilises very quickly for most simulations (stabilised to within the degree of variability shown in the control ensemble) although for very high CS simulations, it may be necessary to run the calibration phase for its full length so that the HFA field reaches a stable level. It may not be possible to truncate the

calibration phase and it is not easy to tell which simulations have converged. The very low HFA fields, in particular, require more heat to be taken out of the oceans during the last 8 years of the calibration phase than the first 8 years, whereas other simulations stabilise almost immediately. Nevertheless, this drift takes place over the first few years of the calibration phase, allowing for the possibility for cutting this phase short, saving computational resources. There are, however, three reasons why cutting the calibration phase short may be undesirable:

1. It may not be possible to know, before running the calibration phase, which model versions will show the change in HFA shown in Figure 5.11.

2. A 15 year calibration phase allows ease of comparison to the control and doubled CO_2 phases. It might be useful to run the calibration phase for a full 15 years for this purpose, even if the HFA has converged very early in the phase.

3. The final 8 years of the phase are taken to provide an estimate of the equilibrium HFA field. This reduces the internal variability of the HFA and relies on the assumption that the HFA has nearly stabilised after 8 years. It should be noted that the internal variability of the HFA can be further reduced by taking the ICE mean. This averaging over ICEs may not be easy to achieve within the experimental design of a distributed computing experiment. On the other hand, a future experiment could run the first member of an ICE with a calibration phase and apply this for all subsequent simulations using the same model versions i.e. after the first, simulations would contain no calibration phase.

That the HFA reduces progressively throughout the calibration phase might be an indication for a tendency of these simulations to warm up quickly or uncontrollably and might be a factor leading to their high GMST increases under a doubling

of CO_2 concentration. A relationship between CS and the HFA is shown to exist in Section 5.4 and discussed.

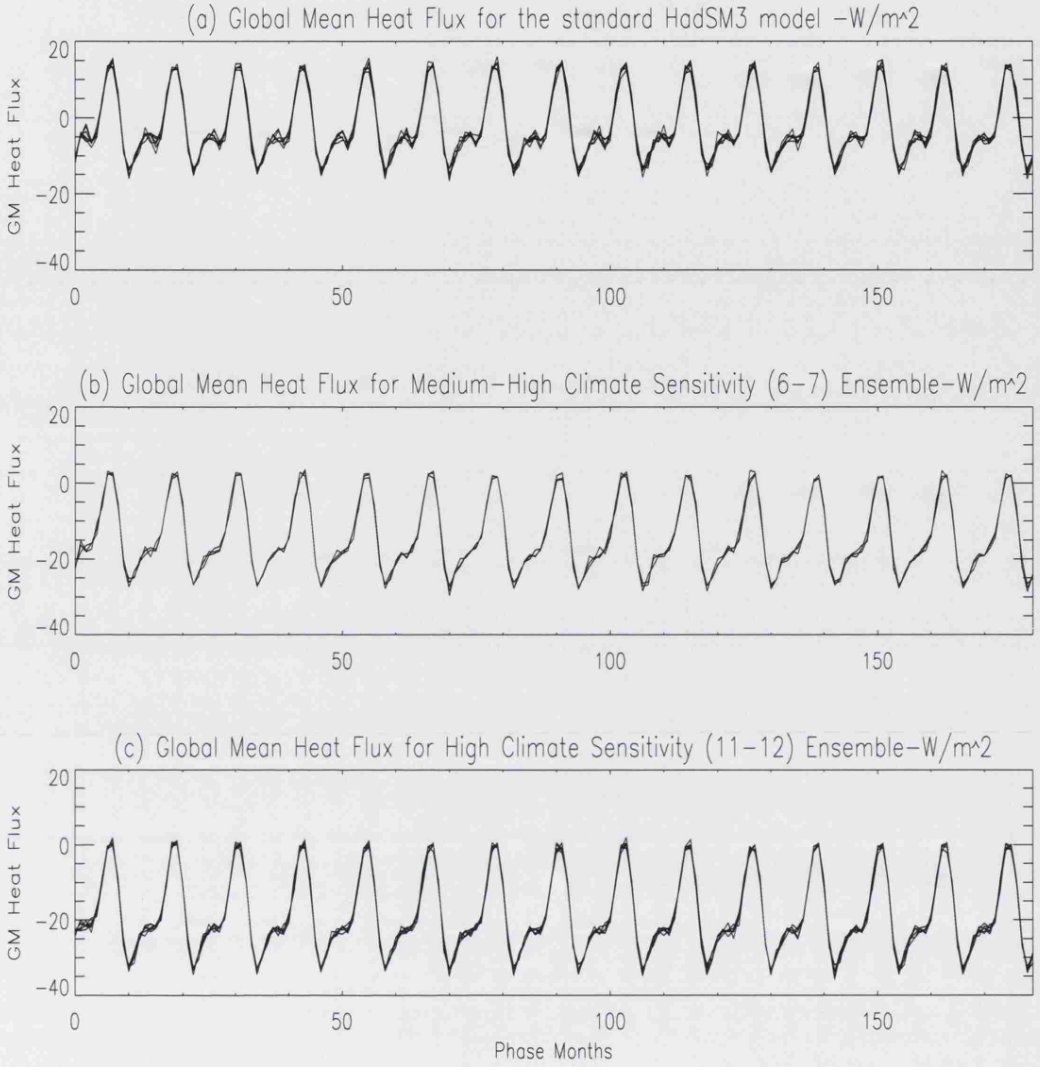


Figure 5.9: These graphs show the global mean HFA for the calibration phase. Time runs in months throughout the phase. Panel (a) shows the control ensemble (of 6 simulations, with an average CS of 3.4 degrees Celsius), panel (b) a randomly selected ensemble whose CS is 6.4 degrees Celsius (3 simulations) and panel (c) a 11.1 degree CS ensemble (7 simulations).

5.3 Seasonality in the HFA

This Section studies seasonal variability in the HFA, in particular it is argued thats:

1. The HFA acts to mimic the ocean's effect on the seasonal cycle, by transferring heat out of the ocean during warm seasons and into the ocean during cold

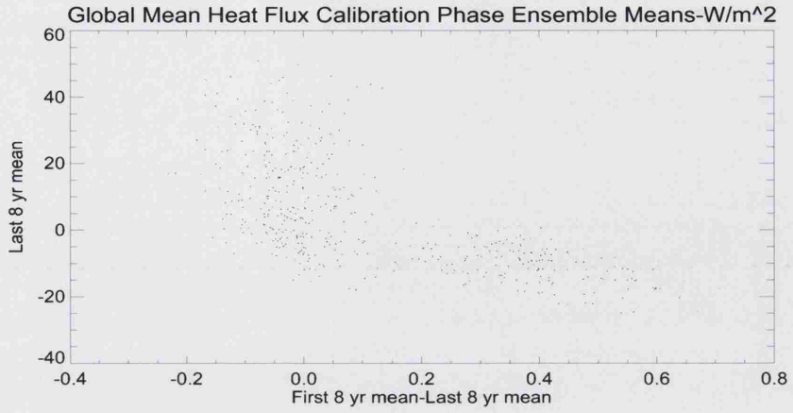


Figure 5.10: The y-axis shows the final 8 year mean global mean HFA for each of 484 model versions. The x-axis shows the *first* 8 year mean minus the last 8 year mean. These values are fairly close, but with a tendency for simulations with negative values of global mean HFA to remove more heat during the last 8 years than the first 8 years.

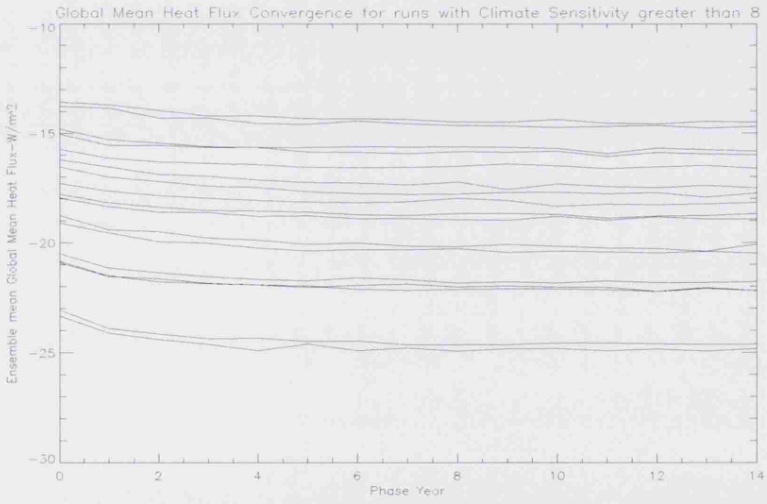


Figure 5.11: The global mean HFA is shown here for model versions with an average CS of 8 degrees or higher. There is an initial drop in the global mean HFA field, followed by a stabilisation.

seasons. This transfer of heat is similar to the oceans' dampening effect on the seasonal cycle.

2. The magnitude of HFA can be larger when analysed as seasonal than annual means. This fact could have an impact on the effect of HFA on instigating

systematic regional biases.

Data from the last 8 years of the calibration phase is used in this Section; these data were available as monthly means. A distinct seasonal pattern can be seen at the global mean level from Figure 5.9. In order to better understand the seasonal variability on a spatial level, the 8 year mean fields are looked at in Figure 5.12. Figure 5.12 shows the HFA field, averaged over the Standard ICE, for each season. In particular, the HFA transfers heat into the oceans in the winter and takes heat out of the oceans in the summer. This is probably due to the use of a slab ocean which requires seasonally dampening to re–produce the observed seasonal cycle; the regulating effect of a deep ocean on the seasonal cycle is not present in a slab model. Figure 5.12 shows that heat is taken out of the ocean all year round around the equator. There are also clear seasonal patterns. During December, January and February (DJF), the Northern hemisphere shows large influxes of heat, especially in coastal regions. During the DJF season, the Southern hemisphere generally has a negative HFA. This means that the HFA field adds more heat to the oceans during the colder months. Similarly in the warm months (DJF in the Northern Hemisphere, JJA in the Southern) heat is taken out of the oceans. This suggests that the HFA acts to dampen the magnitude of seasonality in the model. This effect might be expected due to the role that oceans play in dampening the Earth's seasonal cycle. Since the ocean has a higher heat capacity that land (the ocean warms up and cools down more slowly than land) and there is no deep ocean in the HadSM3 model, the HFA acts as a proxy for the dampening effect of the ocean. It should be noted that the magnitude of the HFA can be larger on seasonal time scales e.g. in large areas of the Northern hemisphere the HFA puts up to $300W/m^2$ into the oceans during the DJF season and takes out up to $200W/m^2$ around the equator during the MAM and JJA seasons.

The East Pacific shows an interesting feature in the seasonal data, obscured in the annual average. From March to August the HFA shows a strong flux of heat out

of the oceans off the West Coast of South America, as shown in panels (b) and (c) of Figure 5.12. The HFA takes up to $200W/m^2$ from the ocean to the atmosphere in this region. This area has been identified in Stainforth *et al.* (2005), as having a tendency for a strong negative feedback. This effect may be amplified or induced by a large negative HFA. Further investigation of this effect is carried out in Section 5.5.

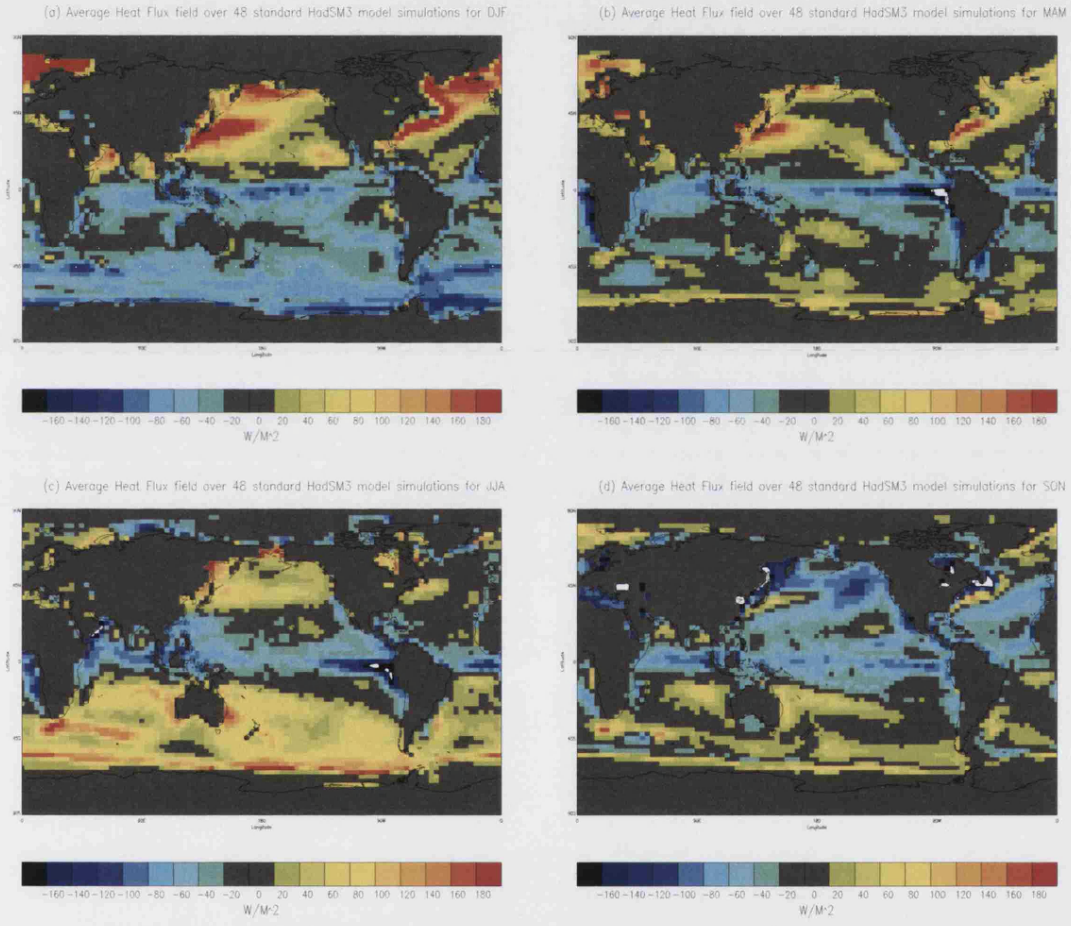


Figure 5.12: The HFA field for (a) DJF, (b) MAM, (c) JJA and (d) SON averaged over the Standard ICE.

5.4 HFA and Climate Sensitivity

The CPDN experiment shows that GCMs are capable of producing values of CS significantly higher than the typical 4–5 degrees Celsius estimated high range Houghton

et al. (2001); Solomon *et al.* (2007a) for CS. Furthermore, it has been widely hypothesised that HFA has no direct impact on CS Changnon *et al.* (2000); Houghton *et al.* (2001). In order to investigate possible relationships between HFA and CS, Figure 5.14 shows the global mean HFA plotted against CS. There is a pattern for simulations with large negative values of the HFA to result in high CS. All simulations in PPE_{2578} with a CS of more than 8 degrees Celsius have a global mean HFA of less than -13 W/m^2. The very tight distribution of the annual mean HFA within each model version is the cause of the apparently streaky structure seen in Figure 5.14.

It is important to try and understand the physical reasons for the relationship between HFA and CS seen in Figure 5.14 in order to interpret simulations with high CS. It can not be concluded directly that the HFA has an effect on CS since there could be some other reason for their correlation or a confounding factor. Figure 5.13 shows the calibration phase total global cloud cover plotted against global mean HFA. There is a clear pattern for simulations with low total cloud cover to require negative HFA. The pattern seen in Figure 5.13 might be due to model versions with low amounts of cloud cover, and thus low reflectivity of solar radiation, requiring HFA as a surrogate cooling mechanism. Figure 5.13 shows that there is a relationship between clouds and the HFA and Figure 5.14 a non–linear relationship between HFA and CS. The HadSM3 model confirms that the amount of cloud cover is a key property for understanding high simulated CS in GCMs, as has been previously noted Solomon *et al.* (2007a). It might then be tempting to use simulated cloud amount to constrain the range of values of CS. It would be statistical bad practice to use the observed relationships between CS, HFA and cloud cover to constrain the range of values for CS without understanding how these relationships arise. Such constraints face the danger of a selection bias whereby constraints are

selected based on their effects rather than their physical basis. The search for observational constraints must be based on the physical relationships between variables, not by searching for observational variables that would produce the desired result. It is essential that expert judgements for constraints should be stated before the simulations are available; this would avoid the problem of cherry–picking patterns in the data that can lead to misleading results.

Various mechanisms could explain the relationship between HFA and CS seen in Figure 5.14. It is possible that simulations with negative global mean HFA produce high values of CS because the HFA is introducing an artificial source of heat in the model atmosphere although, if this were true, we might expect to see a stronger tendency for simulations with a strong positive HFA to cool. It is also possible that the HFA might be acting as a surrogate cooling mechanism for a lack of clouds in the calibration phase (with pre–industrial CO_2 concentrations) but can not produce the same type of feedbacks when CO_2 concentrations are doubled.

In this latter case, it is important to note that there are still large uncertainties in the feedback effects of clouds across different models Cess *et al.* (1989); Webb *et al.* (2006). The type of clouds, their height and regional distribution also play an important role in determining the sign and magnitude cloud feedback effects Ringer *et al.* (2006). The question of constraining the range of values of CS using the HFA is looked at in detail in Chapter 7. It is also possible that the relationship between HFA and CS might be "confounded", to some extent, by another model aspect related to each of these components. This means that the relationship seen might not be a straightforward bias introduced by the HFA field on CS but a more complex mechanism.

It is not argued here that any of these explanations are definitive but that both statistical and physical understanding are necessary to gain insight to model behaviour and interpret important model features, such as high CS simulations.

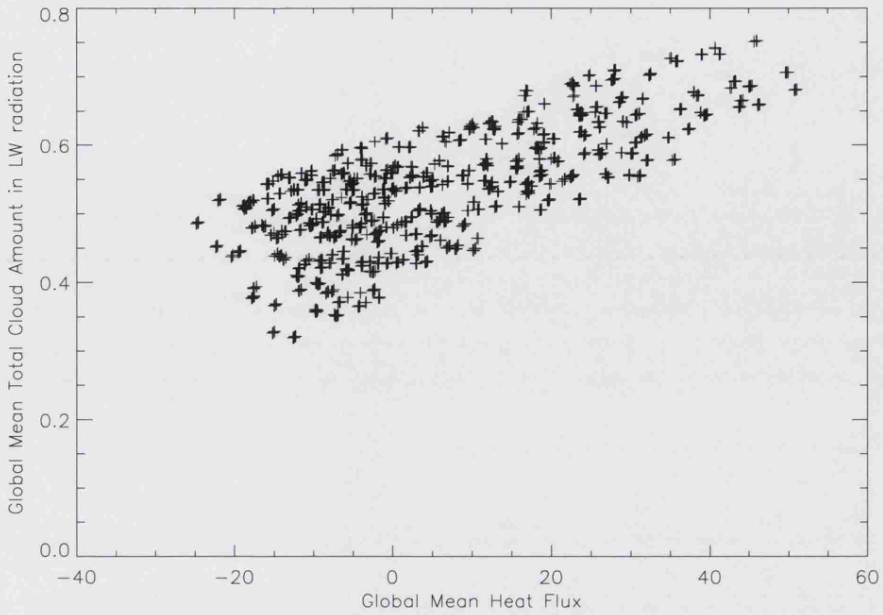


Figure 5.13: Total global cloud cover (as a fraction) is plotted against the global mean HFA for $PPE_{quality}$. There is a pattern for simulations with a low total cloud amount to have negative global mean HFA.

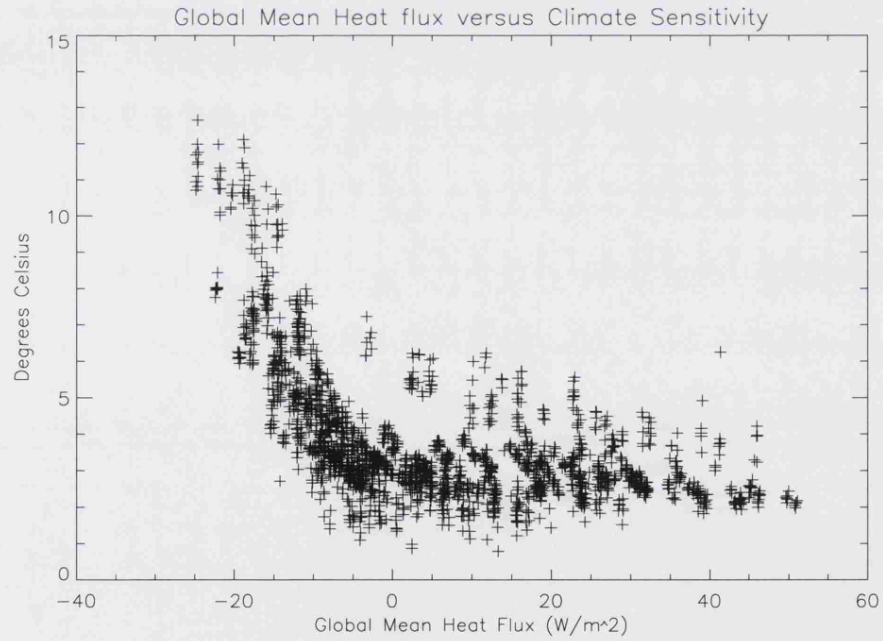


Figure 5.14: The global mean HFA is plotted against CS for $PPE_{quality}$. There is a distinct tendency for simulations with a large negative global mean HFA to produce simulations with very high CS.

5.5 HFA and drift

Despite the use of HFA in the CPDN experiment, a proportion of simulations (~30%) of PPE_{2578} display a strong negative temperature drift[1] during the control phase. These simulations are disregarded from the analysis of the effect of external forcings on model climate since the simulation exhibits significant unphysical climatic changes in the absence of external forcings. A smaller proportion of simulations show a positive temperature drift, but generally of a much lower magnitude than simulations with a negative drift (in PPE_{2578}, of 319 simulations with a significant positive drift, 3 have a positive drift greater than 1 degree Celsius per decade, whereas of 766 simulations with a significant negative drift, 265 have a negative drift of more than 1 degree Celsius per decade). An important negative feedback mechanism, first identified in Stainforth *et al.* (2005) and discussed in this Section, is a key component of negative drift. No such mechanism has been detected that results in unphysical positive feedbacks. Negative drift could arise from a mis–calibration of the HFA field or from some other source. This Section presents evidence to suggest that the HFA field (or its mis–calibration) is unlikely to be the sole cause of such model drift.

It would be expected that, if the HFA were a contributing factor to model drift, that there would be a discernible relationship between HFA and GMST drift. Figure 5.15 shows the proportion of simulations with significant control phase GMST drift plotted against the global mean HFA, for categories of size $2W/m^2$ (proportions are taken since the number of simulations in each category varies) in PPE_{2578}. PPE_{2578} is used in this Section to analyse GMST drift; $PPE_{quality}$ has been quality–controlled has thus does not contain simulations with significant GMST drift. There is no clear pattern for simulations with significantly non–zero global mean HFA to have a greater propensity to drift. Figure 5.15 suggests that a significantly non–

[1]Following Stainforth *et al.* (2005), a significant drift is defined as greater than 0.2 degrees Celsius per decade in magnitude. This drift is calculated by a linear fit to the last 8 years in the control phase.

zero global mean HFA is not an important determinant of GMST drift. Possible relationships between the HFA and model drift in a region previously identified as important for understanding model drift Knight *et al.* (2007); Stainforth *et al.* (2005) are now looked for.

The major known cause of significant GMST drift occurs in the East Pacific, just off the coast of Peru Knight *et al.* (2007); Stainforth *et al.* (2005). A cycle of cool oceans and high cloud cover lead to a progressively colder East Pacific that can cause the temperatures in this region to fall by as much as 27 degrees Celsius. It might be that a strong negative HFA in this region is triggering a cooling feedback process. The negative feedback begins in a particular grid–box in the East Pacific, henceforth called "Area 51". This Area 51 grid–box is used to characterise the cooling East Pacific problem and is used as a proxy for this local cooling feedback. The East Pacific problem is detected, using the temperature anomaly in the Area 51 grid–box, following on from Knight *et al.* (2007). The Area 51 grid–box in the Pacific is identified as (78.75 West, 2.5 North). An Atlantic grid–box of the same latitude (48.75 West, 2.5 North) is subtracted from the Area 51 grid–box. The calibration phase value for this difference is then deducted from the control phase value to give an anomaly statistic. This statistic has been presented in Section 4.4.2 and is used in the process of quality control. It might be expected that if the HFA is influencing the East Pacific negative feedback there would be a relationship between the HFA in Area 51 and the Area 51 temperature anomaly. It could be that the HFA persistently takes enough heat of the Area 51 region so that a local cooling feedback is initiated. Alternatively, the HFA might not initialise temperature drift in this area but may exacerbate an existent problem – since HFA is kept constant for each year of the control and doubled CO_2 phases, once such a negative feedback has begun the HFA will continue to take heat out of the oceans even when they are cooling significantly. If the HFA is a contributory factor to GMST drift, it might be expected that a strong negative HFA in the Area 51 grid–box would display some

relationship with GMST drift (as defined in Section 4.4.2, drift is measure as the linear fit to the last 8 years of the global mean temperature time series in the control phase). Figure 5.16 shows this Area 51 anomaly plotted against model drift. No clearly discernible pattern is shown. This could be because all the simulations have a HFA in this area below a certain threshold level (Williams (1999) points out that using a HFA of less than -40W/m^2 can lead to sea ice feedbacks in the HadSM3 model). Figures 5.16 shows that all simulations have a strong out–flux of heat of between -40 and -170W/m^2 during the summer season.

The association between HFA and model drift can be further explored through the effect of IC and parameter perturbation on model drift. Simulations within an ICE have very close global mean HFA. Figure 5.8 and 5.9 have shown that, on a global scale, the HFA fields are very close between ICE members. On regional scales there is more internal variability, as shown in Figure 5.4. Models sharing the same parameter values are much closer, on a regional level than simulations selected randomly across parameter sets. If the HFA contributes to model drift, it would not be expected that some simulations within a ICE would exhibit significant drift and others not. Such differences are, in fact, found between ICE members, as shown in Figure 5.17. A value of -15 degrees Celsius for the Area 51 statistic is used to define which simulations are acceptable based on investigation of the distribution of this statistic, as explained in Chapter 4. Of 374 ICEs with more than 1 member available (and no quality control applied), 47 (12.47%) had at least one member with an anomaly below this cut–off point of -15 degrees Celsius (no simulations show an Area 51 statistic below -15 degrees since this criteria is explicitly applied during the quality control process). Figure 5.17 shows that there is a considerable range of Area 51 anomaly values across ICs. There are ICEs with no simulations that pass this test for the Area 51 temperature anomaly and have a relatively small range. Other ICEs have a wide range (over 20 degrees Celsius), with some simulations proving acceptable and others being ruled out within the same model version.

Figure 5.17 suggests that, for at least some parameter values, the problem of East Pacific cooling can not be determined exclusively by the parameter values or the HFA in this area using a 15 year control phase. Despite this, for some model versions, all members have a very low Area 51 anomaly with little variability between ICE simulations (less than 1 degree Celsius in some cases). It should be noted that none of the Standard ICE members show a significant GMST drift; this suggests that parameter perturbation is a contributing factor to the Area 51 problem. It could be that some parameter values are more susceptible to this problem in general whereas other simulations fall into this category because of the IC variability alone.

It might be that if the control phase were run long enough the East Pacific problem would manifest itself across all simulations but there is no evidence for this assumption. Alternatively, it is possible that this effect occurs independently over parameter values, the only variable being at what point this problem occurs and can be detected. It is also possible that unphysical drift begins in the doubled CO_2 phase in cases where the HFA is no longer able to prevent instabilities in the new model state (some model simulations that do not exhibit drift during the control phase do show significant regional drift in the East Pacific during the doubled CO_2 phase). There are potentially some parameter values that are largely immune to GMST drift; this is suggested by the absence of significant GMST drift in the Standard ICE. These hypotheses could be tested by running a longer control phase (100 years or longer) for selected model versions that exhibit the Area 51 problem in **1)** No simulations and **2)** Some simulations. The control phase time series of GMST and Area 51 statistics could be stored and compared to analyse whether there are stable sets of parameter values and the extent to which IC perturbation accounts for the presence of GMST drift.

It is judged here likely that the problem of model drift results from underlying errors in the slab model that are exposed by parametrically de–tuning the standard

HadSM3 model from its stable state. Arguably, the HFA might be unable to stabilise the HadSM3 model in some parameter settings.

The HadSM3 model requires large adjustments of HFA, shown in Figure 5.2, particularly in the East Pacific. There may be limits to the extent that HFA can account for lack of a dynamic ocean. In particular, where HFA adjustments of the order of $100W/m^2$ are required, the validity of the HadSM3 model might be questioned. The use of such large corrections may not be justifiable if the HFA does not respond to feedbacks in a physical way or has a direct effect on key model properties, such as CS. Since unstable simulations are eliminated via quality control, it is not thought that subsequent analysis is susceptible to the problems analysed in this Section.

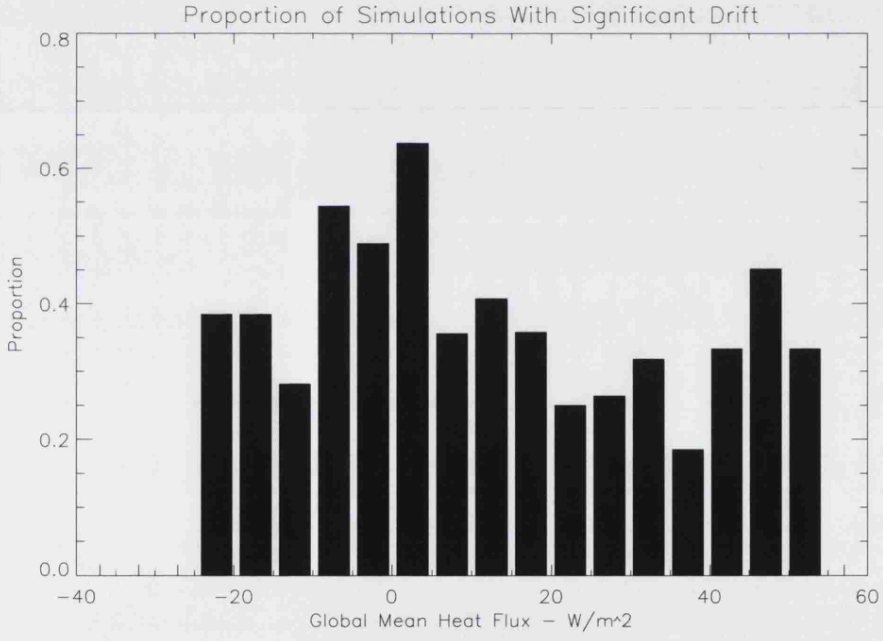


Figure 5.15: The proportion of simulations from PPE_{2578} with significant GMST drift is plotted for categories of global mean HFA of width $2W/m^2$. There is no clear tendency for simulations with a significant negative global mean HFA to have a significant GMST drift.

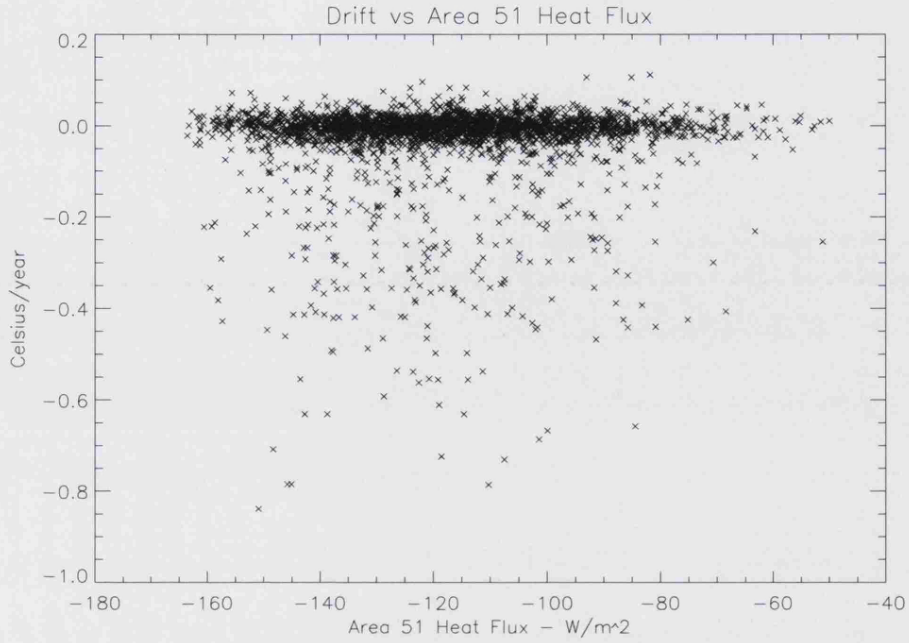


Figure 5.16: The Area 51, JJA, HFA is plotted against the control phase drift for PPE_{2578}. There is no clear pattern for simulations with a strong reduction of heat over Area 51 to have a strong negative drift.

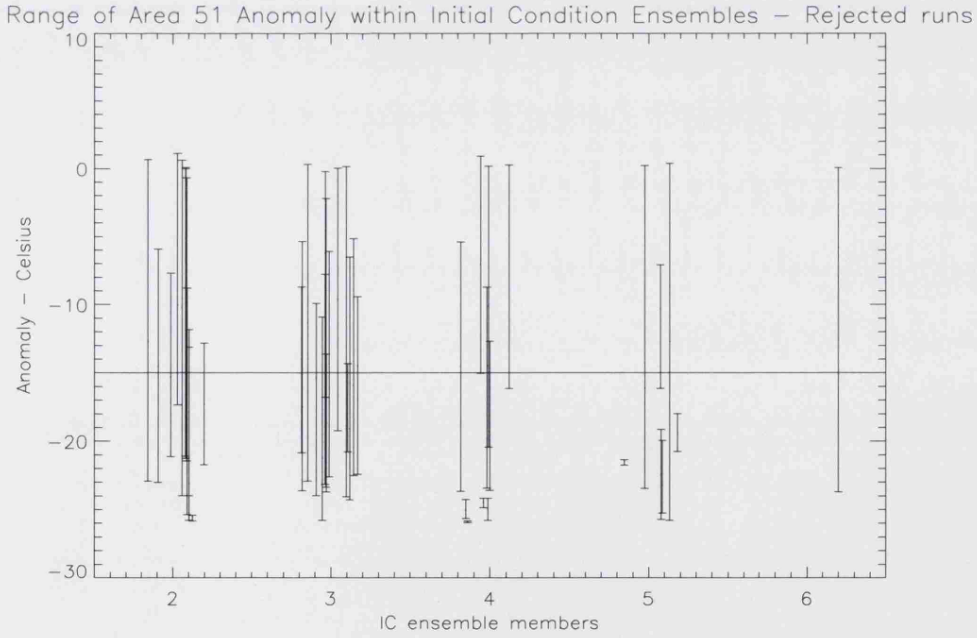


Figure 5.17: The range of anomalies within an ICE is plotted over the problematic grid–box. Of 484 ensembles, only those with at least one unacceptable simulation (an Area 51 statistic less than -15 degrees) are plotted (47 ensembles).

5.6 Conclusion

This Chapter has looked at variations in the HFA from different points of view; a comparison has been made between the impact of initial condition and parameter perturbation on the HFA, the stabilisation of HFA during the calibration phase, seasonal variation in HFA, the relationship between HFA and CS and the possible impact of HFA on model drift.

The relationship between HFA and climate drift has been looked at and, in particular, the question of whether the HFA could cause a specific negative feedback in the East Pacific. The results are inconclusive, although it seems unlikely that the problem of model drift is solely due to a miscalibration of the HFA field. It is more likely that the model structure is not stable over all sets of parameter values, particularly if a certain amount of interactive model structure–parameter tuning has occurred for the standard HadSM3 model. Further work, presented in Chapter 7, will analyse a larger set of CPDN simulations and will examine the use of HFA as a means of constraining model simulations.

Original results presented in this Chapter are:

- Perturbation of Initial Conditions has little effect on the global mean HFA (the greatest difference between IC members across 418 model versions is $0.419 W/m^2$). On the other hand, perturbation of Initial Conditions can lead to differences of up to $40 W/m^2$ (100 times the greatest global difference) on a grid–box level.

- It has been shown by analysing the leading Singular Vector in the HFA fields of ICEs that parameter perturbed model versions of HadSM3 can require significantly different global HFAs. The leading Singular Vector has been shown to explain significantly more variability in the HFA fields where model simulations share parameter values than where simulations are drawn at random. It was also also shown that whilst perturbing Initial Conditions has an impact

of less than $0.5W/m^2$ on the global mean scale, perturbing parameters can lead to changes of up to $70W/m^2$. The HFA should be calibrated for each set of parameter values, but can be averaged over ICE members to provide a more robust estimate.

- Significant seasonal variations exist in the HFA both globally and regionally. This effect is likely mimicking the seasonally-dampening effect of a deep ocean and means HadSM3's seasonal cycle might not respond to rising CO_2 in a physical way.

- There is a relationship between global mean HFA and CS. Simulations with high values of CS (over 8 degrees) tend to have strong negative global mean HFA (less than -$10W/m^2$). This relationship is potentially important for interpreting simulations with very high values of CS (greater than 8 degrees Celsius).

- Relationships between HFA and model drift have been investigated with the use of global mean and refined local statistics. No discernible pattern between global mean HFA and model temperature drift was found. The same model version can produce simulations that either drift or do not drift, suggesting drift is not dependent solely on parameter perturbation.

Chapter 6

ICEs and the Internal Variability of Climate Models

6.1 Overview

Climate is a distribution[1], consisting of a range of possible outcomes. A single climate model can exhibit a range of different states; this range of model behaviour is the internal variability of the model. Climate modelling is thus concerned with a long–term distribution and determining whether or not a range of forcings will result in significant changes in this distribution. This Chapter explains how the distribution of model climate can be evaluated using ICEs.

In a dynamical systems context, this internal variability can be understood in terms of the climate model's "attractor", the set of points to which a model will converge on long time scales. In a dynamical systems' context, internal variability arises from not knowing on which point of the model's attractor the system will lie at any given time point; there are a variety of possible states, depending on the precise specification of the IC[2]. The notion of simultaneously possible states is a multi–world

[1]Hence the saying, "Climate is what you expect, weather is what you get" Heinlein (1973).

[2]In the case of climate simulations, internal variability often refers to behaviour in an unforced model. In this Chapter, internal variability is used in a more general sense to include the variability of simulations under forcings. Such internal variability is not mathematically well–defined but is a useful concept when considering the effect of initial condition uncertainty on climate model

scenario and thus can only be studied in "model land". Unlike observations of the Earth's climate for which estimates of a single, multi–dimensional, observation are available at each time point, climate models can simulate the same time period repeatedly. Thus, variability in model space occurs not only with time, but as a distribution of possible states at each time point.

The existence of internal variability places limits on our ability to make precise statements about the impacts of climate change. Instead, the distribution of model output within an ICE can be used to assess the consistency of model response on both regional and global scales. The internal variability of a model can be studied using ICEs, under the same model and same parameter values. There have been few efforts to run large ICEs for this purpose, due to computational constraints and choice of experimental design. Results based on the CPDN experiment show that internal variability can be non–trivial for the Hadley Centre's HadSM3 model. This has important implications both for decision making and our understanding of the internal variability of climate models.

Some common statistical approaches fail to distinguish between internal variability and other forms of uncertainty that affect the model's dynamics, such as parametric uncertainty. One difficulty arises when the ICs are *mixing* Arnold & Avez (1968) in a sense that the parameter values are not; perturbed IC simulations result in identical long–term statistics, whereas perturbed parameter simulations can have very different dynamics Stainforth *et al.* (2007a). In particular, ICE members sample the same distribution, whereas changing parameter values can change the sampling distribution itself. The two types of ensemble should be kept conceptually distinct. ICEs have been used to reduce the variability of climate model output by taking ICE means Tebaldi & Knutti (2007). This Chapter presents two further uses of ICEs are shown that make use of information regarding the distribution of ICE members. The roles of ICEs presented here include:

response.

1. ICEs are used here to assess variability in the HadSM3 model's response to a doubling of CO_2 concentrations for different regions and variables. An ICE is necessary to understand how the distribution of model climates change in response to some external forcing. This method can be similarly applied to the comparison of different emissions scenarios or models that differ by parameter values or structure.

2. ICEs can be used to test for consistency of information in the model output, as laid out in Chapter 2. Such tests can provide a check for robustness of information in climate predictions through an evaluation of the magnitude of the model's internal variability on various length scales and in different variables.

This Chapter presents the first detailed discussion of the use of ICEs to evaluate the distribution of climate and distinguish between models using their respective levels of internal variability. The magnitude of internal variability seen in climate simulations provides a "strawman" test for the consistency of model projections. This relates to the ICE test outlined in Chapter 2. Decisions made based on climate models should be robust to the magnitude of internal variability seen. It is shown that there is significant variability within the standard HadSM3 model on regional scales in key variables.

The structure of this Chapter is as follows: Section 6.2 introduces the use of ICEs to evaluate the internal variability of climate models. Section 6.3 looks at the impact of IC perturbation under both a control (unforced) climate and a forced scenario (doubled CO_2 concentrations). The effect of IC perturbation, whilst small on global scales compared to parameter perturbation Knight *et al.* (2007), is shown to be non–trivial on regional and local length scales. Section 6.4 shows how ICEs might be used to distinguish between different forcing scenarios, model versions or multi–model ensembles. Section 6.5 discuses the role of ICEs in transient–forcing climate experiments.

6.2 Introduction to ICEs

Uncertainty in climate projections can be ascribed to a number of different sources Stainforth *et al.* (2007a) (also see "Uncertainties and Ensembles" in Chapter 2). These uncertainties vary in magnitude with length scale and the variable of interest. One of these types of uncertainty, ICU, is considered here. ICU is an irreducible uncertainty and is an expression of the model's internal variability.
The distribution of model climate needs to be understood in order to make a judgement on whether the model's response lies outside the range of internal variability. There are at least 2 different ways to estimate the internal variability of GCMs:

1. Internal variability can be assessed using a very long simulation under the same scenario Min *et al.* (2005); this approach is only valid in *equilibrium experiments* and not *transient experiments.*

2. ICEs can be used to assess internal variability in either equilibrium or transient settings. A single, long, simulation can be useful for detecting drift in the model but can not be used in transient experiments. In contrast ICEs can provide a distribution of climate over time.

Unlike in weather forecasting, where much computational effort is spent assessing error growth due to uncertain ICs, there has been little consideration given to the role of ICs in climate modelling. Many methods of analysis implicitly assume that ICs do not have a significant impact on model climate Tebaldi & Knutti (2007) and experimental designs often run only a few members (and frequently only one) Houghton *et al.* (2001); Solomon *et al.* (2007a). This constrains our ability to quantify internal variability in climate change simulations and to distinguish models and forcing scenarios.
A 64 member ICE is studied in this Chapter in order to understand the standard HaDSM3 model better and to provide additional understanding of the model's internal variability.

Since members of an ICE are samples from the same dynamical system (the model attractor, in the case of equilibrium experiments) ICEs might be used to make probabilistic statements about the model attractor. These probabilistic statements are reliant on our ability to assume that members of an ICE are, in effect, independent draws from the same distribution. These assumptions are not tested here, nor might they be easy to test rigorously since this would require further assumptions regarding the nature of the model and its attractor. It is important to note that whilst the terminology of dynamical systems is often used (e.g. climate models are often referred to as "chaotic" Liu *et al.* (2008)), dynamical systems theory is of limited use in the case of climate models; no computer model is truly aperiodic nor would a 10^8 dimensional model attractor be easily amenable to analysis. A pragmatic approach is therefore adopted and terminology used for the sake of convenience.

Other types of ensemble, such as multi–model ensembles, draw from diverse attractors, and thus the statistics produced are critically dependent on the sampling strategy and choice of models used Frame *et al.* (2005). ICEs explore variability *within a particular model*, whereas perturbing parameters or changing model structures explores uncertainty *across models*. Within a single model, it is possible to make probabilistic statements assuming independent and identically distributed samples. In contrast, the metric of model space (or even parameter space) is not well–defined in multi–model ensembles Allen & Stainforth (2002) and thus is not readily amenable to objective *probabilistic* analysis.

6.3 The internal variability of HadSM3

This Section looks at the internal variability of temperature and precipitation on regional scales in the HadSM3 model. The magnitude of internal variability is shown to be large, particularly on small spatial scales (up to 10 degrees Celsius in some cases), on 8 year mean seasonal timescales. The magnitude of internal variability

is compared in temperature and precipitation between pre–industrial and doubled CO_2 concentrations. Furthermore, it is shown here that there is considerable variability within a 64 member ICE in both temperature and precipitation. Whilst the estimates of CS from the standard HadSM3 ICE are close (3.12–3.68 degrees Celsius, with a median of 3.37), warming is non–uniform across the globe and also varies significantly with season.

The ICE mean field can give an indication of the spatial variability of model response. Figure 6.1 shows the 64 member mean change in 8 year temperature from pre–industrial CO_2 (phase 2 of the experiment) to doubled CO_2 (phase 3) for each season (DJF, MAM, JJA, SON). In general, the oceans warm by around 2 degrees Celsius, whilst land areas typically warm by 3 to 6 degrees with the centre of large land masses warming the most. A striking feature of the model is that the high Northern latitudes warm by up to 10 degrees Celsius in the DJF season, but can show very little warming, or even *cooling* (areas shown in white) in large areas during the JJA season. Other areas show a much more consistent warming pattern, such as central North America which warms by by around 5–7 degrees all year round. The ICE contains information beyond the mean – the internal variability of the HadSM3 model can also be evaluated. The variance across ensemble members is shown in Figure 6.2; this variance is a representation of IC uncertainty. The areas in black in Figure 6.2 represent an ensemble variance of less than 0.2 degrees Celsius, occurring mostly over the ocean areas. The oceans warm up by less than land areas (as seen in Figure 6.1) and the ensemble variance is low (mostly less than 0.2 degrees Celsius). Over land there is a typical ensemble variance of 0.5 degrees Celsius, with the high latitudes showing a variance of up to 2 degrees in their winter season. The Arctic region has a high variance where extreme warming is predicted in some seasons and a variance close to 0 where no warming, or a slight cooling is projected in other seasons.

As an alternative to plotting the variance of the ICE, the range of behaviour can

be studied using a bounding box. Figure 6.3 shows the ICE range in 8 year temperature change under a doubling of CO_2 concentrations. This range is calculated by subtracting the minimum temperature change from the maximum at each grid box. Figure 6.3 shows that there can be a difference of up to 10 degrees Celsius in 8 year mean seasonal warming, due to IC perturbation alone. There are large areas in all seasons with 3 degrees or more of disparity between IC members. There are significant variations in the ensemble mean temperature change by region and by season. The internal variability of HadSM3 varies with season and region.

In precipitation, a *democracy plot* is shown in Figure 6.4, showing the percentage of ensemble members at each grid box for which precipitation increases under a doubling of CO_2. A democracy plot shows the extent to which an ensemble agrees on the sign of precipitation change and are most useful when representing binary information (such as change of sign or whether a certain threshold is exceeded) in variables expressed in units that are not readily intuitive (unlike degrees Celsius change for temperature).

Areas in red show a consistent reduction in precipitation over the ensemble; black areas denote regions in which almost all the ensemble members show an increase in precipitation. Over much of the globe, the internal variability of HadSM3 is such that the ICE does not agree on the sign of local precipitation change. It is interesting to note that a similar plot is shown in the IPCC AR4 Summary for Policymakers (SPM), Figure SPM.7. The democracy plot show in the SPM is based on a multi–model ensemble and shows a similar pattern of model agreement in precipitation response to rising CO_2, although the SPM Figure is based on a 10–year mean transient response and Figure 6.4 is based on 8–year mean equilibrium response. This suggests that the apparent disagreement between models seen in the SPM democracy plot might be due to each model having an indeterminate precipitation response, arising from internal variability, rather than due to differences between structural models. This would mean attributing uncertainty in precipitation re-

sponse to IC uncertainty and not to model uncertainty. It would be necessary to study an ICE from each of the constituent models used in the SPM to test this hypothesis.

There are significant regional differences at grid–box to regional length scales for 8 year seasonal means that are hidden in global mean statistics. Averages over larger spatial or temporal scales will likely result in a more conclusive "vote", in the case of precipitation, due to cancellation of differences and a reduction in variance.

There is a significant amount of variability due to IC perturbation alone, especially for regional[1] simulations of precipitation. This suggests that the internal variability of climate models can be large on regional length scales.

6.4 ICEs and Robust Model Response

This Section addresses a key use of ICEs; when comparing different scenarios, it is important to know whether the model responds in a way distinguishable from internal variability or whether the change seen could be a result of chance. It is shown here how ICEs can be used to robustly distinguish between model scenarios or model versions and how to interpret the results probabilistically. The use of ICEs to evaluate the consistency of information in climate model simulations was outlined in Chapter 2.

Given only the ICE mean change, without including information on the magnitude of model internal variability it is uncertain whether the changes in the ICE mean are significant. Here, an alternative view of model response is used that takes into consideration the range of model internal variability. Model output is looked at a distribution of values, not just the ICE mean. Whilst the ICE mean response might be of interest in some cases, the approach adopted here is useful for assessing the consistency of model response. The 64 member ICE of the standard HadSM3 model is analysed in Section 6.4.1, and two smaller (8 and 12 members respectively)

[1]Regional refers to the 22 Giorgi regions Giorgi & Francisco (2000).

HadSM3 model versions, with different values of CS in Section 6.4.2. Distributions of climate change are compared in terms of robustness i.e. a consistent model response that is beyond the range of internal variability. The presence of an overlap between distributions in key variables highlights the limitations in using small ICEs; it is possible for simulations within the same model version to produce different signs of precipitation change. Large ICEs can help to detect where model response is robust.

6.4.1 Regional response in the 64 member HadSM3 ensemble

An important use of ICEs is to distinguish between models or forcing scenarios. In this Section, a 64 member ICE is used to examine whether a significant response can be detected in model behaviour to a doubling of CO_2 concentrations. A single model run under each of these conditions does not allow the two different scenarios to be statistically distinguished since we can not reliably estimate the model's internal variability. With a few simulations under each scenario, it is only possible to make a limited assessment regarding the internal variability of the model under each level of CO_2 and judge whether the two distributions are significantly different. On the other hand, large ensembles enable robust analysis of model response; it is highly likely that two sets of 64 simulations would overlap in distribution if there were no difference in model behaviour between the pre–industrial and doubled CO_2 scenarios. Distributions of model response are compared in this Section in terms of the magnitude of overlap observed; this overlap represents a measure of the distinguishability of these distributions[1]. In the case of the CPDN experiment, the degree of overlap between the pre–industrial and doubled CO_2 phase distributions of the standard HadSM3 model ICE is a quantification of the consistency of model response to doubled CO_2. Thus, when no overlap between model distributions is seen

[1]The overlap between distributions can also be thought of as the chance of observing a response in the opposite direction to the ICE median in the case where only a single simulation is run

under different forcing scenarios, it can be concluded that there is a significant difference between these distributions, with the level of significance depending on the number of ICE members. Clearly, if the number of simulations in each distribution is very low, this significance level would be insufficient to conclude a statistically or physically meaningful result.

The probability of observing no overlap between two distributions of size m and n can be calculated non–parametrically as follows: under the assumption of independent and identical draws from the same distribution, an ensemble of size n is expected to cover $100*(\frac{n-1}{n+1})\%$ of the probability mass[1] Weisheimer *et al.* (2004). Take the ensemble of size n to be the base ICE (e.g. pre–industrial CO_2) to which the distribution of size m is to be compared. The probability of a simulation, drawn from the same distribution as the base ICE, not falling within the range already covered is $\frac{2}{n+1}$ – a probability of $\frac{1}{n+1}$ of falling below all n simulations and a probability of $\frac{1}{n+1}$ of falling above. The probability that m simulations are drawn from the same distribution as the base ICE and all fall outside the range covered would be $2.(\frac{1}{n+1})^m$, which becomes extremely small for large values of n and m. Where n and m are greater than 3 this probability is less than 5% and in the case of the 64 member standard HadSM3 ICE (n=m=64) the probability of observing no overlap, under the assumption that both sets of simulations are drawn form the same distribution, is less than 10^{-117}. Where distributions do overlap, which is almost certain to occur if the 64 member ICEs are drawing from a common distribution the model produces a result that is within the range of internal variability. Clearly, these probabilities only represent whether there is a statistical difference between distributions and not whether the difference is meteorologically significant. In the case of overlap, the level of distinguishability between the distributions can be estimated as follows.

As explained above, the larger the ICE, the greater the chance of an overlap if there

[1]The mean of the distribution of probability mass covered. More detailed statistics of this distribution can be calculated analytically, using the theoretical properties of order statistics, if required.

is no difference between distributions. In order to account for the varying size of ICEs, a distinguishability criteria can be used, such as the overlap in distributions used in Smith *et al.* (2008) and in Chapter 8. The probability of overlap is estimated by sampling simulations at random from each distribution and asking whether one simulation is greater (e.g. hotter or wetter) than the other, where members are drawn at random from the whole ensemble, with replacement and bootstrapped Efron & Tibshirani (1994) to obtain robust estimates. This method will be applied to data shown in Figure 6.5.

Figure 6.5 shows the distribution of temperature and precipitation change for Northern Europe in panels (a) and (b) and Central North America in panels (c) and (d). The two sets of simulations show no overlap in temperature in panels (a) and (c); the model shows a response outside the probable bounds of internal variability. In precipitation, the distributions overlap for the Central North America region in panel (d), but not for Northern Europe (of the 22 land regions used there is an overlap in precipitation in 12 cases; in no region is there an overlap in temperature). The HadSM3 standard model shows the possibility for the level of precipitation to either increase or decrease in Central North America under a doubling of CO_2, within the bounds of the models' internal variability.

In the case of Central North American precipitation change, there is an overlap in distribution; the probability of this overlap can be estimated using the re–sampling method described above. Based on a re–sampling of 10000 times, this probability is $\sim 8\%$ (repeating this method 1000 times in no case gave a probability that differed by more than $\pm 1\%$ from this value), demonstrating that there is approximately a 8% chance that a randomly selected simulation from the pre–industrial CO_2 phase is wetter in the Central North American region than a randomly selected simulation from the doubled CO_2 phase, where the median change of the ICE is for simulations to get wetter in the doubled CO_2 phase. This indicates that the sign of precipitation change can occur in both directions for individual simulations, as well as an overlap

in the distribution of simulated precipitation changes across the ICE.

This method of estimating the probability of an overlap allows the level of distinguishability of two ICEs to be quantified. If there is no overlap in the ICE distributions, and without involving distributional assumptions, it is not possible to know how many further members would be required to see an overlap. To detect an overlap in distribution in Northern Europe DJF temperature and precipitation at least 65 members would be required (at least one more than the available 64), but it is not possible to say exactly how many more. The information provided by these data demonstrates that the standard HadSM3 model responds in Northern European DJF temperature and precipitation significantly outside the range of internal variability.

The method outlined in this Section provides information on **a)** In the case of an overlap, the probability of an overlap and **b)** In the case of no overlap, its expected probability given the number of ensemble members available without making distributional assumptions. This method has advantages over simply testing the whether the mean of distribution has changed – **1)** The non–parametric method proposed in this Section quantifies the consistency of model response; knowing that the mean has changed alone is not be a physically meaningful result for a decision–maker and **2)** The probability of a single simulation showing a response in the opposite direction to the ICE median is relevant to the utility of single simulation ICEs.

In the case of Northern Europe, the probability of no overlap, given that simulations from both scenarios are drawn from the same distribution is $2.(\frac{1}{63})^{64}$. It can be concluded there is a highly significant model response in DJF temperature and precipitation in the Northern European region and in DJF temperature in the Central North America region but not for DJF precipitation in Central North America. In the latter case, the model response to doubled CO_2 lies within the range of the standard HadSM3 models' internal variability.

6.4.2 Comparison of two Perturbed Physics ICEs

Another use of ICEs is to examine whether or not two different models are statistically distinguishable. In this sub–section, two ICEs from different model versions are compared. Taken from the CPDN grand ensemble, these ensembles have 8 and 12 members and a mean CS of close to 3 and 5 degrees respectively. These ICEs were chosen since they are the largest ICEs with CS close to 3 and 5 degrees respectively. Ideally the size of the ICEs would be the same to allow a more direct comparison but it is felt that it is better to use the largest possible ensembles available. The aim of the comparison in this section is to show that simulations from an ICE with a CS close to 3 degrees can be hotter/wetter than simulations from an ICE with a CS close to 5 degrees rather than to directly compare their diversities and hence is not critically dependent on the ICE size. Each ensemble uses the same structural model (HadSM3), but with different parameter values. Only ICs have been perturbed within each of these two ICEs. Taking two ICEs with different levels of global warming allows for a comparison of the distinguishability of the different scenarios simulated. Here, these two ICEs are compared at a grid box level.

Figure 6.6 shows the magnitude of IC variability for each ICE for 8 year annual mean temperature change under a doubling of CO_2. This range is found by subtracting the minimum temperature change (the change is defined by the 8 year mean surface temperature in the doubled CO_2 phase minus the 8 year mean surface temperature in the control phase) from the maximum temperature change at each grid box. For the 3 degree ICE, the magnitude of the model's internal variability is typically between 0.5 and 1.5 degrees Celsius, and can be over 2.5 degrees Celsius. The 5 degree ICE shows a wider range (this is to be expected since it has more ensemble members), with large areas showing a range across ICE members of over 1.5 degrees Celsius. Note that this range is lower than that shown in Figure 6.3 for the 64 member standard HadSM3 ICE. This is partly due to there being fewer ensemble members available for the parametrically perturbed HadSM3 model ver-

sions, and due to annual means being used instead of seasonal means.

The distinguishability of these ICEs can be compared by looking for grid boxes where the two ensembles overlap. This is done by comparing the maximum simulation in the 3 degree ICE to the minimum simulation in the 5 degree ICE. In temperature, the maximum simulation relates to the simulation showing the greatest warming and in precipitation the simulation showing the greatest reduction in precipitation (or the least increase).

This difference is plotted in Figure 6.7 for temperature (upper panel) and precipitation (lower panel). In temperature, there is an overlap (where the difference between the hottest 3 degree and coldest 5 degree is greater than 0) for 6619 of the 7008 grid boxes ($\sim$ 94%), with some regions in the Northern high latitudes showing a significant overlap of over 2 degrees Celsius. In precipitation, there is an overlap between the 3 degree and 5 degree ICE in 6780, or $\sim$ 97% of grid boxes. It is often impossible to robustly distinguish between local precipitation change between 3 degree and 5 degree simulations. These results depend on the difference in the ICE mean CS chosen; it is expected that ICEs with very close values of CS will overlap more than for ICEs with very different values of CS.

Decisions reliant on grid–scale climate change information must take into account at least the variability in the model's response due to ICU (about one degree Celsius over most land masses, as shown in Figure 6.7). This suggests a limit to the potential for the utility of climate models to inform local decisions. It has been shown in this Section how the magnitude of these limits can be evaluated using ICEs.

The detection of robust differences between scenarios is limited by the internal variability of the model. In order to attribute climatic trends robustly, the level of internal variability on relevant temporal and spatial scales must be taken into account. It has been shown that it is not always possible to distinguish model versions or robustly detect climate change in some variables and regions.

6.5 ICEs in Transient Experiments

In equilibrium climate studies, ICEs can be used to understand the model attractor under different forcings. Such simulations do not reflect particular times and dates or precise real world forcing scenarios. In the case of the first CPDN experiment, artificially simple forcing scenarios are used to study model response.

The interpretation of ICs in transient experiments differs from that in equilibrium experiments since there is no fixed model "attractor" to be sampled from. In the transient case, the model's attractor is being continuously changed by the forcing applied; the model has no long–run equilibrium state. In a transient experiment, the distribution of model climate changes with time.

Thus ICEs play a different role in transient experiments; providing insight into how the model's dynamics change under a particular forcing. The model's internal variability may not be directly comparable to the variability in the observed climate[1]. An ICE can help understand the internal variability of the model under transient forcings.

Recent studies Collins & Allen (2002); Cox & Stephenson (2007); Troccoli & Palmer (2007) have discussed the role that ICs play on timescales up to 10 years. There may be predictability beyond seasonal time scales arising from ocean features, such as El Niño or the Atlantic Meridional Overturning Circulation. Combining such sources of seasonal and inter–annual predictability with other forcing effects already included in some climate model simulations, such as variations of solar luminosity and anthropogenic forcings, could lead to decision relevant forecasts on 1 to 20 year time scales. ICEs can be used to understand the potential for providing skillful seasonal-to-climate forecasts through an understanding of the predictability that arises due to natural variability and the predictability that is due to long–term

[1]Aside from model inadequacies, with many processes omitted, such as volcanoes and solar fluctuations, one might not expect the same type of variability in a model as observed in the climate system itself.

forcing effects. The role that ICEs play in NWP evaluating the information in the systems current state can be combined with the use of ICEs to evaluate how the distribution of climate might change over time.

6.6 Conclusion

The internal variability of climate models has been investigated using ICEs. ICEs are often neglected due to sparseness of computational resources or low priority in experimental design. Two of the potential applications of larger ICEs have been presented in this Chapter. Firstly, they can inform modellers as to the natural variability within the model. This is especially important in transient studies, where it is impossible to use a long "control" simulation to evaluate the models' internal variability in time. Running an ICE allows certain specific historical periods to be reproduced e.g. the very hot European summer of 2003 and attempt to understand whether the summer was a "freak" event Stott *et al.* (2004). Such events can only be related to the impact of climate change by their risk of occurrence – no definitive causal statements can be made due to the existence of internal variability.

Secondly, ICEs can be used to compare different models and forcing scenarios e.g. the different scenarios of future forcings given in Nakicenovic *et al.* (2000) can only be distinguished from each other robustly with a consideration of internal variability. This allows an assessment of qualitative changes in the model i.e. the robustness of a precipitation increase under a doubling of CO_2 concentrations on different length scales.

The impact of internal variability has been shown to be significant, especially on regional and grid–box length scales. The relevance of a global metric should be questioned in light of the diverse changes occurring on regional scales e.g. slow–warming oceans, rapid and seasonally–dependent high latitude warming. Regional changes seen in model simulations should be considered as lacking robustness where an ICE fails to agree on the sign and approximate magnitude of the change. This has

shown to be the case when looking at the effect of doubling CO_2 concentrations. In some areas, there is an overlap of ICEs for 8 year mean temperature change. There is a limit to our ability to distinguish between different scenarios of climate change. In another example, two ICEs are compared on a grid box level using model versions with CS of 3 and 5 degrees respectively. Large overlaps in distribution are seen in key variables. The evaluation of ICU is of immediate relevance to decision–makers since **1)** ICEs provide a test for consistency of information that can assess where climate models show responses outside the range of internal variability and **2)** Scenarios of future change can only be robustly differentiated using some measure of the internal variability of climate models. Given these uses, ICEs can be of significant use to climate modellers and decision makers and need not be restricted to weather and seasonal forecasting.

New results presented in this Chapter are:

- The availability of a large ICE allows for a quantification of the HadSM3's internal variability, which has been shown to be significant on length scales relevant for impact studies and adaptation decisions. The various roles of ICEs are discussed and their increased use encouraged. These results are significant since it has typically been assumed that the effect of perturbing Initial Conditions on climate simulations was negligible Tebaldi & Knutti (2007).

- It has been shown for the first time that Initial Condition perturbation can have a significant effect on model behaviour on relevant length and time scales in temperature and precipitation. The sign of change for 8 year mean seasonal precipitation is unanimous in only $\sim 3\%$ of grid boxes. In temperature, 8 year mean seasonal differences is shown to be as large as 10 degrees Celsius in some grid boxes. Such large differences are not usually considered possible to have arisen from Initial Condition perturbation alone and these results could affect the experimental designs and the interpretation of model variability.

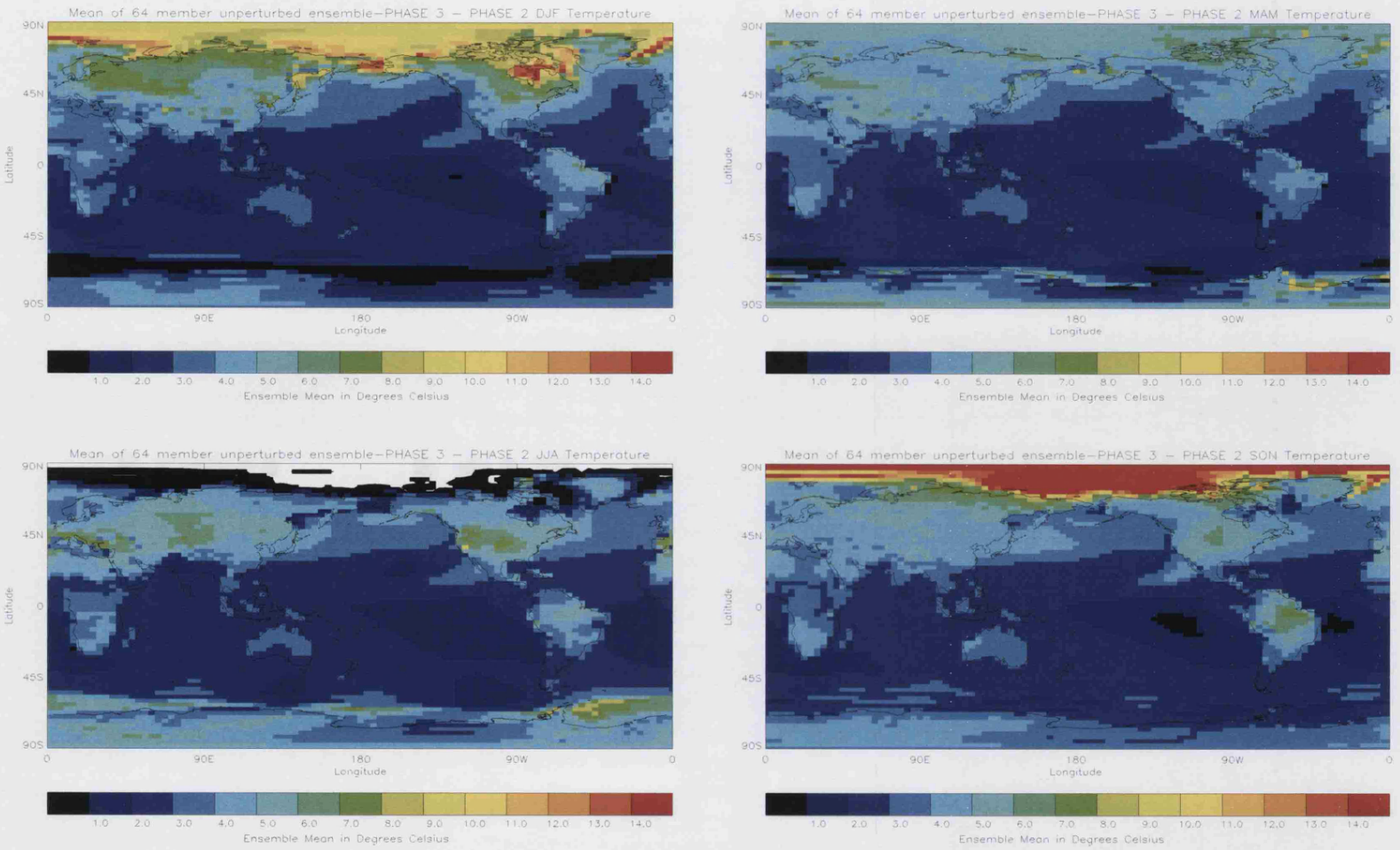


Figure 6.1: The standard HadSM3 model 64 member ICE mean for 8 year temperature change is shown for each season. Black areas show little or no cooling, white areas a cooling. Red areas show very high warming of over 9.5 degrees Celsius. Warming is strongly non–uniform and varies significantly with season.

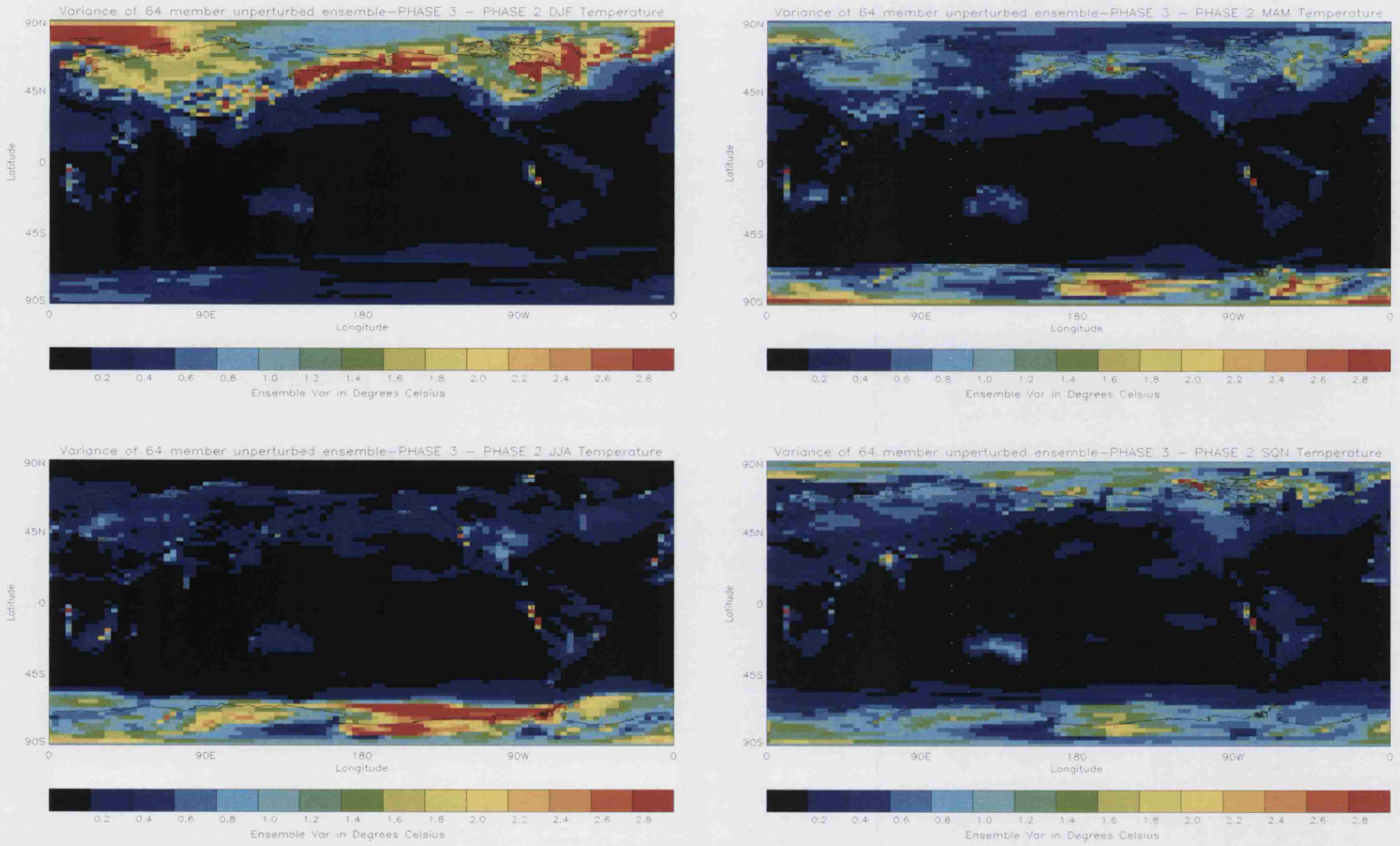


Figure 6.2: The variance in seasonal 8 year mean temperature change fields over the 64 member standard HadSM3 model ICE. Black areas show a variance of less than 0.2 degrees Celsius. Variance is typically higher over land – over 2 degrees Celsius in some cases.

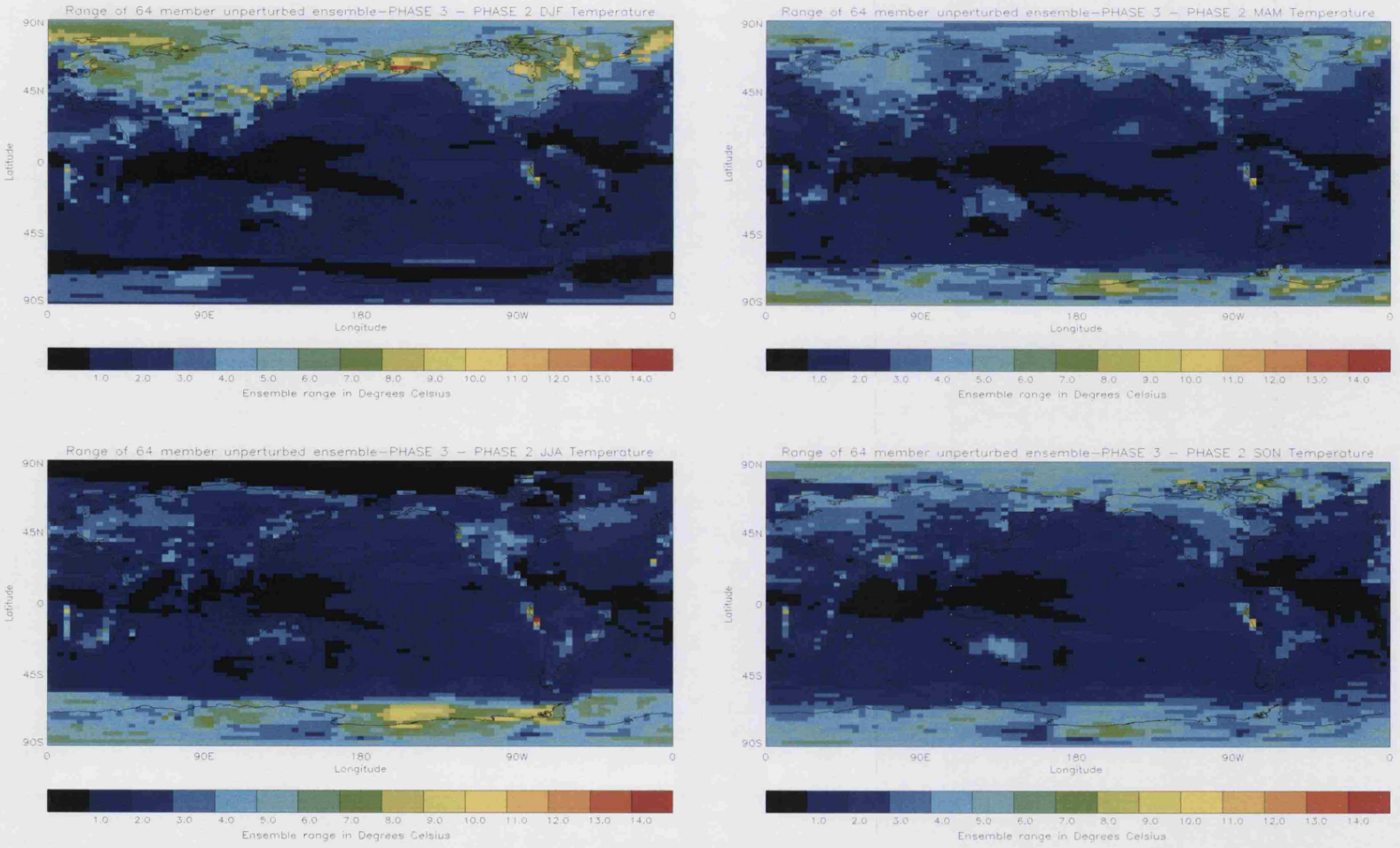


Figure 6.3: The range (maximum - minimum values) for temperature change over the 64 member ensemble. Areas in black indicate a spread of less than 1 degree Celsius over the whole ensemble. The range of seasonal temperature change within this ICE is over 10 degrees Celsius in some cases.

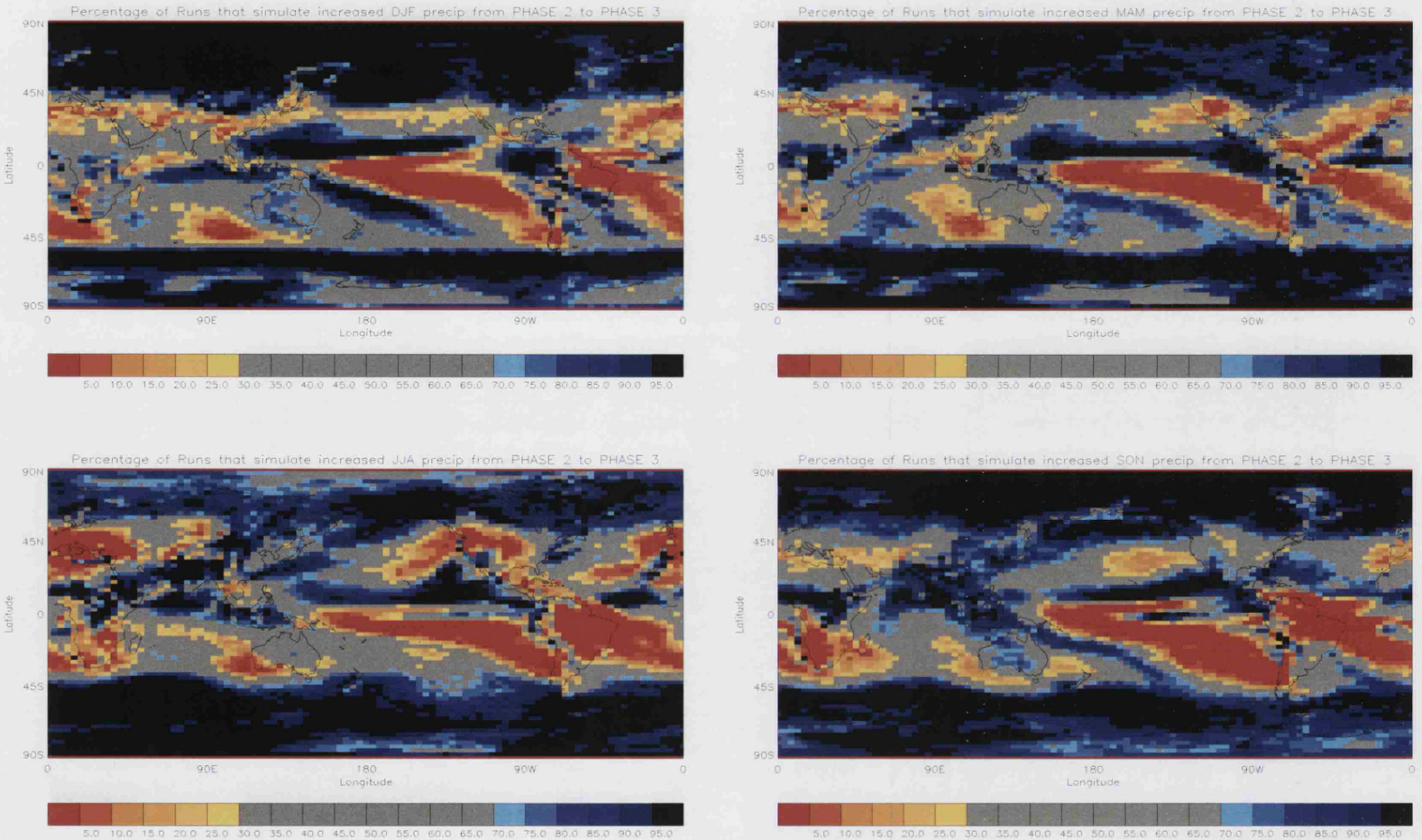


Figure 6.4: A democracy plot of precipitation change. The percentage of simulations (over the standard HadSM3 model ICE for which the 8 year seasonal precipitation increases from control to doubled CO_2. Areas in black (red) indicate that more than 95% of simulations show an increase (decrease) in precipitation. Grey areas indicate that the sign of precipitation change in the standard HadSM3 model is undetermined.

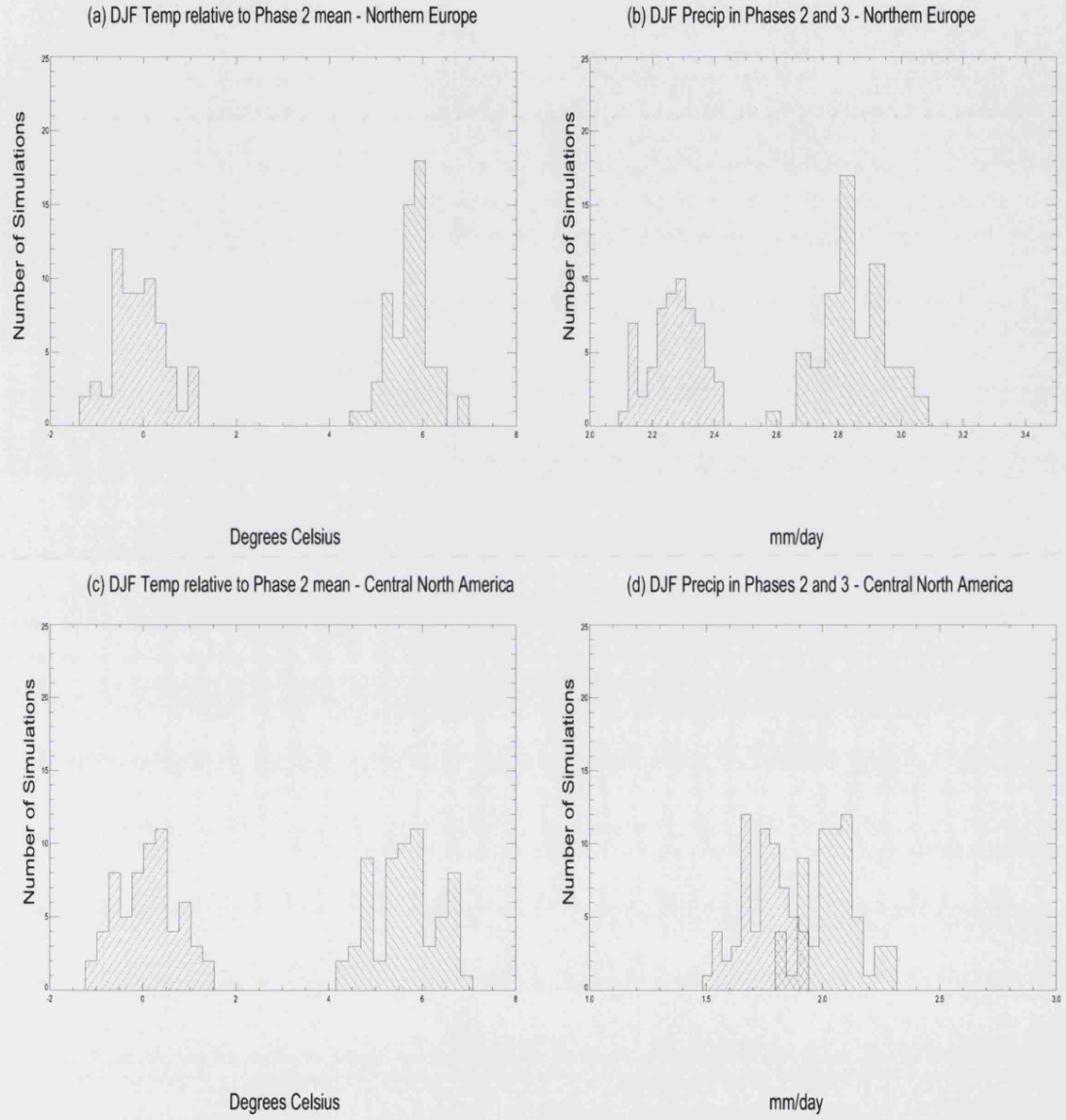
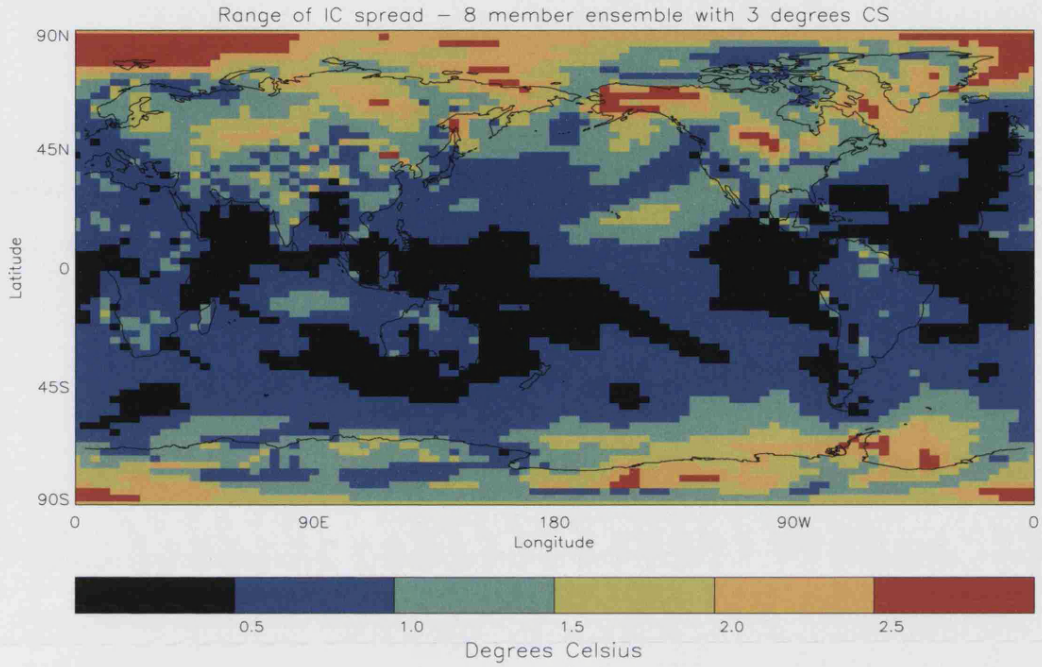


Figure 6.5: The distribution of 8 year means for the 64 members standard HadSM3 model ICE for (a) Northern Europe Temperature, (b) Northern Europe Precipitation, (c) Central North American Temperature and (d) Central North American Precipitation. The control phase is shown in green and the doubled CO_2 phase in red. The presence of an overlap indicates the sign of precipitation change is uncertain in the standard HadSM3 model.



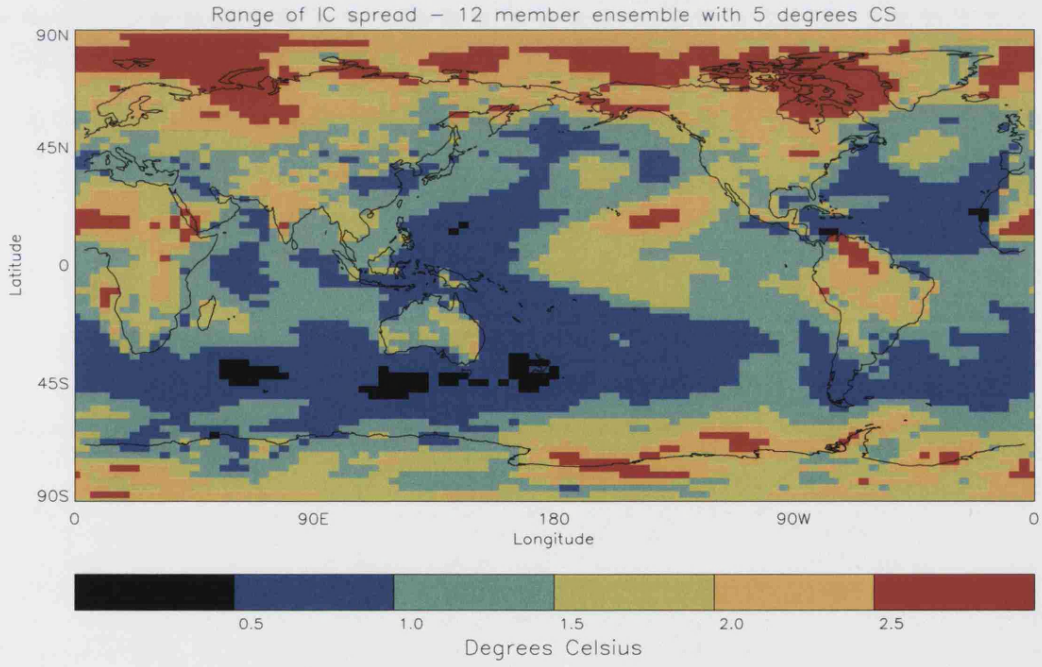
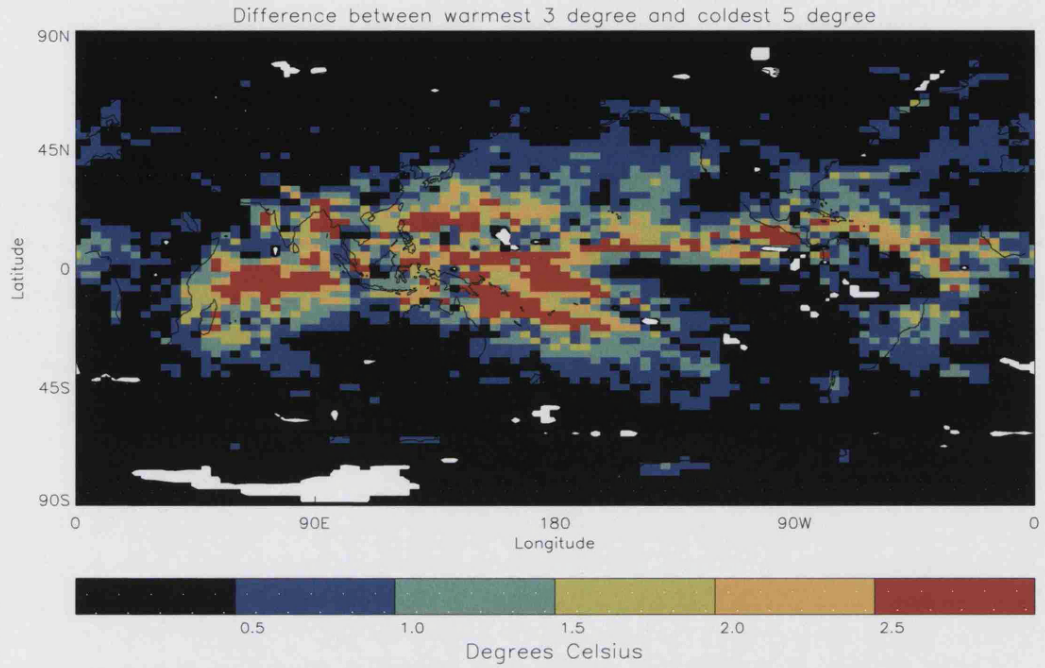


Figure 6.6: Range of 8 year mean temperature change under a doubling of CO_2 for 2 ICEs of 8 and 12 members and 3 and 5 degrees CS respectively. The magnitude of this internal variability is typically one degree Celsius, but can be over 2.5 degrees Celsius, particularly for the larger, 5 degree CS, ICE.



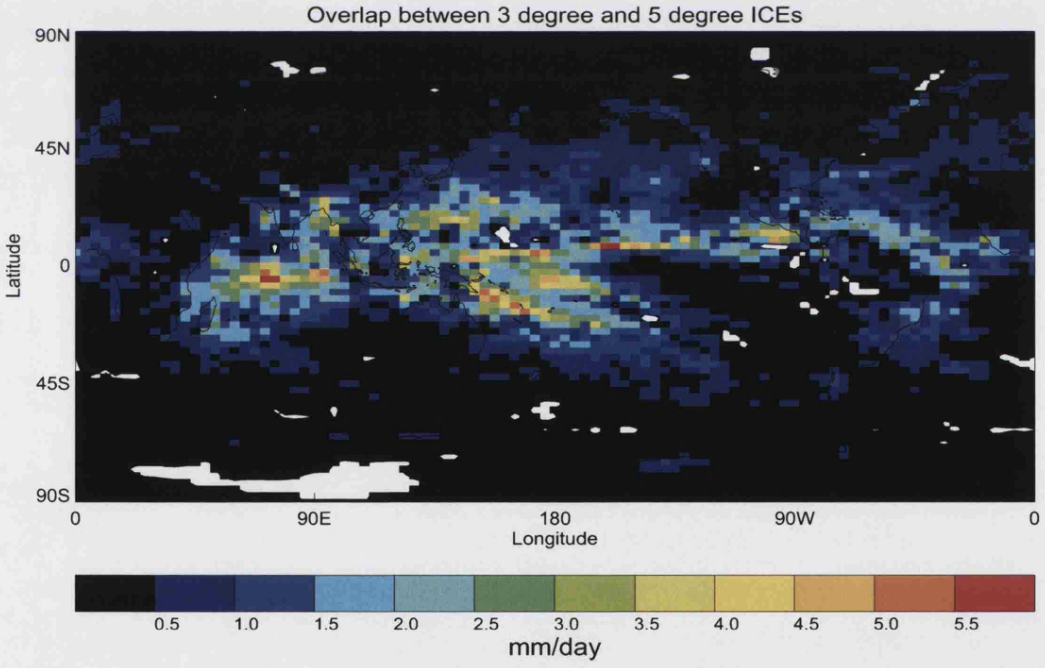


Figure 6.7: The difference in 8 year mean temperature/precipitation change under a doubling of $CO2$ between the maximum of an 8 member ICE with 3 degrees CS and the minimum of a 12 members ICE with 5 degree CS. The extent of this overlap is shown in temperature and precipitation. Positive values in temperature show where the maximum 3 degree simulation is hotter than the minimum 5 degree simulation. In precipitation, values denote the magnitude of the overlap between the driest (wettest) 3 degree simulations and wettest (driest) 5 degree, depending on the median direction of precipitation change form 3 to 5 degrees. Negative values denote areas with no overlap.

Chapter 7

Constraining New Results from the CPDN grand ensemble

7.1 Introduction

Ensembles of climate simulations can produce a wide range of behaviour Stainforth *et al.* (2005). New results from the largest ever climate modelling experiment are presented in this Chapter. A grand ensemble of 45644 simulations from the CPDN experiment Allen (1999); Stainforth *et al.* (2002) is analysed in this Chapter in terms of its uncertainties on global and regional scales. Similar ensembles of climate simulations have shown large variations in estimated CS Sanderson *et al.* (2007); Stainforth *et al.* (2005). Values of CS range from 0.9 to 16.4 degrees Celsius are shown in the data presented, the largest range so far generated by a GCM. Such a range contains simulations with very little change in GMST and those with catastrophic changes (over 20 degrees change in many regions). The presence of such large model diversity calls into question whether some some simulations are responding to increased CO_2 in an unphysical manner. Furthermore, the wide range of estimates for CS places difficulties on decisions that are sensitive to the amount of expected global warming.

Model diversity can reflect our uncertainty in the future, but only in a limited

way. Whilst it is possible to explore uncertainties using ensembles (as described in Chapter 2), it is not possible to fully explore all uncertainties in the case of climate predictions Smith (2002). Thus the diversity of model output places a lower bound on our uncertainty in the future Stainforth *et al.* (2007b). Because of our inability to fully explore uncertainties, model output must be communicated to decision–makers carefully so as to not give a misleading impression of confidence – "the end–user will always assume that the model spread provides some estimate of uncertainty" Knutti (2008a). A wide diversity of model output can encourage decision-makers to consider a wider range of possibilities, whereas low levels of model diversity might sharpen their prior beliefs and focus the decision–making process. The relevance that model diversity has for uncertainty in the future will depend on our confidence that climate models are able to simulate the future Stainforth *et al.* (2007b).

It is prudent to disregard simulations from decision-support in which we have no confidence in their ability to simulate future climate. Therefore, when considering this diversity of model behaviour, it is relevant to ask whether there is a systematic tendency for certain simulations to display unphysical behaviour that should be eliminated from decision–support. If model diversity can be reduced in a physically meaningful way, ensembles that initially produced a wide range of behaviour might be constrained to give robust results that can guide decision–makers towards particular courses of action.

Various tests can be devised that check climate model simulations for unphysical properties and can be used to exclude these unphysical simulations. These tests have been primarily based on observations of past climate Forest *et al.* (2000, 2007); Hegerl (2006); Piani *et al.* (2005); Sanderson *et al.* (2008) but can also be based on other qualities of physical coherence in the model simulations.

The use of methods used to constrain climate model ensembles is investigated in this Chapter in light of principles of statistical good practice. Prior to an analysis of three different methods for constraining model diversity, this Chapter looks at

the magnitude of uncertainties present in a grand ensemble of CPDN simulations. Constraining is understood here as attempting to reduce uncertainties by eliminating simulations[1]. Mathematically, simulations receive an effective weight of either zero or one. Whilst fractional weights have been attempted Murphy *et al.* (2004) the simpler, yet still challenging, approach of assigning weights of zero or one is used in this Chapter. The difficulties in constraining the range of results from climate ensembles are highlighted in this Chapter in the context of their physical rationale and statistical good practice.

The range of CS can be reduced by applying constraining procedures, although results may not be consistent across different methods. For example, it is shown in this Chapter that when using an observational constraint the distribution, and range, of CS depends on which variable is chosen as an observational constraint.

Three methods for constraining model simulations are carried out in this Chapter:

- *Constraining Parameter Values.* It is possible to change the distribution of CS by constraining the range of parameters used in the experimental design. This can be done by eliminating simulations with parameter values that are considered inadmissible on physical grounds. Of all the parameters perturbed in the CPDN experiment, the *Entrainment Coefficient* has been shown to have the greatest effect on CS in Knight *et al.* (2007); Sanderson *et al.* (2007). It could be that particular values of the Entrainment Coefficient used in the CPDN experiment are physically unjustifiable and therefore these simulations could be disregarded from decision–support. All parameter levels for the CPDN experiment were chosen by expert elicitation and there was no prior indication that the values chosen of the Entrainment Coefficient were any more or less realistic than other parameters. The selection of this particular parameter as a potential source of bias, only after seeing its effect on CS, is therefore

[1]It is interesting to note that previous analysis of model diversity has been focused on the problem of reducing model diversity and little attention has been paid to the question of how we might extend model diversity to reflect known uncertainties not yet explored by ensembles.

statistical bad practice. Despite these concerns, the distribution of CS for simulations with low, standard and high values of the Entrainment Coefficient are considered. Whilst the distribution of CS does indeed change when looking at simulations with a fixed Entrainment Coefficient, high CS simulations (over 8 degrees Celsius) exist for all three levels of this parameter.

- *Heat flux adjustment.* The HadSM3 model used in this experiment uses a Heat Flux Adjustment (HFA) (details have been presented in Chapter 5). It might be suggested that a global mean value of this HFA close to zero would be a desirable property of the model – the global mean HFA is sometimes forced to equal 0 in fiux–adjusted experiments Min *et al.* (2005). In particular, if the global mean HFA is close to zero this would suggest the atmosphere and ocean components of the model are in close equilibrium during the calibration phase. Furthermore, Chapter 5 established a relationship exists between HFA and CS; this might be used as the basis of a post-hoc constraint on CS. Such post–hoc filters are not recommended in this Thesis, rather this method is applied to make two points: **a)** It is statistical bad practice to choose a method for constraining CS *a posteriori* and **b)** The effect of ruling out simulations with a significantly non–zero global mean HFA is shown to have limited effect on the range of CS. Even when a stringent HFA filter is applied, simulations with high values of CS (over 8 degrees Celsius) are still admitted.

- *In–sample fit.* An important requirement for models to provide useful predictions is their ability to match the past, as discussed in Chapters 2 and 3. One approach to constraining model simulations could then be to disregard simulations that do not capture past observations within a certain level of accuracy. There are a number of different ways to do assess such in–sample fit. A simple criteria is set in this Chapter, using the in–sample RMSE error, as in Reichler & Kim (2008); Stainforth *et al.* (2005), compared to the observational fields for 1961–1990. This RMSE is used to constrain model

simulations in 7 different variables. The effects of constraining simulations on the distribution of CS are shown to depend on the variable used, demonstrating that the resultant distribution can depend on subjective choices. In particular, using a constraint in temperature tends to skew the distribution of CS towards high values and and precipitation towards low values.

In addition to the methods for constraining model simulations used above, a process of quality control is used to eliminate simulations that are not internally inconsistent e.g. simulations with a significant decadal drift in GMST during the control phase are ruled out. This process of quality control has been explained in Chapter 4. Whilst both quality control and constraining methods result in eliminating simulations from further analysis their aims are different. There is an important distinction between quality control and the constraining methods analysed in this Chapter. Quality control seeks to purge internally inconsistent simulations whereas post–hoc constraining methods seek specifically to reduce model diversity.

The range of model behaviour seen, after any suitable constraining procedures have been applied provides an estimate of the uncertainty present in model projections. The range of resultant model diversity provides a test for consistency of information, as described in Chapter 2 (Model diversity test). It is shown that model diversity is large in the case of the CPDN PPE of parametrically perturbed versions of the HadSM3 model.

In general, a larger ensemble will produce a wider range of CS. Similarly, a method that rules out many simulations is likely, on average, to constrain the range of CS by more than a method that rules out only a few due to simple counting statistics, regardless of the physical basis for this reduction in uncertainty. Since a very large ensemble is analysed in this Chapter high CS values are still admitted for many methods, even once constraining procedures are applied. The analysis presented in this Chapter shows that high CS simulations can remain, after a wide range of constraining methods have been applied. As more simulations are run, it should

be expected that the range of values of CS will increase. Since large ensembles will explore the tails of the distribution of CS it is expected that very high values of CS will be seen, even if this occurs only seldom.

The layout of this Chapter is as follows: Section 7.2 introduces the data set analysed, Section 7.3 looks at the distribution of CS found in the data. Section 7.4 examines the range of behaviour seen on regional scales; these uncertainties are often larger than on the global scale and vary with region. Section 7.5 applies three methods of constraining that attempt to reduce the uncertainties in estimated CS. The consequences of the results are discussed and conclusions given in Section 7.6.

7.2 The Data Set

The first CPDN experiment aims to understand the effect of doubling CO_2 concentrations on key variables such as GMST and has produced the largest set of climate simulations to date. For more details on CPDN experimental design see Chapter 4. Analysis of data from the CPDN experiment has previously been carried out by Knight *et al.* (2007); Knutti *et al.* (2006); Piani *et al.* (2005); Sanderson *et al.* (2007); Stainforth *et al.* (2005, 2007a,b). In this Chapter, data from the first 45644 simulations are analysed. For some simulations, only a sub–set of data is available i.e. there is data missing for some simulations. Rather than throwing away these simulations completely, analysis is carried out in each case on as many simulations as possible. Applying this principle results in a different number of simulations being used for different analyses; 31818 of these simulations had adequate data to calculate CS, and 22711 simulations passed quality control (22723 had a full time series global mean temperature).

7.3 Climate Sensitivity

CS is a key statistic in understanding climate change over the 20th and 21st centuries and has received much attention in the literature Annan & Hargreaves (2006); Annan *et al.* (2006); Forest *et al.* (2002); Frame *et al.* (2005); Gregory *et al.* (2002); Roe & Baker (2007); Schwartz *et al.* (2007); Solomon *et al.* (2007a)[1]. The IPCC Fourth Assessment Report said that CS is "likely to be in the range 2 to 4.5 degrees Celsius with a best estimate of about 3 degrees, and is very unlikely to be less than 1.5 degrees. Values substantially higher than 4.5 degrees cannot be excluded, but agreement of models with observations is not as good for those values."[2]. There have been studies showing a wider range of values of CS; using simple climate models Andronova & Schlesinger (2001), and using GCMs Stainforth *et al.* (2005), that have presented estimates of CS as high as 10 and 12 degrees respectively. This Section presents results from the CPDN experiment that show a range of ICE mean CS from 0.9 to 16.4 degrees. The implication of such a wide range of CS is discussed in this Chapter as well as three possible methods to constrain this range.

CS is an equilibrium statistic calculated from a time series of GMST values. This time series has not always reached an equilibrium by the end of the final doubled CO_2 phase of the experiment. This can be seen in Figure 7.1, which shows the time series of GMST quality controlled simulations over the three phases of the experiment. During the calibration and control phases (with constant, pre–industrial CO_2 concentrations) the distribution of GMST is fairly stable, with slightly more variability in the control phase. The effect of instantaneously doubling CO_2 at the beginning of the third phase is characterised by a rapid and sustained warming in most simulations. Some simulations do not have a smooth time series. In some cases, like the notable dip at years 38-40, this is due to an unusually cold month

[1]Climate sensitivity has also been studied without the use of climate models. Morgan & Keith (1995) showed that there can be disagreement amongst climate scientists' subjective estimates of CS.

[2]The terms "likely" and "very likely" are interpreted as corresponding to a probability of greater than 66% and 90% respectively in the IPCC AR4 Solomon *et al.* (2007a)

skewing the annual mean time series. These features could be a result of the model's internal variability or a numerical error. Whilst quality control explicitly picks up simulations with seasonal jumps of over 20 degrees Celsius, some simulations with instabilities on shorted timescales slip through.

The distribution of CS is important for understanding the likelihood of various levels of warming. Figure 7.2 shows the range of CS for all simulations with quality control (panel (a)), without quality control (panel (b)) and for ICE means (panel (c)). The application of quality control reduces the number of simulations available from 31818 to 22711 but does not appear to significantly alter the shape of the distribution. Quality control does not appear to qualitatively change the distribution, in particular a wide range of CS values is admitted both with and without quality control. The similarity of the three distributions of CS is shown in panel (d) of Figure 7.2. Quality control seems have little effect on the distribution of CS, nor does the taking of model version ICE means.

There are three notable features of the distributions shown in Figure 7.2;

1. The peak is around 3.5 degrees for all three distributions of CS, close to the value of CS for the standard HadSM3 model (about 3.4 degrees Celsius, as stated in Chapter 6) and consistent with estimates from other state–of–the–art models Solomon *et al.* (2007a). This suggests that the effect of simultaneous parameter perturbations often cancels out (or the effect of perturbing some parameters is small) for GMST change (this need not be the case for other variables and length scales), often resulting in values of CS close to the standard model version.

2. The distribution is fairly smooth, with no obvious discontinuities or multi–modality. There is a wide range of values, especially at the high end. Rather than a few isolated high values, there is a smooth tail, with a monotonically decreasing number of very high CS simulations.

3. There are no negative values of CS. Despite having 22711 quality controlled simulations available, using 5983 different sets of parameter values, and exploring the range of possible high CS in great detail, the lowest CS found was 0.9 degrees Celsius. This result has important consequences for understanding the nature of feedbacks in the model – it seems very difficult for the HadSM3 model to exhibit cooling behaviour under a doubling of CO_2 in any simulation with even a very large sample size and a wide range of parameter values. In the absence of feedback effects, the temperature rise resulting from the radiative forcing equivalent to doubling CO_2 has been estimated to be $\sim$1 degree Celsius Colman (2003); Roe & Baker (2007); Solomon *et al.* (2007a). The lowest values of CS are close to this "zero feedback" level, implying that feedbacks in the HadSM3 model are substantially positive for a wide range of parameter values.

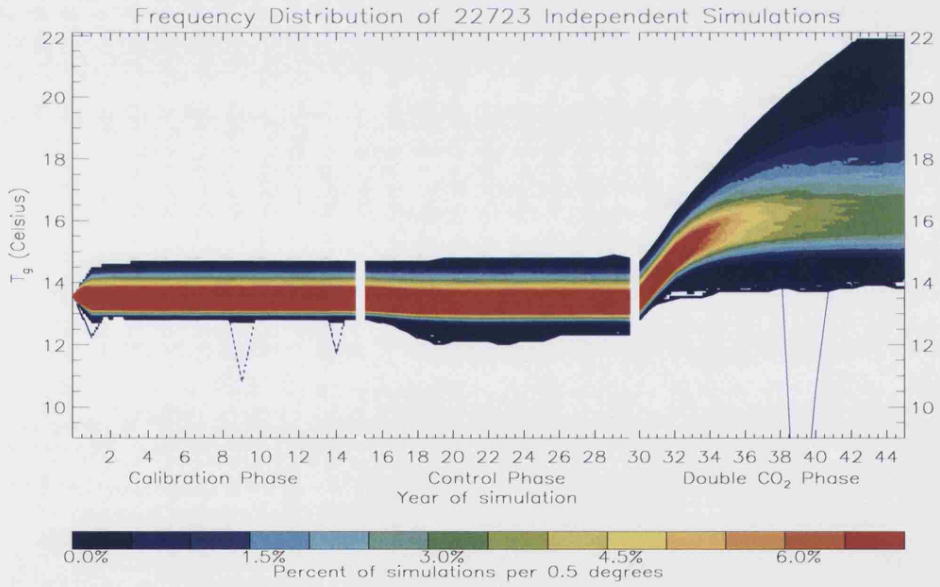


Figure 7.1: The time series of model version mean (averaged over available quality controlled ICE members) for the three phases of the experiment. Most simulations warm rapidly in the final phase, some by over 8 degrees by the end of the 15 year doubled CO_2 phase. There are some simulations with unsmooth trajectories.

In order to understand the effect of parameter perturbation on model GMST response to increased CO_2 it is possible to compare the time series of GMST from

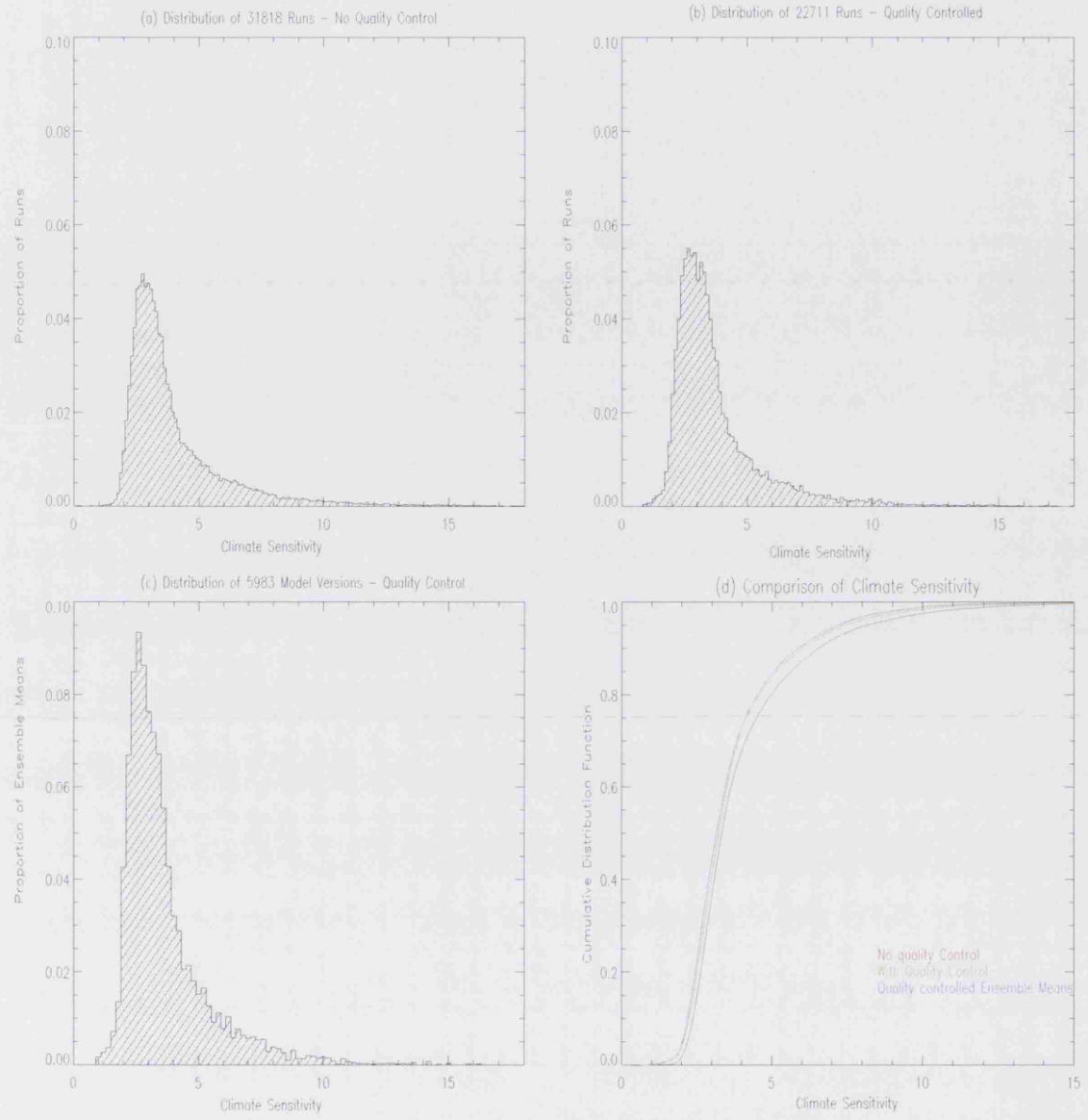


Figure 7.2: The distribution of CS in the CPDN PPE. Panel (a) shows the distribution of all simulations, panel (b) the distribution of quality controlled simulations. Panel (c) shows the ICE mean over all model versions, for quality controlled simulations. Panel (d) shows a comparison of the three different distributions as CDFs. The highest model version mean CS is 16.4 degrees Celsius.

the parametrically perturbed CPDN experiment to other slab–model equilibrium experiments made with different structural models. It is shown here that the large, parameter perturbed CPDN ensemble produces a much wider range of behaviour than different structural models. It should be noted that it is easier to create a large parametrically perturbed ensemble than a new structural model since only

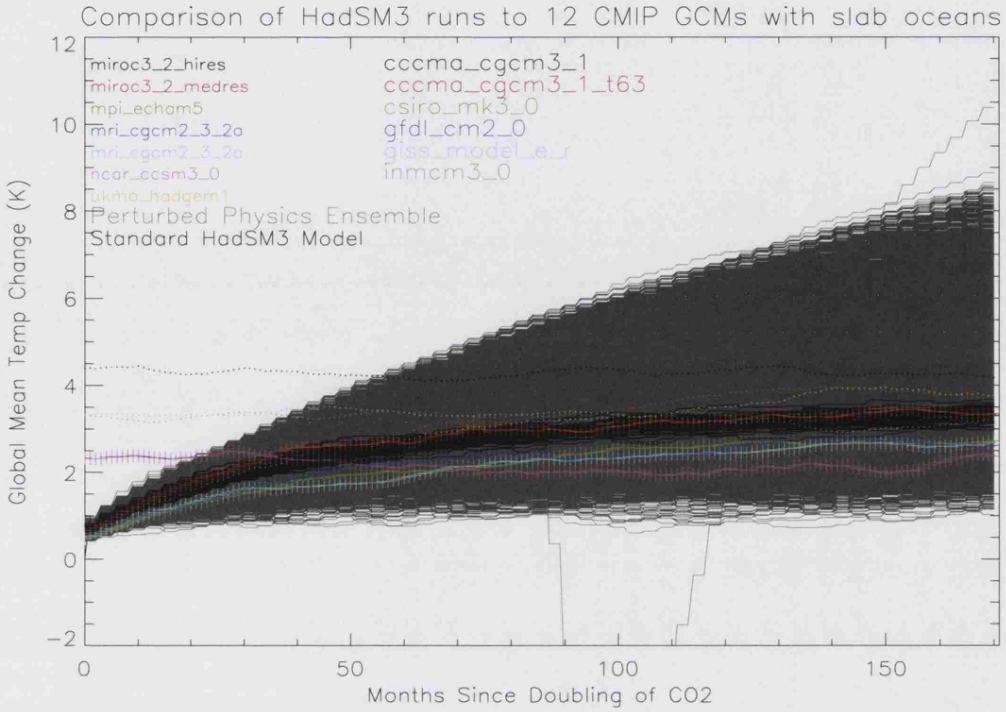


Figure 7.3: The change in temperature following a doubling of CO_2 is shown for 12 CMIP II models and the CPDN ensemble. For some CMIP simulations, data pertaining to the transient period of warming immediately following a doubling of CO_2 was not available.

parameter values need to be changed, rather than a re–working of the model's physical structure.

Figure 7.3 shows the doubled CO_2 phase of the CPDN PPE (grey). Also shown is an ensemble of simulations from the HadSM3 Standard model (black), also from the CPDN experiment, and comparable slab model simulations used in the Second Coupled Model Intercomparison Project (CMIP) Covey *et al.* (2003), in colour. This graph shows that all models and parameter values warm under a doubling of CO_2. The range of warming across the Standard HadSM3 model is approximately half a degree, compared to over 2 degrees across different structural models used in CMIP. For some of the CMIP models data for the transient part of the simulations is not available (e.g. the black dotted line for the **miroc3-2-hires** model) – Figure 7.3 shows only the available times series after simulations have reached an equilibrium. The CPDN PPE shows a far greater range of warming – from 1 degree to over

8 degrees over the 15 year period. Note that there is one simulation that dips significantly below the range of the other models and proceeds to warm rapidly, showing the greatest amount of warming by the end of the phase. This is an example of a simulation that passes quality control but that one might not wish to consider useful for decision support. This simulation appears to be unstable; it begins to warm, then cools, then warms up rapidly (over 15 degrees in 6 years). Whilst simulations can not be ruled out simply for producing unexpected results, it is possible that the cause of such erratic behaviour is unphysical and thus might be more relevant for model developers than decision makers.

The set of simulations shown in black are generated from a 64 member ICE using the HadSM3 standard model. It is clear that the range seen within this ICE is less than across different structural models. The range of warming shown within the standard HadSM3 model ICE shown is $\sim$0.5 degrees Celsius, compared to $\sim$2 degrees across different structural models and almost 8 degrees across the CPDN PPE. These results show that changing parameter values can have a more significant effect on model behaviour than changing ICs and that parameter perturbation has explored a wider range of model behaviour than different structural models to date. The response to doubling CO_2 is much more varied across the CPDN PPE than the CMIP multi–model ensemble.

7.4 Sub–global Behaviour

Section 7.3 has looked at the CS; this Section now looks at the sub–global responses simulated in the CPDN PPE. It is important to consider model behaviour on sub–global length scales since the response to increasing CO_2 varies with region and can not necessarily be inferred directly from global means Smith *et al.* (2008). Local climate changes are particularly relevant to impact studies and adaptation measures, which can require information on the length scales as a model grid–box or finer e.g. assessing the impact on climate change on downscaled precipitation over the River

Thames Wilby & Harris (2006). The range of model behaviour is shown to be large for temperature and precipitation on grid–box length scales, but more so at the grid–box level.

Figure 7.4 reflects the spatial variations in temperature change for each grid box and the associated variance. Figure 7.4 shows the mean (upper panel) and variance (lower panel) of the change in 8 year annual mean 1.5 metre surface temperature between the control and doubled CO_2 phases for 22698 quality controlled simulations[1].

The ocean typically warms by 1–3 degrees Celsius (shown in dark blue), the centre of large land masses by around 4–6 degrees Celsius (shown in green and yellow, although lower in Central Africa) and the Arctic by up to 8 degrees. The variance in ocean warming is around 1–2 degrees Celsius. Note that the variance in land warming is higher than over the ocean – typically about about 2-4 degrees Celsius with significant regional differences; the Amazon region has a particularly high variance of about 6–7 degrees Celsius. There is also high variance in the Arctic region where the most extreme warming is simulated.

Figure 7.5 shows the mean, democracy plot and extremal fields for change in 8 year annual mean precipitation between the calibration and the doubled CO_2 phases for 22698 quality controlled simulations. Mean changes are calculated as the percentage change between the ensemble mean precipitation in the last 8 years of the doubled CO_2 phase and the ensemble mean precipitation in the last 8 years of the control phase. Percentage change is used here rather than the absolute change since it is felt to be more intuitive and is consistent with common practice Solomon *et al.* (2007a). The use of percentage change can mean that in areas with very low precipitation in the control phase, small absolute changes appear as large relative changes.

The mean precipitation difference shows that, for large areas, there is less than 20%

[1]22698 simulations are used in this Section in contrast to the 22711 quality–controlled simulations used in Section 7.3 since 13 simulations have requisite data for the calculation of CS but not for producing the required regional fields. Similar variations exist elsewhere in this Thesis due to a corresponding effect.

change in precipitation. In other areas, mostly the high latitudes precipitation is simulated, on averaged, to increase by 20% or more. The democracy plot shows the percentage of simulations that simulate an increase in precipitation. Areas in black, occurring mostly in high latitudes (50 degrees or higher), show that over 95% of simulations indicate an increase in precipitation. Conversely, areas in red show that less than 5% of simulations simulate increased precipitation. Areas are coloured in white where between 40 and 60% of simulation show increased precipitation i.e. the ensemble is not consistent in simulating the change of sign of precipitation. It is not clear what level of agreement should be taken as a robust signal, but it is apparent that large parts of the USA are not robustly simulated by this model – roughly half the ensemble simulates more precipitation, and half less. Furthermore, these annual means do not help understand how precipitation events will change e.g. "no change" annually might mean half the number of rainy days with twice the frequency or many other types of behaviour.

In order to better understand the uncertainty in simulated precipitation change, the minimum and maximum percentage difference between the control phase and the doubled CO_2 phase at each grid box is plotted in Figure 7.5. These plots show that, at every grid box, some simulations show at least a 40% drop in precipitation and others at least a 20% increase in precipitation; there is no grid box for which every simulation is either wetter or drier. Looking at the Middle East, simulations indicate a range of precipitation changes from minus 80% (over a five–fold reduction) to over 300% increase. The wide range of model behaviour means that the HadSM3 model can not guide impact studies reliant on regional precipitation information but does indicate the need to make flexible decisions and monitor climate responses carefully since large changes in precipitation are possible.

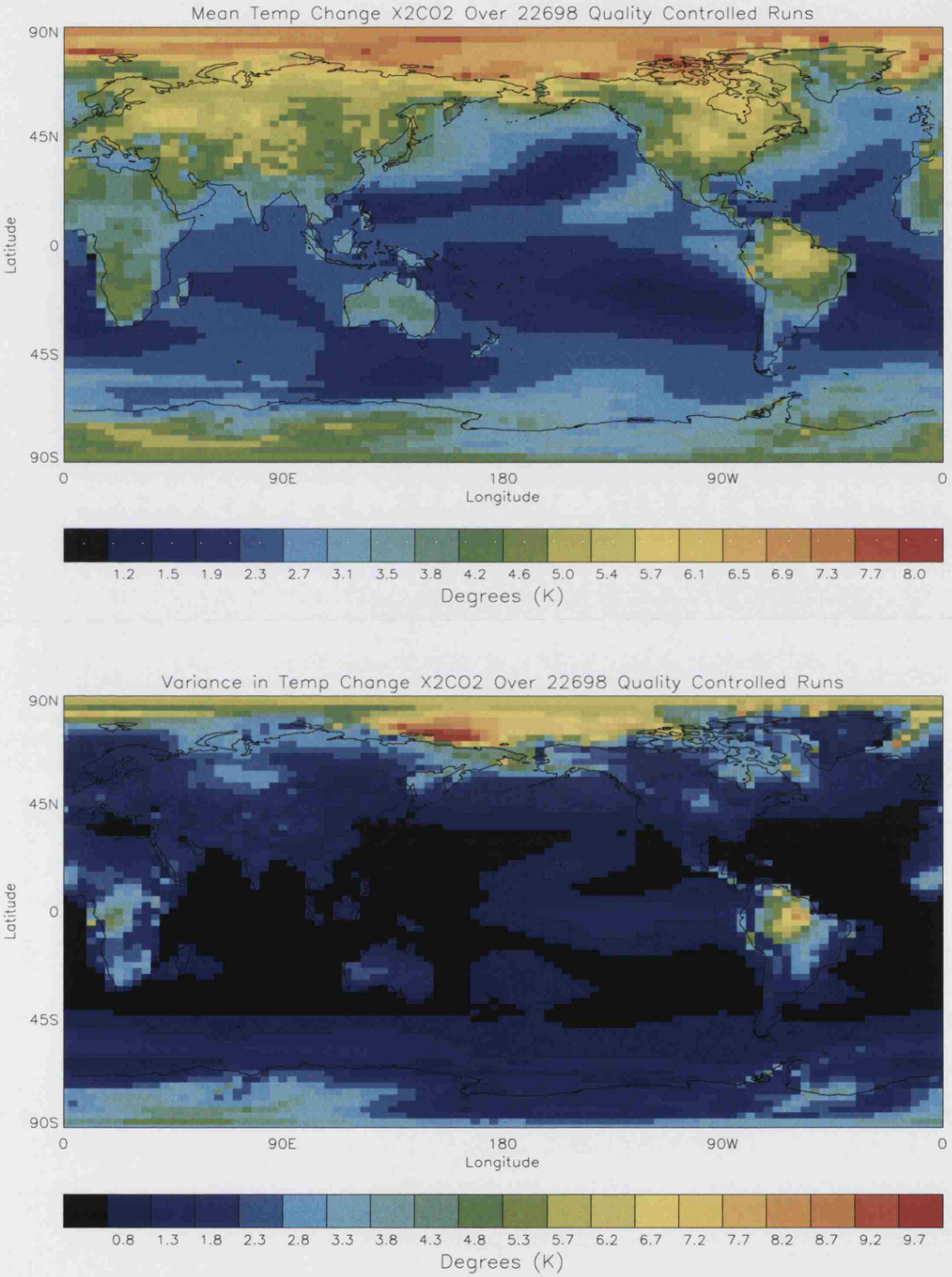


Figure 7.4: The mean (upper panel) and variance (lower panel) of 8–year mean annual mean temperature change between the pre–industrial CO_2 calibration phase and the doubled CO_2 phase over 22698 simulations. Warming is greater in the centre of large masses and in the Northern high latitudes. Warming over the ocean is typically between 1 and 3 degrees Celsius, compared to 6 to 8 degrees in the Arctic.

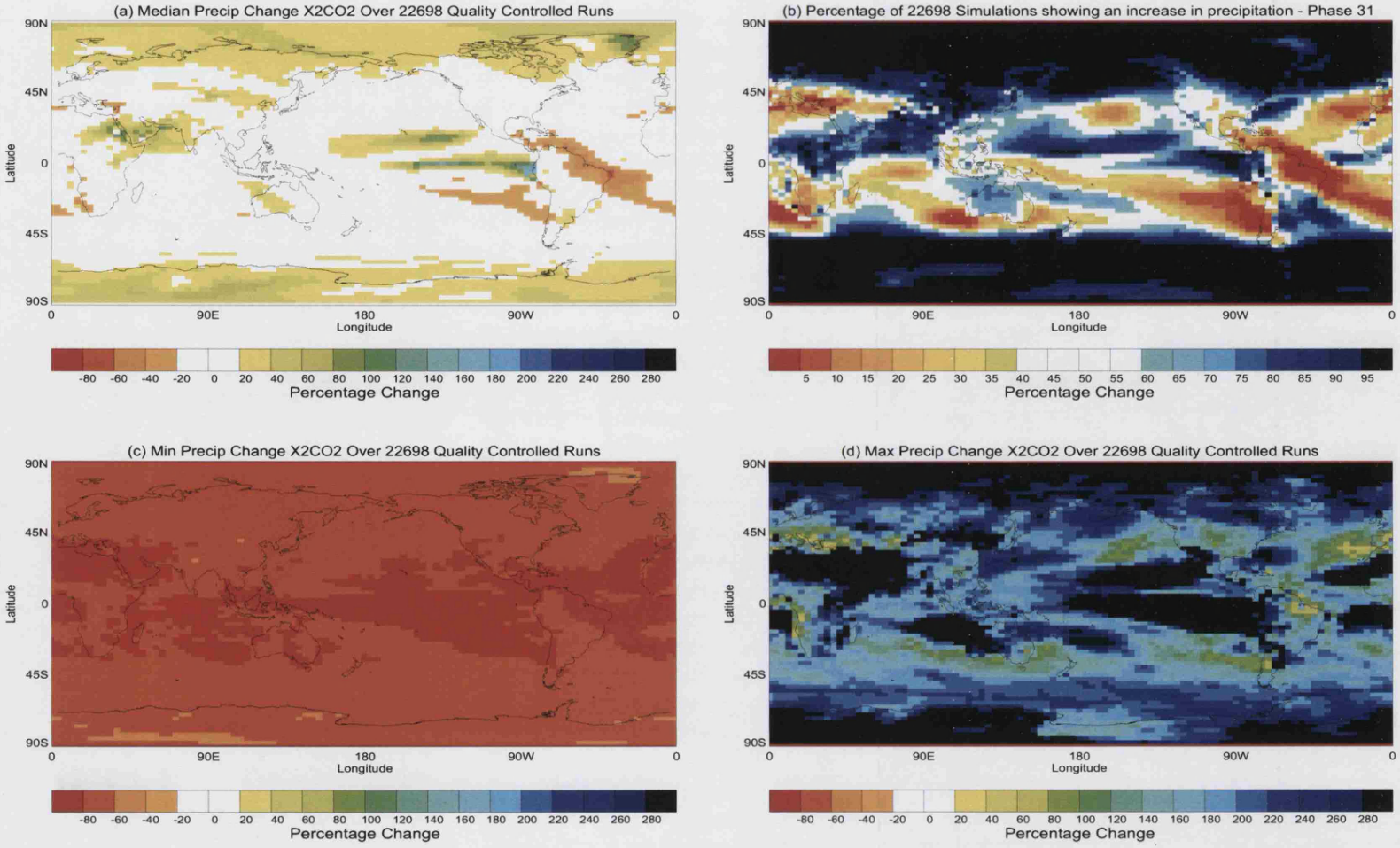


Figure 7.5: The mean (a), democracy plot (b) and bounding box of 8–year (the minimum is shown in panel (c) and the maximum in panel (d))mean precipitation change between the pre–industrial CO_2 calibration phase and the doubled CO_2 phase over 22698 simulations. The democracy plot shows the percentage of simulations with an increase in precipitation at each grid box.

The range of behaviour across the CPDN grand ensemble is now examined in terms of the spatial pattern of temperature change in two extremal simulations. In Figure 7.6 the DJF and JJA seasonal mean temperature change is shown for each of two simulations with 1.2 and 16.9 degrees CS. It is notable that in the 1.2 degree simulation, the centre of large land masses warms by up to 8 degrees Celsius in the JJA season and large parts of the Arctic by 6–10 degrees Celsius in the DJF season. For the 16.9 degree simulation, large areas warm by over 20 degrees Celsius, particularly in the Northern high latitudes and the centre of large land masses. There are some areas where regional warming is far less than might be expected from the global mean; there are land areas with less than 5 degrees of warming in both seasons. It is important to consider the different regional responses when planning adaptation strategies e.g. if the 1.2 degree simulation, with a modest magnitude of global warming, is to be believed, some regions might still expect to experience 8 degrees of warming. Such information is vital for impact assessment and decision makers.

The magnitude of regional warming seen across the PPE is related to the amount of global warming. If CS were to be known to within a degree Celsius, or less (few studies afford such small uncertainty on CS even when heavily constrained by observations), the range of regional changes might be reduced. The question of the relevance of global means for regional impact analysis analysed with in detail in Chapter 8.

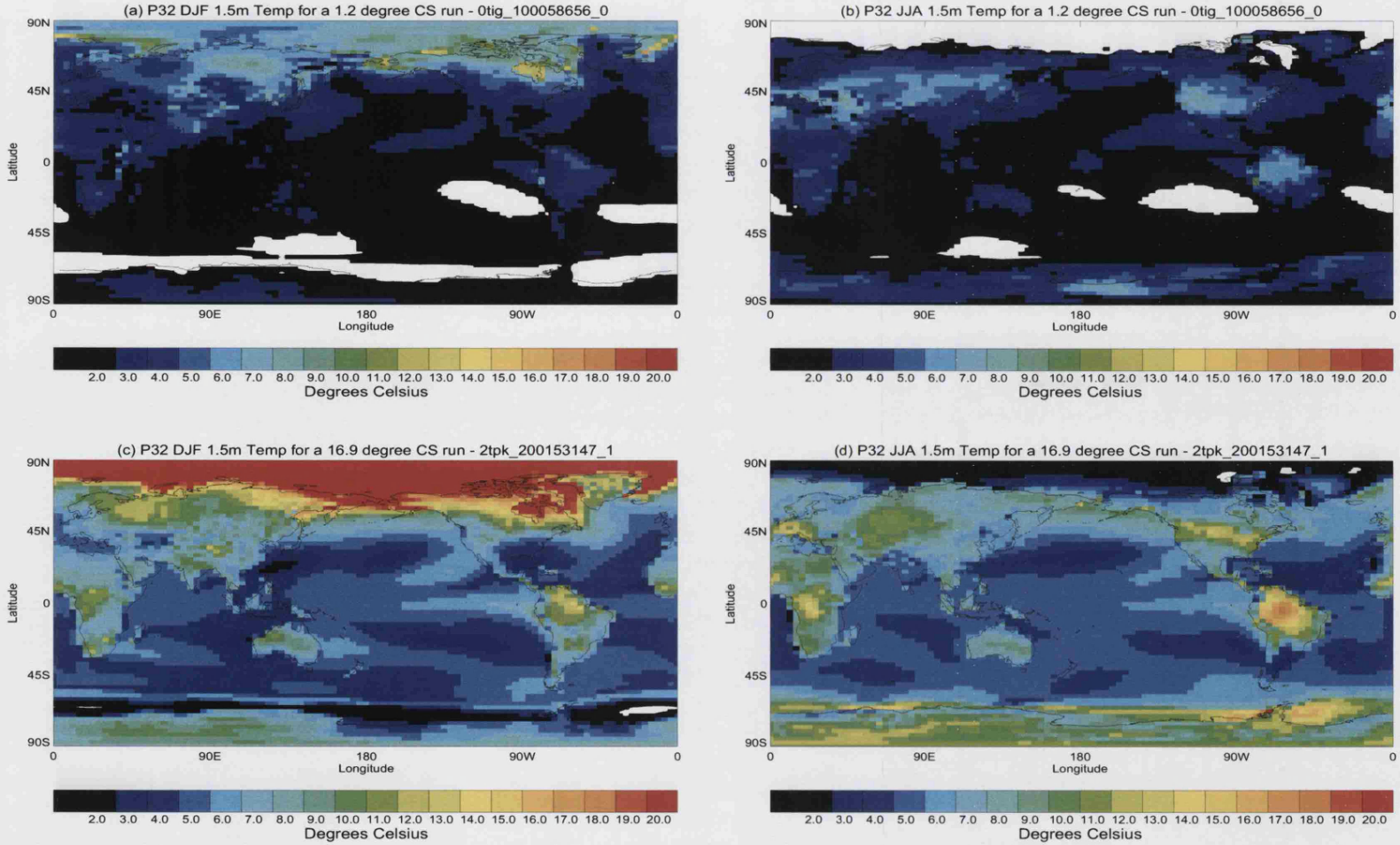


Figure 7.6: The change in temperature from phase 2 to phase 3 is shown for two simulations. These simulations were selected on the basis of having very low and very high climate sensitivities of 1.2 and 16.9 degrees, respectively. Panel (a) shows the DJF change for the 1.2 degree CS simulation and panel (b) the JJA change. Panel (c) shows the DJF change for the 16.9 degree CS simulation and (d) the JJA change.

7.5 Constraining Model Simulations

This Section looks at three methods that can be used to constrain model diversity in the CPDN PPE. The range of temperature change across this data set has been shown to be very large in Sections 7.3 and 7.4 (see Figure 7.2). In fact, the full range of behaviour is so large that, in the absence of constraining methods (such as Annan *et al.* (2006); Forest *et al.* (2007); Hegerl (2006); Knutti *et al.* (2006); Piani *et al.* (2005)), this set of simulations can not guide decision making reliant on regional (and even global) information towards specific courses of action. This Section applies three constraining methods using **a)** the Entrainment Coefficient, **b)** global mean HFA and **c)** in–sample fit. For methods **b)** and **c)**, a 64 member ICE from the Standard HadSM3 model is used to set a benchmark for model performance. This benchmark is explained in Section 7.5.3.

These three methods are now applied to the quality controlled CPDN PPE in order to attempt to assess whether the behaviour in the CPDN PPE should be considered over–dispersed. These methods are not expected to represent the full range of possible constraining procedures and alternative methods are likely to show different results. The three methods presented and applied here are rather used to discuss some of the potential difficulties faced when attempting to evaluate the model diversity climate model simulations in light of statistical good practice.

It is argued that methods that are used to change the model diversity of an ensemble should be based on **1)** physically meaningful grounds and **2)** statistical good practice i.e. the investigation and design of physical tests should not be aimed at achieving a particular result, such as a reduction of model diversity, a posteriori.

7.5.1 Constraining using the Entrainment Coefficient

The first method used to attempt to constrain CS uses the Entrainment Coefficient (a parameter that describes how quickly a convective cloud mixes in clear air – see Section 4.4.3 in Chapter 4). The value of the Entrainment Coefficient is rele-

vant to important cloud feedbacks and has been shown to be the parameter whose perturbation has the greatest effect on CS in the CPDN experiment Knight *et al.* (2007); Sanderson *et al.* (2007, 2008). The distribution of CS is examined under the three values of this parameter – where the Entrainment Coefficient is at its low (0.6), standard (3, the unperturbed value) or high (9) values. Comparing these distributions enables an examination of how changing the Entrainment Coefficient alters the distribution of CS.

It is shown in Figure 7.7 that the Entrainment Coefficient has a significant qualitative effect on the distribution of CS. Figure 7.7 shows that the distribution of simulations with a low Entrainment Coefficient is highly skewed, with a significant number of simulations having CS above 10 degrees and some over 14 degrees Celsius. For the "standard" value of the Entrainment Coefficient (unperturbed from its standard value) or the high value of this parameter, the tail reduces considerably. The distributions for all three values of the Entrainment Coefficient centres about 3.5 degrees, but the distribution of CS has a significantly heavier tail above 8 degrees where the Entrainment Coefficient is low. Despite this, there are simulations with a high value of the Entrainment Coefficient with CS over 8 degrees. Where the Entrainment Coefficient is at its standard or high values the distribution of CS appears more symmetric and less right skewed than where the Entrainment Coefficient is low. This suggests that the HadSM3 model has a propensity to produce fewer high CS simulations for the low value of the Entrainment Coefficient.

It has been disputed how the shape of the distribution of CS should be interpreted – the distribution has been shown to depend on subjective choice of prior distributions Frame *et al.* (2005) and that recent warming does not provide a strong constraint on CS Allen *et al.* (2006). Furthermore, it has been argued in more detail that the right skewed shape is an unavoidable feature of the nature of CS Roe & Baker (2007). It is clear in this case that if the experimental design had been different e.g. with no perturbation to the Entrainment Coefficient, a very different distribution

of CS would have been obtained.

Constraining model output using the Entrainment Coefficient has been shown to significantly alter the distribution of CS. In particular, ignoring simulations with a low Entrainment Coefficient results in eliminating most simulations with a high CS. This method for constraining model diversity does not meet the requirements of statistical good practice since the choice of parameter and the values deemed unrealistic were cherry–picked in order to rule out most simulations with CS above 8 degrees Celsius. If *ad hoc* methods are permitted, a similar constraining procedure could be adopted that rules out simulations with low values of CS and exagerrating the right–skewed distribution. If simulations with a low Entrainment Coefficient are to be eliminated the justification for doing so must be physically based (e.g. if these model simulations do not re–produce key properties of the seasonal cycle Knutti *et al.* (2006)) and this test must be carried out in light of good statistical practice i.e. not designed to produce some desired effect such as eliminating simulations with high values of CS.

In contrast to the results presented here, Roe & Baker (2007) presents a reason why CS should have a characteristically right–skewed distribution of a particular analytic form. Here these claims are refuted using evidence from the CPDN PPE for the first time. In Roe & Baker (2007) the equilibrium change in GMST in response to a radiative forcing is taken to be $\frac{T_0}{1-f}$, where T_0 is the reference CS. Reference CS is the temperature response in the absence of any feedbacks. In the equation above, f is the feedback parameter. The authors then explain, assuming feedbacks approximately follow a Gaussian distribution ("Although the general features of our results do not depend on this assumption, it facilitates our analysis" Roe & Baker (2007)), why CS is right skewed. This right-skewness is claimed to be a direct result of the non–linear transformation of Gaussian feedbacks to calculate CS. The fact that the distribution of CS will be skewed in an inverse–Gaussian manner assuming observational errors are Gaussian was also made in Piani *et al.* (2005). The claim that CS

is inherently right–skewed due to its non–linear relationship to Gaussian feedbacks can be examined using data from the CPDN grand ensemble. This is done by taking three sets of simulations from the CPDN experiment (Entrainment Coefficient equal to its low, standard and high values). The distribution of CS should be largely similar under different parameter values ("The basic shape...is not an artifact of the analyses or choice of model parameters. It is an inevitable consequence of a system in which the net feedbacks are substantially positive." Roe & Baker (2007)). From the three empirical distributions of CS, the implied distribution of feedbacks was calculated. Roe & Baker (2007) suggest that the distribution of feedbacks can be assumed to be Gaussian without introducing any critical biases. The distributions of CS and the feedback parameter when holding the Entrainment Coefficient fixed are shown in Figures 7.7 and 7.8. The distributions of CS, shown in Figure 7.7, do indicate a long tail, although the shape of the distribution itself (and the skewness of the distribution) is highly dependent on the choice of parameter values. Due to the skewed distribution of feedbacks, a small number of negative feedbacks, shown in Figure 7.8 are simulated in some simulations.

The distribution of feedbacks, shown in Figure 7.8, appears approximately Gaussian in the case of a high Entrainment Coefficient, less so for a standard value and highly left skewed for a low value. Where such highly skewed feedback distributions are possible, the assumption of Gaussianity may no longer be tenable. With highly skewed feedback distributions, the distribution of CS need not have a long tail towards high values of CS, as was argued to be the case in Roe & Baker (2007), but can take any other distribution.

It is concluded here that important features of these distributions of CS vary with the parameter values chosen. Furthermore, if the characteristics of the distribution of CS are subjective it does not make sense to attempt to interpret model output as an objective PDF.

7.5.2 Constraining using HFA

In this Section the use of Global Mean HFA (GMHF) to constrain the range of CS is discussed. The HFA is calculated in the calibration phase as the flux of heat between the ocean and the atmosphere required to force model SSTs to match climatological values, as explained in Chapter 5. The HFA acts as a surrogate for missing ocean processes. Where the GMHF is far from 0 the ocean and atmospheric components of a model are not in balance and thus might be thought to be prone to instability due the artificial imposition of a systematic radiative flux. It should be noted that GMHF need not be close to 0 in a perturbed physics model version since different parameter values might not have a non–zero global radiation balance Collins *et al.* (2006). In fact, it has been shown in Chapter 5 that there are simulations with significantly non–zero GMHF that do not show signs of instability in the time series of GMST. It is not clear that requiring parameter perturbed model versions to have a GMHF close to 0 is physically necessary.

It has been shown in Chapter 5, and in Figure 7.9, where GMHF plotted against CS, that there is clear structure between the GMHF and CS. In particular simulations with a strong, negative GMHF tend to have very high values of CS. It would not be statistical good practice to choose a GMHF filter to rule out high CS simulations based on this relationship alone; such a filter is ad hoc and similar filters could be devised to obtain different results. Instead a physical explanation for this relationship must be found and such relationships must not be sought specifically to gain a desired result, as previously state din this Chapter. It is shown in this Section that even if such a filter were used, with a stringent requirement to have a GMHF close to 0, it would still admit simulations with values of CS over 8 degrees Celsius.

GMHFs within the 64 standard HadSM3 ICE range from -2.18549 to -1.94776 W/m^2. This narrow range arises as a result of the low variability in GMHF within ICEs, as shown in Chapter 5. Applying a filter that requires the GMHF to be as

close to 0 as the standard HadSM3 ensemble (of absolute value less than 2.18549) results in a constraining method that allows simulations within the two vertical lines (denoting the range of GMHF in the standard HadSM3 ICE) shown in Figure 7.9. Whilst this filter eliminates some very high CS simulations, the range of CS values admitted is still large; from 1.6 to 8.2 degrees Celsius. Using $2.18549 W/m^2$ as a critical level for the GMHF rules out $\sim$ 92% of quality controlled simulations and might be considered as a fairly stringent requirement. The range of post–filter CS values is over twice as large as the interval of 1.5 to 4.5 degrees Celsius given in the IPCC Fourth Assessment Report Solomon *et al.* (2007a). Figure 7.9 suggests that whilst the GMHF is related with CS, but that a filter based on requiring GMHF to be close to 0 can not be used to reduce the range of CS simulated significantly. Furthermore, it would be statistical bad practice to use the GMHF to constrain CS only after having seen that a relationship exists.

7.5.3 Constraining using in–sample fit to observations

The third constraining method examined in this Chapter is based on comparing each model simulation to climatic observations. The method of calculating in–sample performance is similar to that applied in Stainforth *et al.* (2005), although here the in-sample performance of simulations is considered in up to 7 different variables unlike Stainforth *et al.* (2005) in which o in–sample performance is aggregated across 5 variables. Other attempts to constrain climate model output using observations has been carried out in Knutti *et al.* (2006) and Piani *et al.* (2005). Piani *et al.* (2005) uses CPDN data to search for observational constraints on climate model behaviour, in particular CS and the feedback parameter (defined as the inverse of CS). Whilst such search methods can help to identify correlations between observations and future behaviour it is important to further investigate these correlations to form consistent physical explanations. It is shown in this section that there may not be a single observational variable that can constrain CS but that

different observational variables can indicate different values of CS. This point has been noted previously in Sanderson *et al.* (2008) in which it is shown that different observational fields can give different constraints on CS. This section extends the results of Stainforth *et al.* (2005) to a set of 22712 simulations and confirms the results of Sanderson *et al.* (2008) that the choice of variable used to constrain CS can be critical.

The root mean square error (RMSE) compared to gridded observations, relative to the ICE mean of 64 standard HadSM3 simulations is used here to compare model simulations to observations. The scaling of RMSE error relative to the standard HadSM3 model does not affect any results but provides a more intuitive interpretation of the RMSE. The RMSE gives an indication of the relative in-sample skill of simulations in various variables, but the absolute values of these metrics should not be interpreted in terms of likelihood, as argued in Stainforth *et al.* (2005). In particular, it is not proposed here that there is a specific value at which model simulations should be considered realistic in the sense of RMSE proximity to observations. In each of the seven chosen variables, the CPDN PPE is pruned to those simulations that perform as well as the worst member of the 64 member standard HadSM3 ICE. The level below which a model should be dismissed is an important question, but is not dealt with here.

Initially, the performance of a 64 member ICE of the standard HadSM3 model is considered in 7 variables, followed by the CPDN PPE as a whole. This is done in order to evaluate whether there is a relationship between CS and in–sample fit and to analyse the effect of the choice of variable on any relationships that are found to exist. It is shown here that the choice of variables can effect the resultant distribution of CS when constraining model simulations.

The Metric

When eliminating simulations using in–sample fit or the HFA it is necessary to choose a critical value, above which models are ruled out. The worst member of the standard HadSM3 model will be used as a benchmark for the performance of the perturbed physics model simulations.

The 7 variables used to eliminate simulations here are : latent surface heat flux from land, latent surface heat flux from sea, sensible latent heat flux from sea, total cloud amount, total precipitation rate, sea surface pressure and 1.5 metre surface temperature.

The RMSE for a particular simulation, s, in variable j, is defined as:

$$\epsilon_{s,j}^2 = \sum_i w_i (m_{s,j,i} - o_{j,i})^2 \tag{7.1}$$

where $m_{j,i}$ is the simulation value for variable in j grid–box i averaged over the last 8 years of the control phase of that simulation, $o_{j,i}$ is the observed value for variable j in grid–box i and w_i is an area weighting for grid–box i[1]. The error for each simulation is then divided by the RMSE of the standard HadSM3 model to give the Relative RMSE score as for each simulation in variable j as:

$$\theta_{s,j}^2 = \frac{\epsilon_{s,j}^2}{\epsilon_{h,j}^2} \tag{7.2}$$

Where $\epsilon_{h,j}^2$ is the mean of $\epsilon_{s,j}^2$ for the standard HadSM3 model, averaged over the 64 ICE members. Hence the relative RMSE for each simulation is expressed as a scalar statistic representing the average disparity between the control phase simulation and observations.

In the case of constraining model diversity using in–sample fit, a simulation is required to match observations in its control phase at least as well as the worst simulation from the standard HadSM3 model in order to be admitted. Simulations

[1]Grid–boxes in the HadSM3 model correspond to different area of different size; grid–boxes are larger near the equator and smaller in the high latitudes.

that are worse than all 64 of the standard HadSM3 model simulations are rejected. This criteria is arbitrary, but the results are not believed to depend crucially on the precise level of RMSE used. The idea is rather to look at the use of different variables to constrain CS and whether it is possible to reduce uncertainties in a way consistent with statistical good practice and physical understanding.

Perturbed Physics Ensemble

The performance of simulations form the CPDN PPE is examined in this Section. Figure 7.10 shows the mean RMSE over 5 variables (to allow comparison of RMSE to CMIP simulations, 5 variables are used here) of 22712 simulations versus CS. The 5 variables used here are: annual mean temperature, sea level pressure, precipitation, and atmosphere–ocean sensible and latent heat flux; these were the only five variables were available from the CMIP II simulations used for comparison[1]. The CMIP GCMs are taken from the Coupled Model InterComparison Project II experiment Covey *et al.* (2003), developed at modelling centres across the world. The mean RMSE is found by taking the arithmetic mean of RMSE over these 5 variables, as in Stainforth *et al.* (2005). The use of such a composite score is not recommended here, since differences in the variability of RMSE between different variables could, without standardisation, lead to some variables gaining more influence than others. Nevertheless, this mean RMSE score is shown in order to highlight the importance of considering the performance of simulations in the individual variables that comprise the mean score.

The quartiles of the distribution of RMSE are coloured in Figure 7.10 to aid visualisation – the bottom quartile is shown in red, the 25th percentile to the median in green, the median to the 75th in dark blue and the upper 25th percentile in light blue. The mean RMSE shown is relative to the standard HadSM3 model; a model that performs similarly well on average to the standard HadSM3 model will score about 1 on the y–axis. Panels (a), (d) and (e) show that there are variables where

[1]I would like to thank David Stainforth for providing data for the CMIP II GCMs.

high CS simulations have a tendency to perform better than low CS simulations – surface temperature, atmosphere–ocean sensible and latent heat fluxes. In contrast to this, high CS simulations seem to perform worse in panels (b) and (c) (showing precipitation and sea surface pressure) although there is also a pattern for simulations with CS lower than 3 degrees to also perform worse. When using the mean of these five variables, the details average out and there seems to be a tendency for high CS simulations to perform worse in this metric. Figure 7.10 shows that the mean RMSE obscures that fact that high CS simulations tend to perform well in some variables and worse in others.

The PPE simulations perform well against 13 CMIP models (shown in black diamonds), with no CMIP models scoring in the bottom quartile of the PPE in all variables except surface latent heat flux in panel (e). The CPDN PPE model versions perform well against these GCMs in four of five variables, with 10, 11, 9, 10 and 2 of the 13 CMIP models falling into the worst quartile of the CPDN PPE in each of the 5 variables shown in panels (a) to (e).

Figure 7.10 demonstrates an important result since it suggests that there is no robust lack of compatibility with observations for simulations with high CS. This is not to say the same will be true for different types of modelling experiments and observational data sets, especially as more data becomes available throughout the 21st Century. It might be that observations that take into account dynamic changes due to CO_2 forcing will tend to favour particular values of CS in a more robust way than shown in Figure 7.10.

Figure 7.11 shows the resulting cumulative distribution functions (CDF) of CS from the CPDN PPE after constraining in 7 different variables. Some variables, such as precipitation (light blue, bottom panel), constrain the high end of CS to within about 7 degrees Celsius, whereas temperature (dark blue, top panel) does not constrain the range at all. Different variables constrain the range of CS by different amounts and "rule out" different numbers of simulations; 11032 simulations remain

after constraining with 1.5m surface temperature, whereas only 2876 remain after constraining with latent surface heat flux [1]. A stricter constraint was applied where a simulation is admitted if it performs as well as the worst member of the 64 member standard HadSM3 ICE in all 7 variables.

Applying the constraint in all variables (a simulation must pass in all 7 variables to be included) give the top–most CDF in Figure 7.11, shown in the upper panel; the highest admitted CS is about 5 degrees Celsius. This shows that it is possible to constrain model diversity by using more tests and drastically reducing the number of simulations (408 pass the test in all variables – ~1.4% of the original number). Constraining based on a single variable gives different results depending on which variable is used. Using more variables applies a stricter constraint but it is not clear how many should be used. Using all the data available to constrain model simulations would result in the CDF tending towards the standard model, since this is standard for in–sample fit used as the reference point[2]. If enough variables were included, at a sufficiently strict level of compatibility with observations, all simulations could be ruled out. It is not clear what level of model–observation agreement should be demanded when constraining model simulations.

Using different variables results in a different CDF for CS. Furthermore, the seven variables used eliminate different numbers of simulations. The greatest number of simulations pass the constraint using total cloud amount (~71% of simulations), the fewest pass the constraint based on latent surface heat flux from land (~11%). Constraining using multiple variables provides a stricter constraint (stricter only in the sense of ruling out more simulations) on CS; when all variables are used only 1.4% of simulations pass. Using multiple constraints fewer simulations are allowed; it could be that if enough constraints are added, the number of simulations admitted could be reduced to a very small number, even 0.

[1] A similar result was found in Sanderson *et al.* (2008) where it is shown that top of atmosphere radiative flux provides a stronger constraint on CS than temperature or precipitation.

[2] It is interesting to note that the method used in Piani *et al.* (2005) produced a best estimate of CS of 3.3 degrees Celsius – very close to the median value of the CPDN CDF used here, when constrained by all 7 variables.

It is important to consider whether the performance of model simulations is correlated in different observational variables. In particular, it would be misleading to prefer one simulation to another on the basis of its ability to simulate observed temperatures, then use this model to simulate precipitation, without considering the relative performance in precipitation. In order to look at this, Figure 7.12 shows the relative RMSE in precipitation versus temperature over all 22711 quality controlled simulations. The structure indicates a pattern for simulations with worse in–sample precipitation fields to also have worse temperature fields. This relationship is far from universal – there are simulations with an relative RMSE better than the standard model in temperature (an RMSE of less than 1) and a relative RMSE of 3 or more in precipitation. There is a relationship between in-sample performance in precipitation and temperature, suggesting that constraints are not independent. Simulations that perform well in temperature, have a tendency to also perform well in precipitation although, since this is not always the case, it would seem sensible to look for hindcast skill in all variables of interest.

As, shown in Figure 7.11, precipitation is a stricter constraint than temperature and constraining using these variables has different effects on the shape of the CDF of CS. In particular, constraining using temperature seems to skew the CDF towards higher values and precipitation towards lower values. This suggests that the effect of constraining model output using these variables is not equivalent to eliminating simulations at random, since the shape of the CDF changes with the variable chosen.

The performance of model simulations in–sample has used the global mean RMSE relative to the standard HadSM3 model. Three important limitations to using global mean relative RMSE as a guideline for model performance are:

1. By using the standard model as a benchmark, the relative RMSE shown does not give an indication of systematic bias present in all model simulations, nor any idea of the magnitude of model errors.

2. The global mean statistic of RMSE used is here can only be a rough guide to model performance. More detailed spatial and temporal analysis is required for decision–relevant model evaluation.

3. The constraint is based on a static climate; no dynamical changes are included. Arguably, the ability of models to match dynamical changes in the Earth's climate system are the most important test of likely out–of–sample skill. Such a test can not be carried out using the data analysed here since CO_2 is instantaneously doubled rather than increased transiently. Attempts by Annan & Hargreaves (2006); Knutti *et al.* (2006) to constrain CS using dynamical changes have yielded a reduction in high CS values, although the shape of this distribution itself is, to some extent, an artifact of the experimental design Frame *et al.* (2005).

7.6 Conclusion

The range of values for CS shown in the CPDN PPE is unprecedented and poses new challenges for the interpretation of model output for decision makers. Attempts to constrain such a large range of model behaviour face several difficulties. Appropriate variables must be chosen for which to evaluate the model output. To conform with statistical good practice, these variables must be chosen with relevance to out–of–sample model performance and not because they reduce model diversity by the largest amount. The importance of this principle is demonstrated using the examples of constraining with the Entrainment Coefficient, global mean HFA and model in–sample fit.

The use of global mean heat flux (GMHF) to constrain the range of values for CS, having seen the GMHF is related to CS, does not allow us to eliminate all high CS simulations; simulations with values of CS over 8 degrees pass this test. Furthermore, this would be statistical bad practice since we would be choosing

the constraint based on the results we would like to see. The criteria used for constraining CS should be ideally chosen before running the experiment (or at least seeing the effect that various constraints would have on the data). Despite the difficulties in applying a constraint on CS based on the GMHF, it is useful to see the relationship between GMHF and CS since this might lead to an improved understanding of why this relationship occurs and why some simulations show such a large degree of warming. A physical understanding of why simulations have different behaviour is paramount to constraining the range of behaviour.

The use of in–sample fit as a means of assessing model diversity is analysed in this Chapter. The choice of variables is shown to have an effect on the distribution of CS. In particular, using temperature as a constraining variable tends to permit more high CS simulations and using precipitation tends to permit more low CS simulations. The more observational variables simulations are required to perform well in, the fewer simulations will be admitted. In the limit, all simulations can be ruled out since none are realistic in every relevant sense. Attempts to constrain the model simulations here are based on a small set of observed data (30 years long) during which no significant climate change occurred. The ability for models to produce useful forecasts of climate change might better be assessed using a transient experiment and data during which the climate changes e.g. using observations of the 20th Century. This would test the dynamical strength of the models and provide a more relevant test for out–of–sample model performance.

In the next Chapter, a sub–set of simulations is analysed that all have very similar values of CS. This sub–set can be thought of both as a reflection the range of behaviour in simulations after successfully CS to a narrow range, even if such a constrain could be robustly achieved, and as an analysis of the utility of global mean metrics for decision–support.

New results from a grand ensemble of 45644 simulations of HadSM3 presented in this Chapter are:

- The range of behaviour shown in an ensemble of 45644 of GCM simulations is unprecedented, with estimated CS ranging from 0.9 to over 16 degrees Celsius.

- By comparing sub–sets of the CPDN PPE it has been shown that the shape of the distribution of CS is not an inevitable feature resulting from an approximately Gaussian distribution of feedbacks, as was proposed in Roe & Baker (2007). The distribution of CS can be changed substantially by a different choice of experimental design.

- The global mean HFA and the Entrainment Coefficient can be used to change the distribution of CS, but simulations with CS over 8 degrees are still admitted. The use of such post-hoc filters is criticised on the basis of bad statistical practice.

- When comparing model performance in–sample in 7 different variables, results depend on the choice of variable. For example, constraining in temperature tends to admit more high CS simulations and constraining in precipitation more low CS simulations.

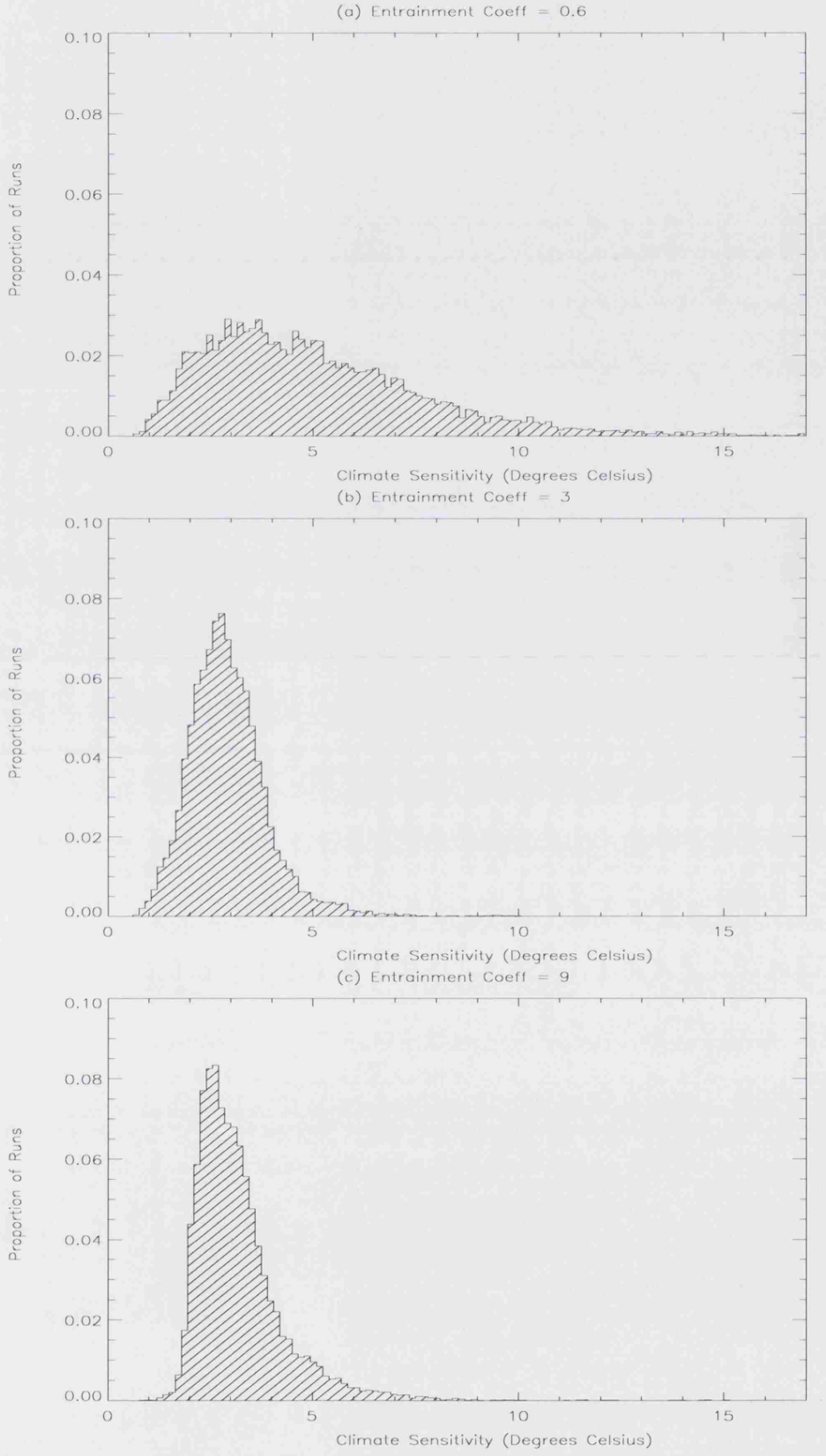


Figure 7.7: The distribution of CS is shown for all quality controlled simulations for three different values of the Entrainment Coefficient - 0.6 (low) in panel (a), 3 (standard) in panel (b) and 9 (high) in panel (c).

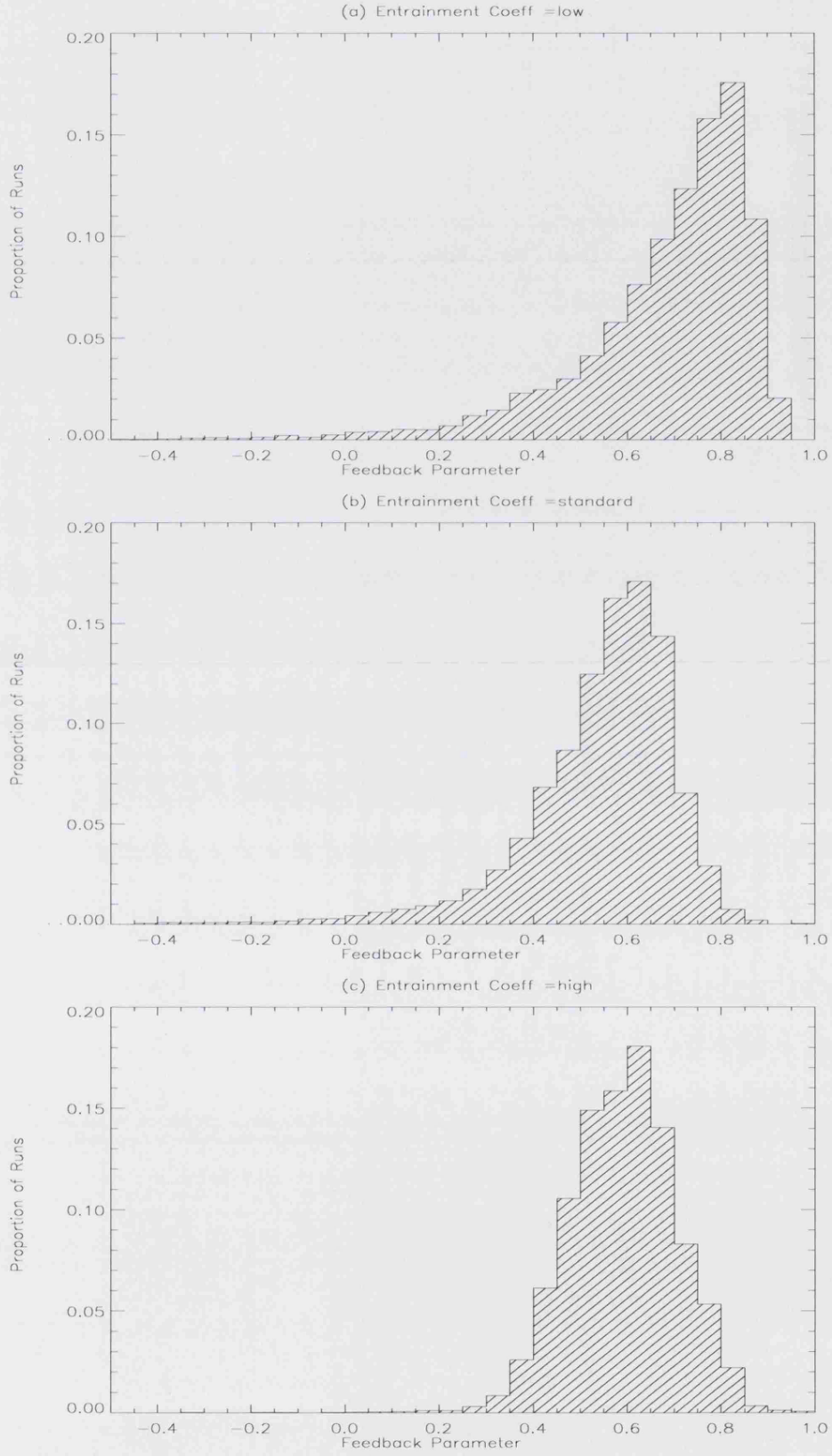


Figure 7.8: The implied distribution of feedbacks for three different values of the Entrainment Coefficient

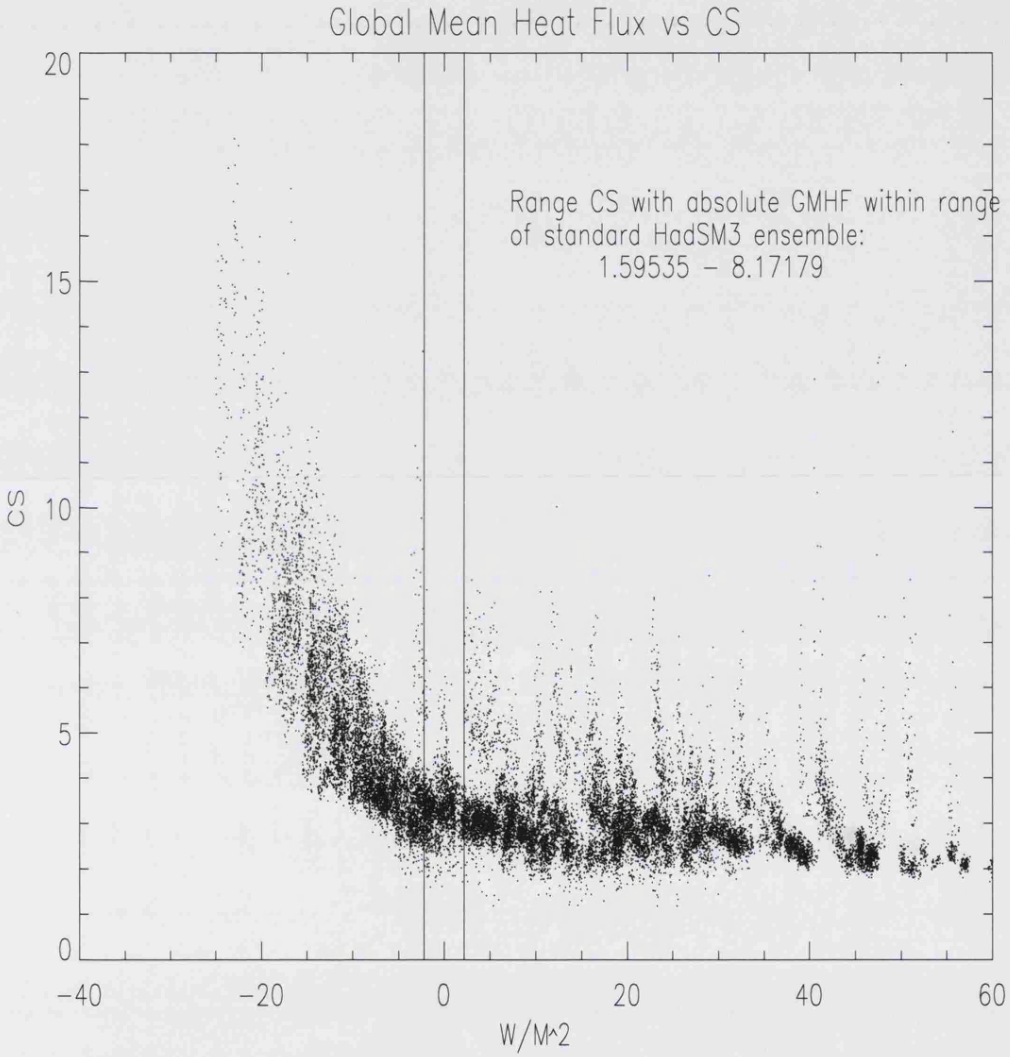


Figure 7.9: The global mean HFA is plotted against CS. The vertical lines denote the largest absolute values of global mean HFA in the 64 members standard ensemble. The range of CS captured by this range is (1.59535, 8.17179).

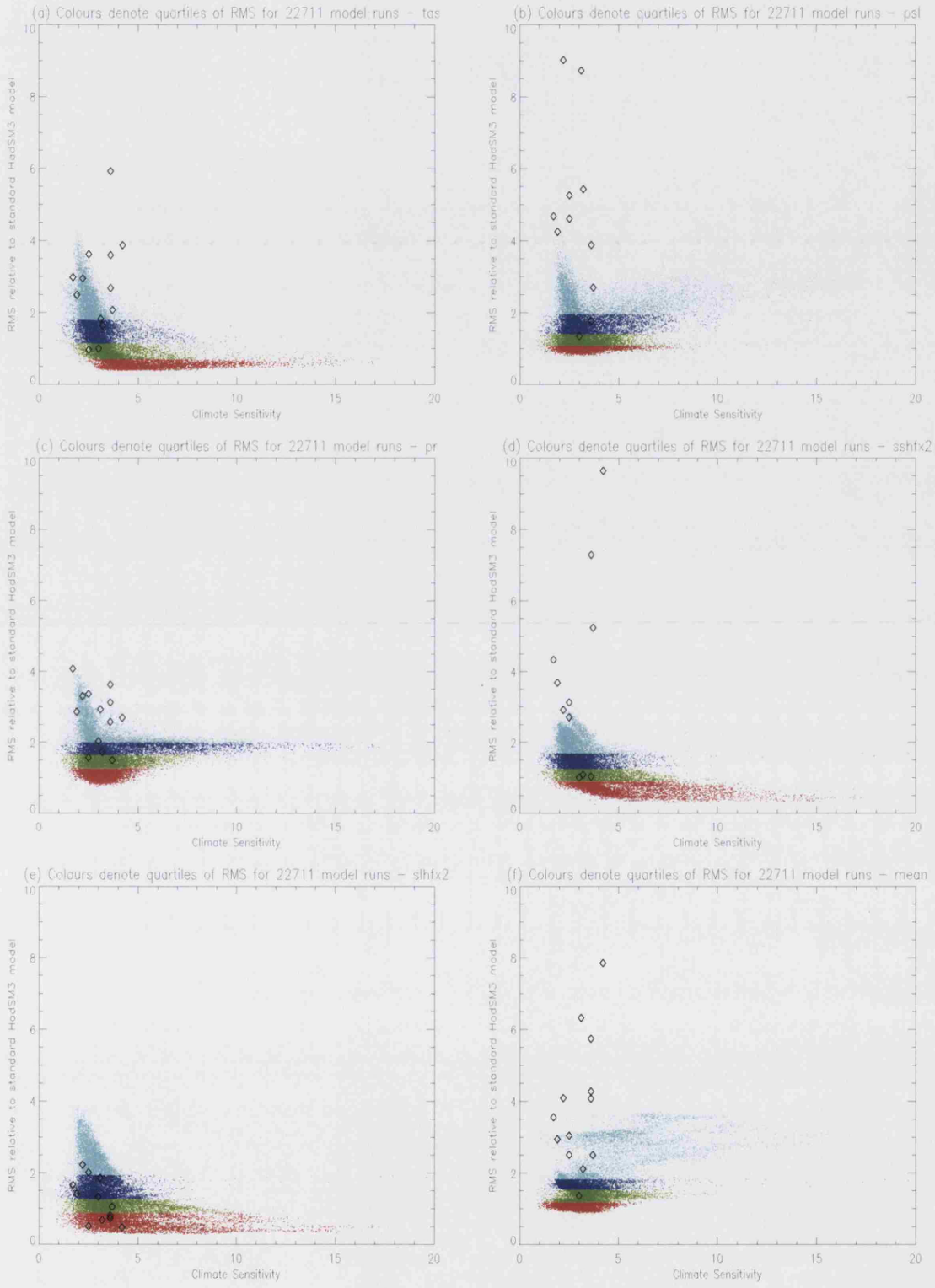


Figure 7.10: The RMSE, relative to the standard model, is plotted against CS for 22712 simulations in five different variables – (a) surface temperature, (b) sea surface pressure, (c) precipitation, (d) surface sensible heat flux from sea and (e) surface latent heat flux from sea. Panel (f) shows the average RMSE error over these five variables. The values for 13 GCMs taken from the CMIP II project are plotted as black diamonds.

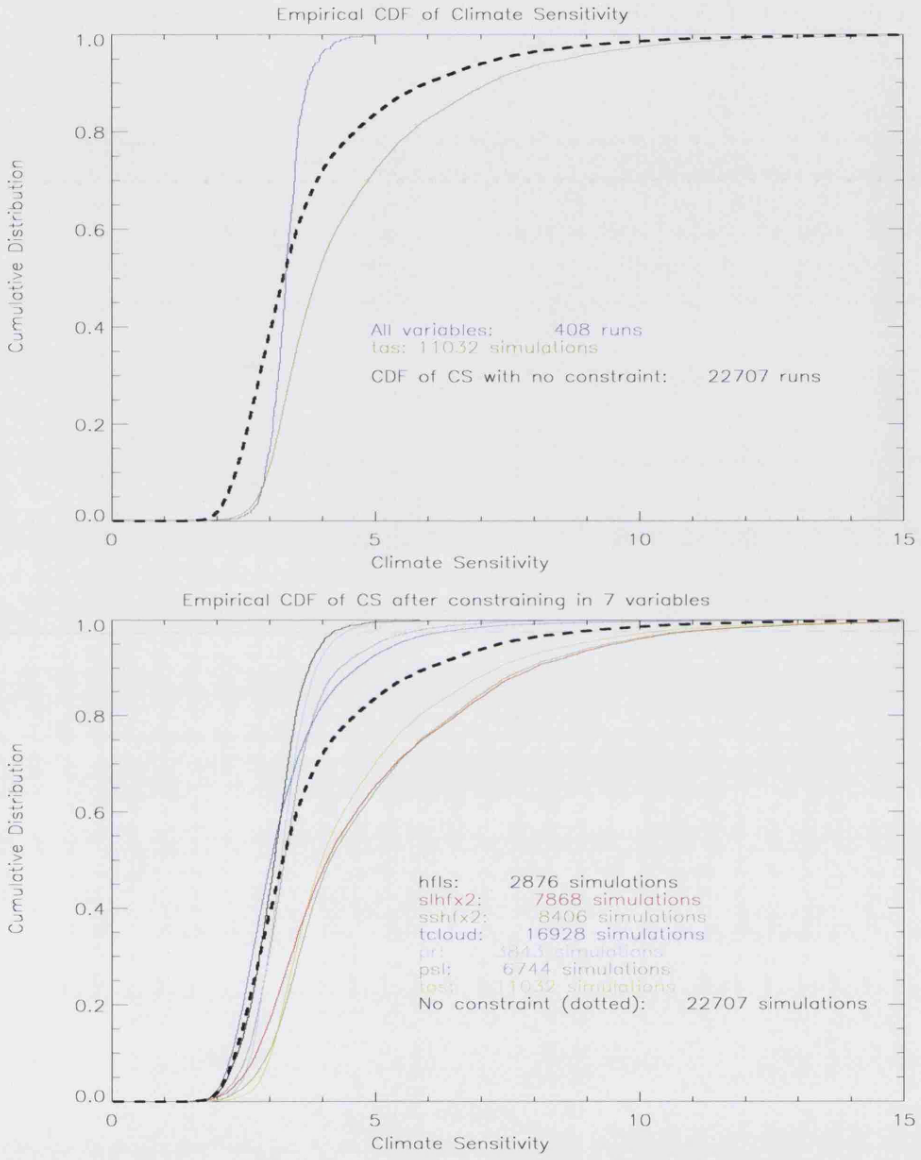


Figure 7.11: The Cumulative Distribution Function of CS for the grand ensemble, including only simulations with an RMS error no higher than the worst member of the 64 member standard ensemble. In the top Figure, temperature is used as the observational constraint (green). Also shown is the effect of constraining in 7 different observational variables simultaneously. The variables shown are heat flux latent surface, total precipitation rate, sea surface pressure, 1.5m temperature, surface sensible heat flux from sea, surface latent heat flux from sea, total cloud amount. The number of simulations left after applying constraining in each variable is shown adjacent to each variable's name.

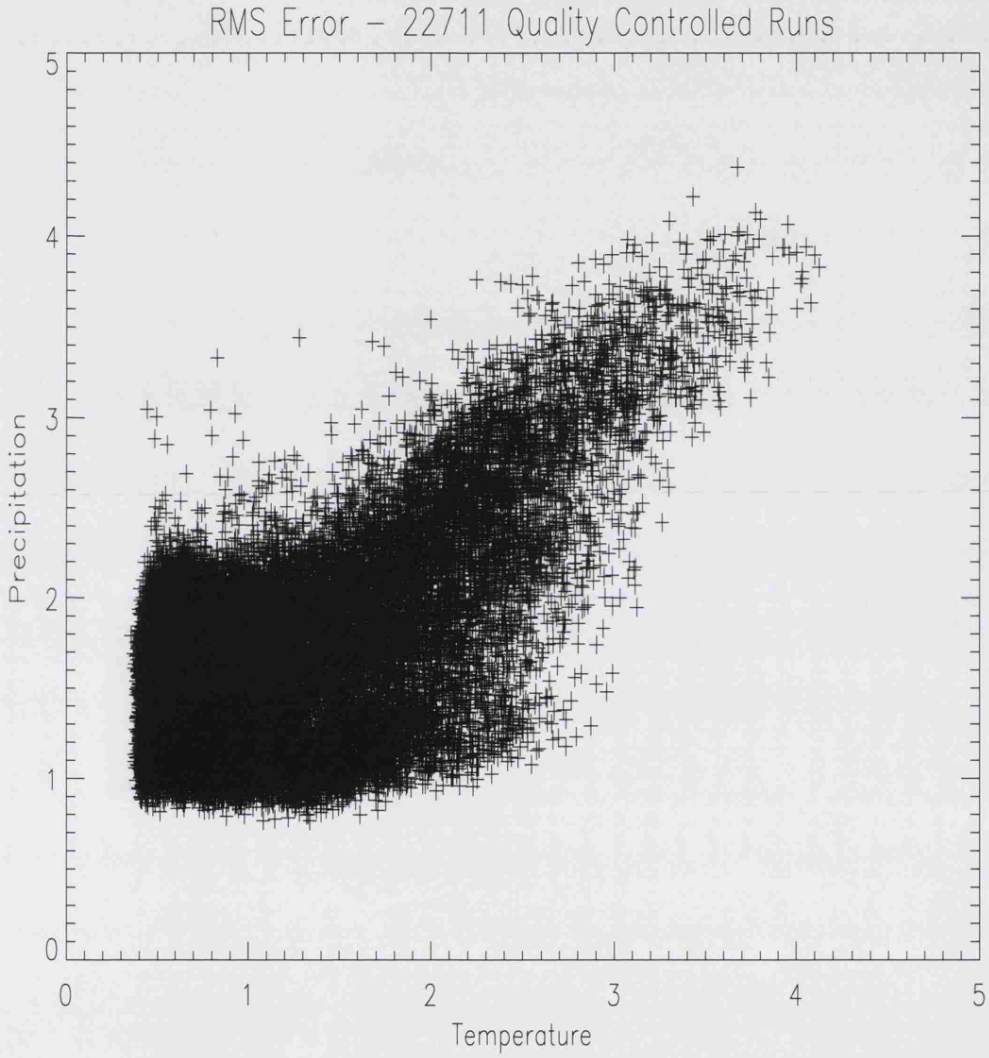


Figure 7.12: RMSE, relative to the standard model, for precipitation and temperature for 22711 quality controlled simulations. There is a pattern for simulations with a worse score in one variable to have a worse score in the other, although a number of exceptions exist.

Chapter 8

The relevance of global means for climate policy

8.1 Introduction

Climate simulations might provide information on strategies to mitigate or adapt to climate change. This Chapter looks at the utility of GMST (Global Mean Surface Temperature) as an index variable for policy and impact studies. The relevance of GMST as an index is then discussed from the point of view of mitigation and adaptation policies.

In recent years, there has been a significant rise in interest in potential impacts of climate change. Such impacts inevitably occur on sub–global length scales. Examples include the recent UK Climate Impacts Program (UKCIP), the recent IPCC Fourth Assessment Report Solomon *et al.* (2007a) and the Stern report on the Economics of Climate Change Stern (2006). It is often the case that climate simulations are taken as the basis for impact studies or adaptation strategies and important uncertainties are ignored. An inadequate treatment of inter–model differences and known uncertainties can result in over-confidence in the use of climate simulations, especially if uncertainties are lost either by assuming future climate is known (using a single, deterministic simulation) or by assuming a one–to–one relationship between

GMST and local impacts. This could lead to a lack of flexibility in planning or commitment to sub–optimal (or damaging) action due to over-confidence.

Plans for stabilisation of GHGs are often expressed in terms of GMST commission (2007); Parry *et al.* (2007); Solomon *et al.* (2007a); Toth *et al.* (2003); Yohe *et al.* (2004). GMST is used in impact studies for a number of reasons; it is supposed that models can produce realistic hindcast simulations of GMST Solomon *et al.* (2007a) and it is a fairly intuitive statistic to deal with. Many impact studies relate changes in GMST directly to impacts Stern (2006) (p.180), Parry *et al.* (2007). This approach poses a number of problems that will be discussed in this Chapter. In particular, information on the regional responses to climate change and the associated uncertainties is lost. It is shown here that relating GMST change directly to sub–global climatic impacts can be highly misleading. Climate can be forecast, at best as a distribution Smith (2002) with its inherent uncertainties; moreover these uncertainties are larger on smaller length scales.

As a consequence of the analysis presented in this Chapter, it is shown that constraining GMST (or CS), even to within a very narrow range, may not provide a strong constraint on regional changes. The consequences of this are discussed.

Using data from the CPDN experiment, the problem of how GMST change relates to regional change is looked at in this Chapter. For some variables and length scales, such as GMST, the models produce consistent results. In other cases, such as regional precipitation changes, the models produce inconsistent results and raise important questions about the degree of flexibility required in adaptation decisions. The extent to which GMST changes provide information on the length scales most obviously relevant for impact studies is investigated. The cause of climatic change is global but the impacts are always local. Whilst there are climatic changes that are local in extent, such as land–use change and volcanic activity, the simulations used here look at the possible effects of doubling atmospheric concentrations of CO_2, a GHG that mixes throughout the atmosphere and thus has a global impact. Nowhere

is affected by a global mean *per se*; rather each region and location experiences its own local climate. Whilst global mean statistics do have an important part to play in climate science, their relevance in terms of their direct use for climate impacts is limited.

Two specific questions are discussed in this Chapter;

1. What does a 2 degree rise in GMST mean for decision–makers? Different spatial patterns of change, and associated uncertainties, can be averaged out to give the same value of GMST. It is shown in this Chapter that many different local changes in climate can result in a 2 degree rise in GMST. The uncertainty in response to a doubling of CO_2 is looked at on a variety of length scales for simulations with a fixed GMST response. This allows for an assessment of how much information GMST provides on climatic changes on various length scales.

2. How different is a rise in GMST of 2 degree Celsius compared to a rise of 3 degrees? This question has importance both for policy decisions and localised impact assessment. In the case of adaptation, it is important to know how quickly our adaptation plans may have to change – if 3 degrees is much worse than 2 degrees, we should prepare accordingly or least be flexible enough in our plans to adapt to the consequences of greater warming effectively. It may be that the local effects of a monotonic rise in GMST change result in non–linearities in regional climate response and adaptation methods planned accordingly. Another example of the limitation of annual mean global mean metrics, shown in this Chapter, is that there can be different effects in summer and winter e.g. a drier winter and a wetter summer should not be assumed to be equivalent to "annually no change" for impact assessment.

Planning mitigation strategies can be helped by understanding how policies targeting GMST may result in different regional responses. For example, if it is

thought that rising GMST might lead to an increased risk of flooding in London, policies could be adopted to mitigate this risk. It would then be useful to know how much more risk is attributable to a 3 degree rise compared to a 2 degree rise. This work is also relevant to the problem of constraining CS in previous work in Chapter 7 and in the literature in general Annan & Hargreaves (2006, 2007); Forest *et al.* (2007); Hegerl (2006); Knutti *et al.* (2006), where attempts are made to constrain uncertainties in the value of CS. The utility of such global constraints for decision–makers reliant on regional climate information is questioned in this Chapter. Regional uncertainties are examined here for a narrow range of GMST change (0.2 degrees Celsius). Approaches that constrain climate change using CS may not reduce regional uncertainties to manageable levels; indeed is it shown that for simulations with 2 degrees GMST rise the range of 8 year mean DJF warming simulated in the Central North American region is from 1 to 7 degrees Celsius.
The structure of this Chapter is as follows. Section 8.1.1 describes the data sets used in this Chapter. Section 8.2 answers question **1)** above by analysing the distribution of temperature and precipitation changes on various length scales for a set of simulations with 2 degree GMST rise. Section 8.3 answers question **2)** above, the difference between simulations with 2 degrees GMST rise and 3 degrees GMST rise. Section 8.4 examines the assumption that regional responses are linear with respect to GMST rise. Section 8.5 discusses the results presented in this Chapter and the implications for mitigation and adaptation decisions. Section 8.6 gives conclusions.

8.1.1 Data Sets Used

Data from a grand ensemble of 45644 simulations is analysed in this Chapter. Model simulations within a narrow range of 0.2 degrees GMST change are analysed in terms of their regional distributions. Where simulations are said to have a GMST rise of X degrees, this refers to all simulations with a GMST change of between X

±0.1 degrees Celsius. Changes in 8 year mean temperature and precipitation (the regional time series are described in Chapter 4) are studied on the regional scale (regions are defined as in Giorgi & Mearns (2000)), and the grid–box level (in the HadSM3 model, grid–boxes are 3.75 degrees in longitude by 2.5 degrees in latitude – roughly 200 by 200 kilometres) for seasonal temperature and precipitation. The set of 402 quality controlled simulations with 2±0.1 degrees of GMST rise (defined by the difference between the last 8 years of the doubled CO_2 phase and the last 8 years of the control phase for comparability with the available regional data) is referred to hereafter as the 2 degree set. Similarly , further sets of 2441 and 795 quality controlled simulations with 3 and 4±0.1 degrees of GMST rise are referred to as the 3 and 4 degree sets respectively.

8.2 What does a 2 degree rise in GMST mean?

This Section looks at the level of regional uncertainty associated with the 2 degree set of simulations, as described in Section 8.1.1. This analysis is motivated by two questions that are important for assessing the utility of GMST as a target, or basis, for decisions:

1. To what extent can mitigation strategies based on GMST change reduce significant regional risks?

2. To what degree of certainty can impact studies link GMST to particular impacts?

These questions can be informed by looking at the distribution of climate response on various length scales when considering simulations with a given GMST rise. Model diversity is now examined beginning on hemispheric length scales, then smaller regions. It is shown that model diversity increases significantly when analysing the distribution of temperature and precipitation response on regional length scales.

8.2.1 From global to super–continental length scales

This Section looks at seasonal changes for the 2 degree set in temperature and precipitation on global and super–continental length scales: namely, the Northern and Southern hemisphere, the tropics and the extra–tropics.

Figure 8.1 shows the change in global mean, tropical, Northern hemisphere extra-tropics and Southern hemisphere extra–tropics for temperature and precipitation. The changes shown are in terms of the 8 year seasonal mean under a doubling of CO_2 concentrations. For each region, four plots are shown – DJF (December through February mean), JJA (July through August mean) for temperature and precipitation. The x–axes in Figure 8.1 are deliberately set at a wider range than the distributions of change might appear to justify so that similar distributions on smaller length scales can be compared on the same x–axes.

The range of temperature change is from 1.5 to 2.5 degrees Celsius in the global mean and 2 to 4 degrees in the Northern hemisphere in both seasons. Southern hemispheric change in temperature is lower – 1 to 2 degrees in DJF and about 2 degrees in JJA. The lower levels of warming simulated in the Southern hemisphere is partly due to there being more ocean in the Southern Hemisphere than the Northern hemisphere (and oceans typically warm by less than land masses). The range of precipitation change is greater on the hemispheric level than the global level; 0–20% in the Northern hemisphere and -10–+10% in the Southern hemisphere. In the summer months, the magnitude of warming is less than in the winter. The global mean range is very tight, showing a increase of about 0–5%. Inferring temperature changes on a hemispheric scale from the GMST change is robust across the 2 degree set. This suggests that is not unreasonable to infer global or hemispheric temperature response from GMST rise. In contrast, precipitation changes are always not robust – even the sign of the change is uncertain in the Southern hemisphere in the 2 degree set despite all simulations having a GMST change within a range of 0.2 degrees Celsius. This suggests that it is not possible to robustly infer

the direction of Southern hemispheric precipitation response based on GMST rise in the HadSM3 model. This difference can be seen from the variable magnitude of temperature change between the DJF and JJA seasons in the Northern and Southern hemispheres. These hemispheric differences will often cancel out on the global scale (in both DJF and JJA it will be summer in one hemisphere and winter in the other).

There are two reasons for the narrowness of these hemisphere–averaged ranges. First, averaging over many grid–boxes is expected to reduce the sampling variability considerably. Second, regional variations in the sign of precipitation change partially off–set each other e.g. an area with reduced simulated precipitation and an area with increased simulated precipitation can cancel out, when averaged, to show little or no change. Large changes of different sign in regional precipitation change can cancel out; information that is hidden when using global mean statistics. Figure 8.2 shows the same statistics as Figure 8.1 but for different areas: the tropics, Northern hemispheric extra–tropics and Southern hemisphere extra–tropics. These areas are smaller than the global and hemispheric areas shown in Figure 8.1. The range of temperatures for these areas is similar in magnitude to the hemispheric averages, although the Northern extra–tropics does show a wider range of temperature response (the range has a width of about 2 degree Celsius in both seasons). Furthermore, the magnitude of warming in the Northern hemisphere extra–tropics is higher than on global and hemispheric scales – about 3 degrees in JJA and 4 degrees in DJF. Notably, precipitation in the tropics is very tight in distribution.

On super–continental length scales, the sign and magnitude of temperature change is robust across the model ensemble, with a range of up to 2 degrees Celsius in some areas and seasons. Precipitation change is generally robust, but in regions such as the Southern hemisphere the response is more uncertain (the change in precipitation response varies from $\sim$-15% to $\sim$+10%). When taking averages over large areas, significant changes in precipitation are cancel out. For example, it is shown

in Figure 8.8 that the tropics show precipitation responses in different directions on a grid–box level, but the cancellation of these changes gives the tight distribution shown in the Tropics in Figure 8.2.

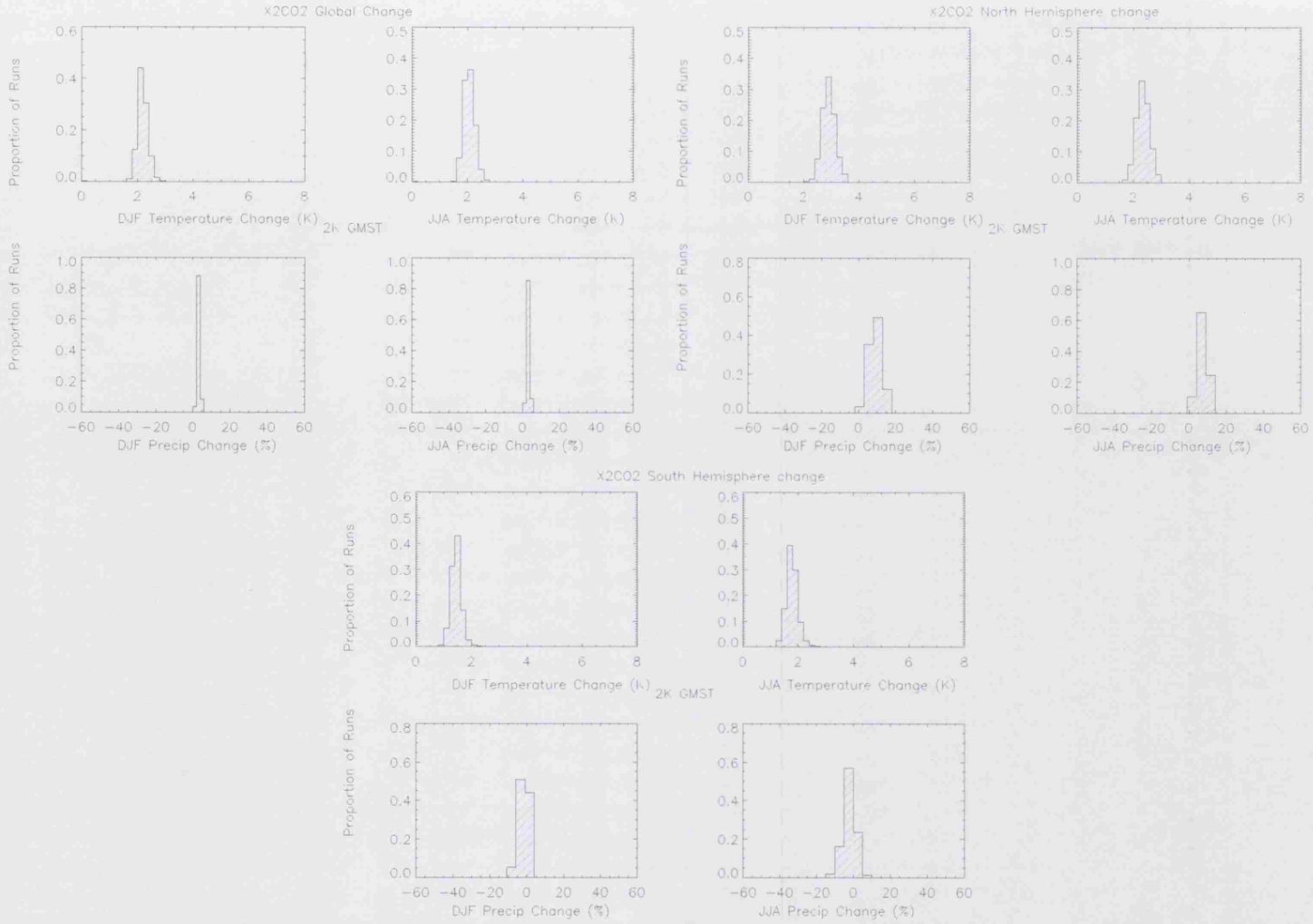


Figure 8.1: The distribution of 8 year DJF (JJA) temperature (precipitation) is shown for simulations for the 2 degree set. The x–axis range is maintained for following regional plots for ease of comparison. Note that there is little variance in precipitation where averages are taken over these large areas.

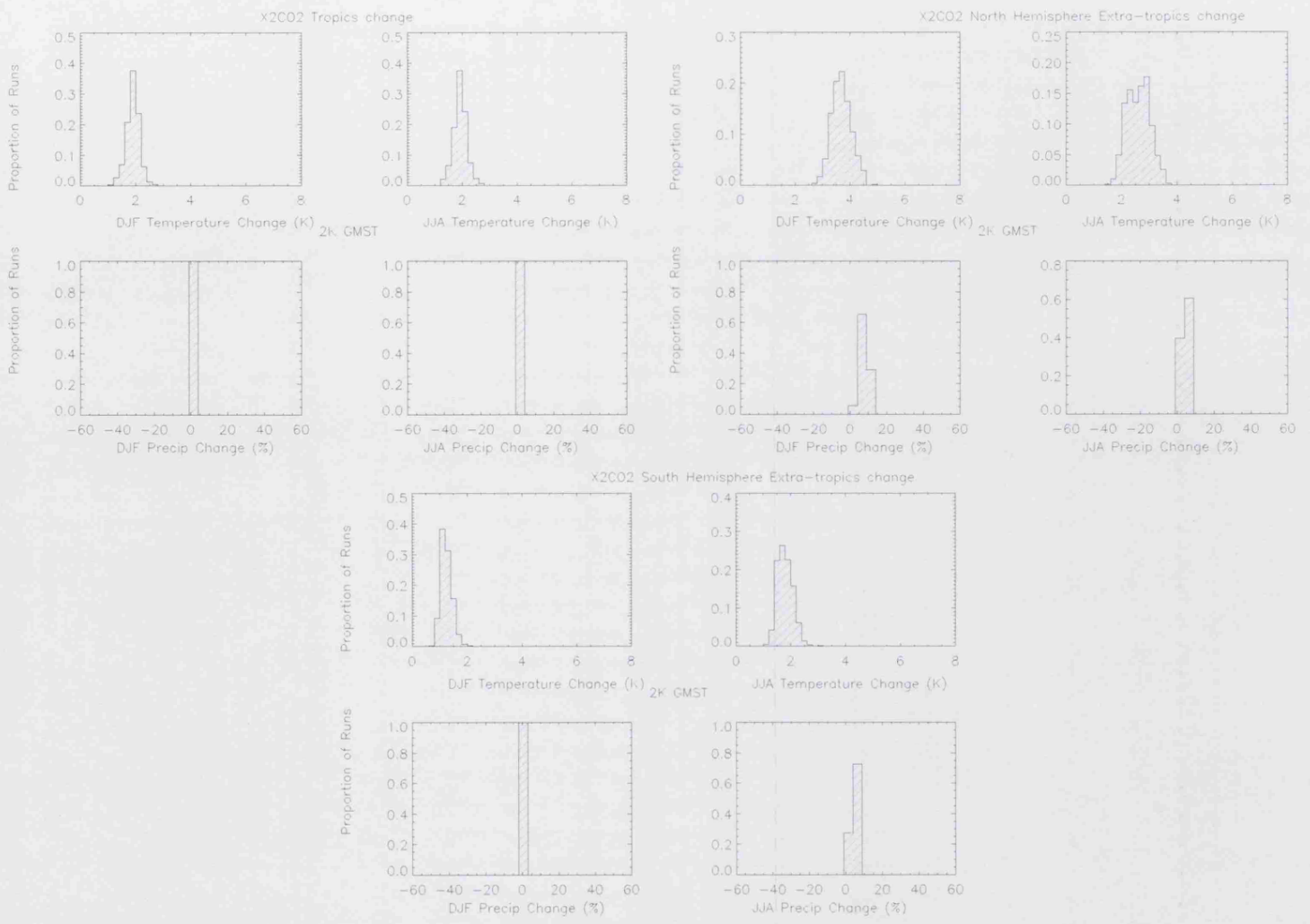


Figure 8.2: The distribution of 8 year DJF (JJA) temperature (precipitation) is shown for simulations for the 2 degree set. The x–axis range is maintained for following regional plots for ease of comparison.

8.2.2 Regional Impacts

Arguably, almost all climate impact decisions require information on regional or lower length scales. This subsection shows that model diversity increases significantly when looking at regional temperature and precipitation change compared with continental and super–continental areas. Regional precipitation change is shown in this Section to be particularly uncertain with a lack of robustness in simulating the sign or magnitude of change.

Figure 8.3 shows the seasonal change in regional temperature and precipitation. Three regions are selected for comparison – Australia, Central North America and Northern Europe. All the land regions' (plus the global, hemispheric, tropical and extra–tropical areas) minimum and maximum change in seasonal temperature and precipitation for the two degree set are shown in Tables 8.1, 8.2 and 8.3. In no cases is there a fall in temperature in either season. In almost every case, the sign of precipitation change is uncertain; the exceptions are Greenland, Antarctica and the Global mean. The regional changes in temperature and precipitation are dealt with in turn.

Region	Variable	Season	Min	Max
Australia	Temp	DJF	0.957	3.00
Australia	Temp	JJA	0.859	2.79
Australia	Precip	DJF	-28.3	59.8
Australia	Precip	JJA	-33.4	50.6
Amazon Basin	Temp	DJF	1.23	5.42
Amazon Basin	Temp	JJA	1.00	5.72
Amazon Basin	Precip	DJF	-25.9	13.2
Amazon Basin	Precip	JJA	-30.8	15.8
Southern South America	Temp	DJF	0.650	2.69
Southern South America	Temp	JJA	0.640	3.09
Southern South America	Precip	DJF	-7.13	18.8
Southern South America	Precip	JJA	-19.5	26.9
Central America	Temp	DJF	1.24	3.77
Central America	Temp	JJA	1.64	4.00
Central America	Precip	DJF	-42.0	105.
Central America	Precip	JJA	-37.8	78.7
Western North America	Temp	DJF	0.320	4.56
Western North America	Temp	JJA	2.12	5.30
Western North America	Precip	DJF	-28.3	41.2
Western North America	Precip	JJA	-26.4	18.3
Central North America	Temp	DJF	0.887	7.22
Central North America	Temp	JJA	2.29	8.48
Central North America	Precip	DJF	-34.2	46.3
Central North America	Precip	JJA	-44.8	30.0
Eastern North America	Temp	DJF	1.03	5.43
Eastern North America	Temp	JJA	1.95	4.58
Eastern North America	Precip	DJF	-16.0	43.6
Eastern North America	Precip	JJA	-10.8	30.8
Alaska	Temp	DJF	0.00134	9.57
Alaska	Temp	JJA	0.130	3.11
Alaska	Precip	DJF	-8.41	68.9
Alaska	Precip	JJA	-9.72	27.8
Greenland	Temp	DJF	1.98	6.78
Greenland	Temp	JJA	0.963	2.86
Greenland	Precip	DJF	5.15	28.4
Greenland	Precip	JJA	1.98	24.4
Mediterranean Basin	Temp	DJF	1.40	4.11
Mediterranean Basin	Temp	JJA	1.81	5.80
Mediterranean Basin	Precip	DJF	-25.5	26.3
Mediterranean Basin	Precip	JJA	-58.2	2.93

Table 8.1: The minimum and maximum change for DJF and JJA seasons for temperature and precipitation for the 2 degree set.

Region	Variable	Season	Min	Max
Northern Europe	Temp	DJF	1.32	6.31
Northern Europe	Temp	JJA	0.987	4.11
Northern Europe	Precip	DJF	-2.97	40.0
Northern Europe	Precip	JJA	-23.8	27.0
West Africa	Temp	DJF	0.992	3.28
West Africa	Temp	JJA	0.708	3.30
West Africa	Precip	DJF	-62.4	108
West Africa	Precip	JJA	-38.6	56.0
East Africa	Temp	DJF	0.827	4.28
East Africa	Temp	JJA	1.34	4.16
East Africa	Precip	DJF	-11.8	40.7
East Africa	Precip	JJA	-25.4	58.8
Southern Africa	Temp	DJF	0.929	2.60
Southern Africa	Temp	JJA	0.975	2.65
Southern Africa	Precip	DJF	-14.8	29.4
Southern Africa	Precip	JJA	-36.6	4.35
Sahara Region	Temp	DJF	1.09	4.42
Sahara Region	Temp	JJA	2.14	4.32
Sahara Region	Precip	DJF	-77.9	146
Sahara Region	Precip	JJA	-40.2	121
South East Asia	Temp	DJF	1.24	2.46
South East Asia	Temp	JJA	0.978	2.78
South East Asia	Precip	DJF	-14.4	44.6
South East Asia	Precip	JJA	-30.8	33.7
East Asia region	Temp	DJF	1.83	4.99
East Asia region	Temp	JJA	1.83	4.55
East Asia region	Precip	DJF	-20.3	53.5
East Asia region	Precip	JJA	-16.1	38.5
South Asia region	Temp	DJF	1.56	4.01
South Asia region	Temp	JJA	0.715	3.23
South Asia region	Precip	DJF	-46.8	206
South Asia region	Precip	JJA	-12.0	50.6
Central Asia region	Temp	DJF	0.848	5.31
Central Asia region	Temp	JJA	2.01	5.81
Central Asia region	Precip	DJF	-33.9	55.3
Central Asia region	Precip	JJA	-36.1	64.6

Table 8.2: The minimum and maximum change for DJF and JJA seasons for temperature and precipitation for the 2 degree set.

Temperature

Three regions are selected for more detailed examination in this Section – Australia (with relatively low warming of ~3 degrees or less in DJF and JJA for the two degree set), Northern Europe (medium levels of warming of 4.1 degrees in JJA and 6.3 degrees in DJF) and Central North America (high levels of warming of 8.5 degrees in JJA and 7.2 degrees in DJF). All three regions selected for comparison warm in both seasons, but with considerable differences in magnitude. Australia warms by about 2 degrees in both seasons (plus or minus about 1 degree) and Central North America typically warms by over 4 degrees, with a wide range of change from 2 degrees up to 8 degrees. Warming is generally greater during the winter months (DJF season for Central North America). Northern Europe similarly warms more during the winter, but typically by around 4 degrees in winter and 3 degrees in summer. The range of temperature change seen in regions such as Central North America raises difficulties in evaluating whether 2 degrees of global warming is a relevant target for local adaptation strategies. Considerably different policies are required to adapt to 2 degrees of regional warming than for 8 degrees.

Precipitation

Unlike temperature, even the sign of precipitation change is uncertain in all four regions. In the Australia region, the range of precipitation change in both seasons ranges between roughly -30% up to +50% with the peak of the distribution close to 0 (no change). The distribution of Central North American precipitation is centred about 0 in the summer months and at about +10% in the winter with a range of (-34%, 46%) in DJF and (-45%, 30%). Most simulations show an increase in Northern European precipitation in both summer and winter, but with a wide range of values for the magnitude of this change. In these regions, especially Australia and Central North America, it is difficult to say what a 2 degree Celsius in GMST would mean for precipitation. It seems the change could be up to 50% in many regions,

Region	Variable	Season	Min	Max
Tibet	Temp	DJF	2.67	5.79
Tibet	Temp	JJA	1.94	5.58
Tibet	Precip	DJF	-26.8	65.8
Tibet	Precip	JJA	-14.0	16.5
North Asia region	Temp	DJF	1.54	7.55
North Asia region	Temp	JJA	1.06	4.74
North Asia region	Precip	DJF	6.71	46.0
North Asia region	Precip	JJA	-2.99	17.2
Antarctica	Temp	DJF	0.246	2.28
Antarctica	Temp	JJA	1.07	5.28
Antarctica	Precip	DJF	0.162	19.9
Antarctica	Precip	JJA	6.24	27.9
North Hemisphere	Temp	DJF	2.10	3.58
North Hemisphere	Temp	JJA	1.63	2.97
North Hemisphere	Precip	DJF	-1.19	21.2
North Hemisphere	Precip	JJA	-0.690	17.0
South Hemisphere	Temp	DJF	0.814	2.22
South Hemisphere	Temp	JJA	1.22	2.72
South Hemisphere	Precip	DJF	-10.1	6.34
South Hemisphere	Precip	JJA	-14.6	9.10
Tropics	Temp	DJF	1.16	2.66
Tropics	Temp	JJA	1.21	2.78
Tropics	Precip	DJF	-0.855	4.73
Tropics	Precip	JJA	-0.938	4.82
North Hemisphere Extra-tropics	Temp	DJF	2.66	4.84
North Hemisphere Extra-tropics	Temp	JJA	1.58	3.75
North Hemisphere Extra-tropics	Precip	DJF	-0.164	15.7
North Hemisphere Extra-tropics	Precip	JJA	-0.00845	8.37
South Hemisphere Extra-tropics	Temp	DJF	0.751	2.01
South Hemisphere Extra-tropics	Temp	JJA	1.12	3.00
South Hemisphere Extra-tropics	Precip	DJF	-1.28	6.77
South Hemisphere Extra-tropics	Precip	JJA	-0.223	8.30
Global	Temp	DJF	1.64	2.80
Global	Temp	JJA	1.59	2.73
Global	Precip	DJF	1.31	5.48
Global	Precip	JJA	0.656	5.62

Table 8.3: The minimum and maximum change for DJF and JJA seasons for temperature and precipitation for the 2 degree set.

although it is not clear whether it will get wetter or drier.

The regional distributions, in general, show a wider range of behaviour in temperature and greater uncertainty in the sign of precipitation response than the histograms for larger areas. These results imply that that decision makers should not assume a robust relationship between the extent of global warming and regional precipitation response.

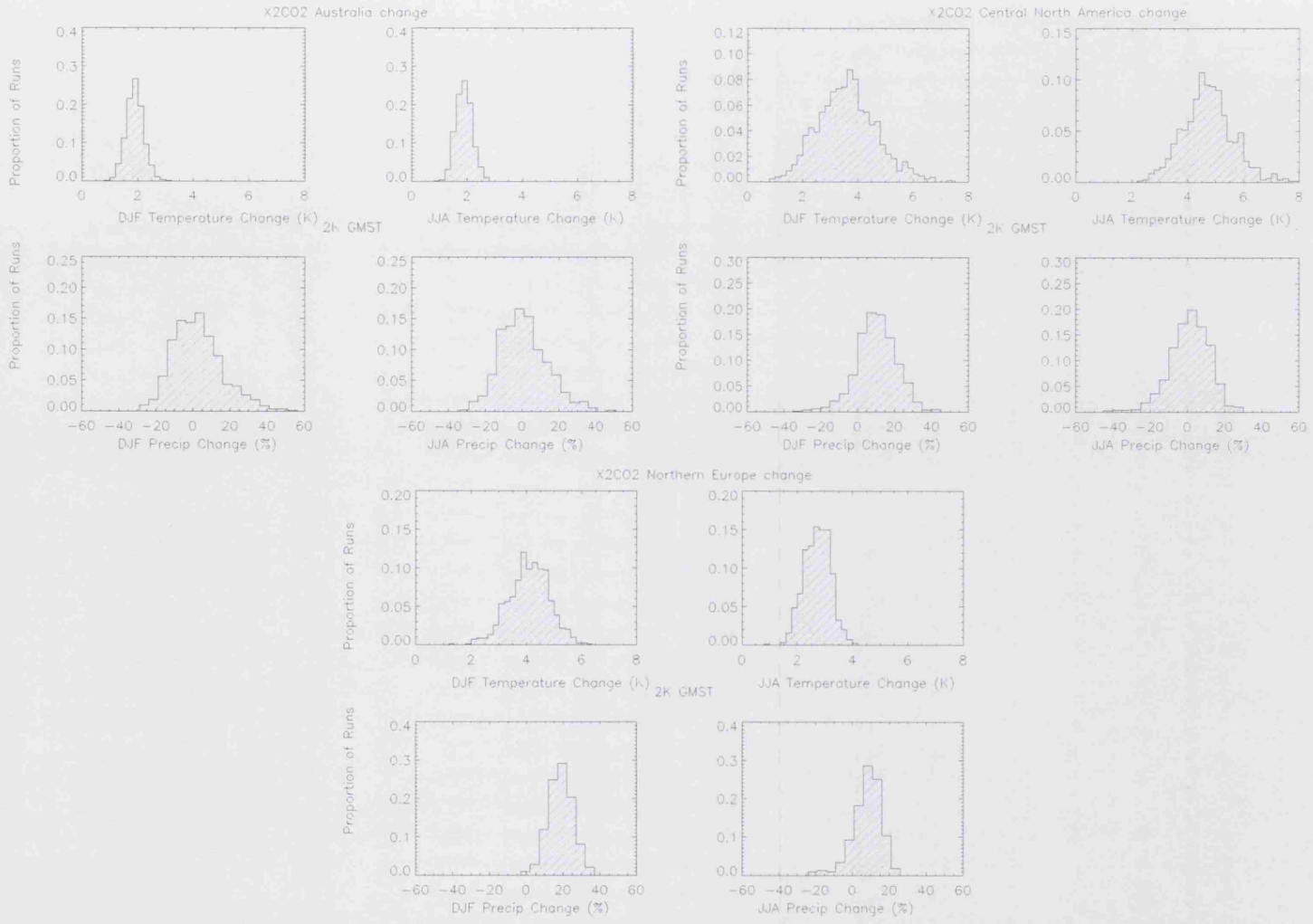


Figure 8.3: The distribution of 8 year DJF (JJA) temperature (precipitation) is shown for the 2 degree set. Whilst all simulations show an increase in surface temperature, the ensembles disagree on the sign of precipitation change in all regions and seasons shown.

8.2.3 Grid–scale Impacts

The impact of a 2 degree rise in temperature is important on scales as fine (and finer) as the HadSM3 grid resolution, as well as on regional scales. A land grid–box is typically of the order of $10,000 km^2$ in area for the HadSM3 model. Although this resolution is insufficient for many detailed impact studies, grid–box information is used in downscaling or impact assessment model, potentially introducing further uncertainties. Such downscaling procedures are only useful for guiding decision–makers when climate changes are simulated robustly on the grid–box level.
Figure 8.4 shows the median and size of the 10% – 90% central interval (this central range was chosen to give a conservative estimate of the diversity regional change and exclude any potential outliers that might have unphysical regional responses) for the 2 degree set for 8 year mean seasonal temperature change. The median change fields (DJF in panel (a) and JJA in panel (b)) show three features obscured when using GMST as the basis for decision–making:

1. Warming over land (~3–5 degrees) is greater than over the oceans (~1–3 degrees).

2. The magnitude of regional change varies considerably and, for the two degree set, simulations show up to 8 degrees of warming in Northern high latitudes' DJF season.

3. There is considerable variability with season. The most marked change is in the Northern high latitudes, showing ~8+ degrees warming in the DJF season and almost no warming in the JJA season.

The width of the central 80% interval is shown in panels (c) and (d) of Figure 8.4. The width of this interval is typically 1–2 degrees over the oceans and 1–4 degrees over land. Northern high latitudes, where the greatest warming occurs, show a range of up to 6 degrees in the DJF season. The model diversity of up to 4 degrees in simulated land temperatures, based on the 2 degree set, is considerable.

The magnitude of grid–box level variability is further examined in Figure 8.7 in Section 8.3.2.

There is a large difference between the percentiles shown in panels (c) and (d) Figure 8.4 in terms of impacts; differences that are invisible when using a global mean. The range of variability present within each grid–box is much larger than for regional means – indeed, statistically, this must be case. It should also be remembered that the fields shown in Figure 8.4 represent contributions from many different simulations and thus the patterns of change may not be consistent with any individual model simulation.

There is more variability at the grid–box level than at the regional level. Information on fine spatial levels is often highly uncertain, especially for precipitation. This suggests that decision–makers should consider the possibility that large regional changes are consistent with apparently low levels of global mean change e.g. panel (b) of Figure 8.4 shows a median Summer temperature response over the USA of 4–6 degrees Celsius for the 2 degree set. If these model results are taken to be indicative of the real world, adaptation strategies would be necessary even under modest levels of global warming.

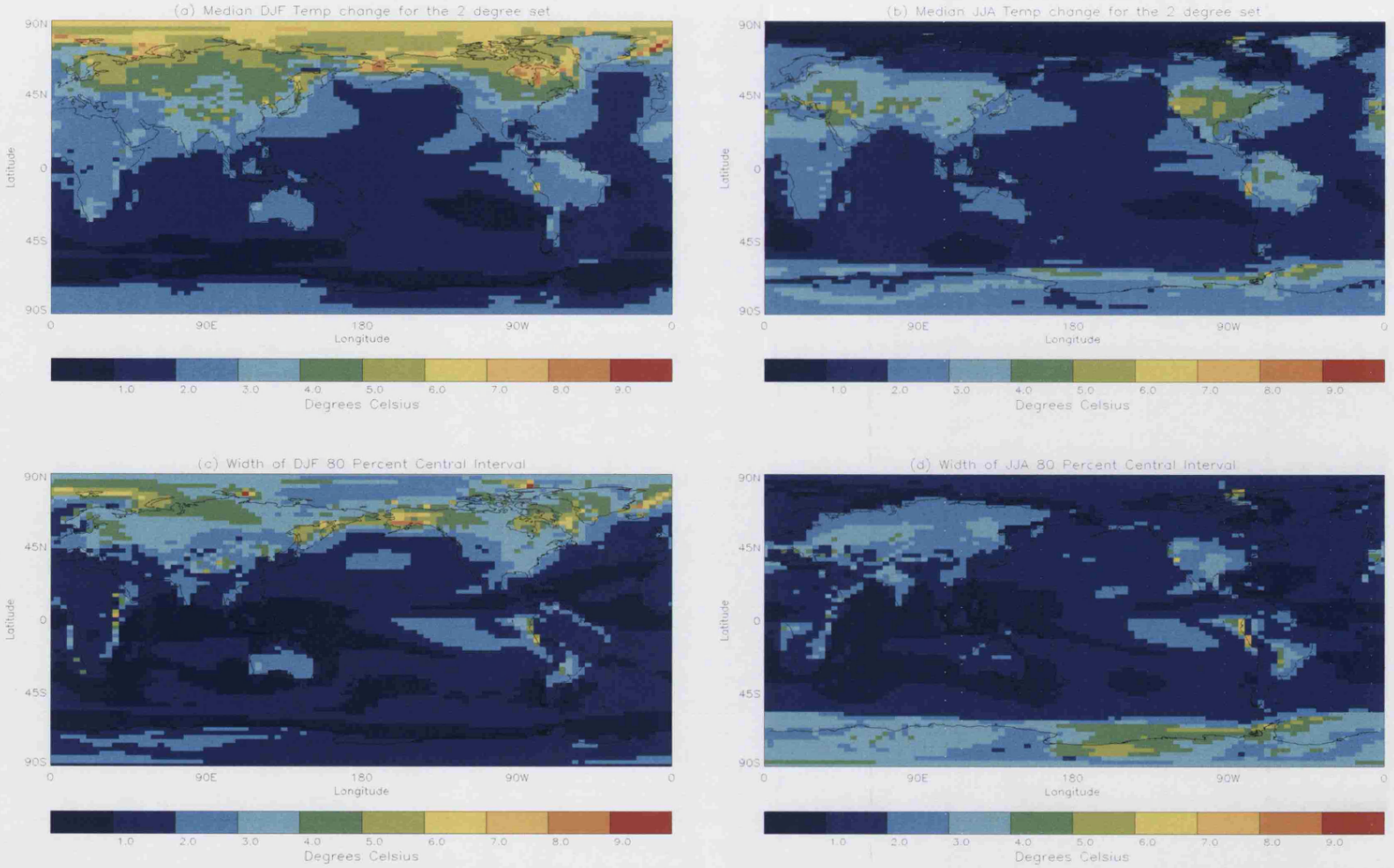


Figure 8.4: Median temperature change for the 2 degree set for the DJF (panel (a)) and JJA (panel (b)) seasons. Also shown in panels (c) and (d) are the widths of the central 80 percent (10th–90th percentiles)of temperature change, respectively for the DJF and JJA seasons.

8.3 What is the difference between 2 and 3 degrees GMST?

8.3.1 Regional differences

Model diversity in the 2 degree set has been looked at from global to grid–box length scales. In this Section, a comparison is made between the 2 degree set and the 3 degree set. Our ability to distinguish between simulations with different GMST rise is often implicitly assumed in impact studies Lynas (2007); Solomon *et al.* (2007a); Stern (2006). There is an important question to what level we can distinguish between such scenarios of GMST rise. Here, the question is asked "What is the difference between the 2 degree set and the 3 degree set on a regional level?"

It is shown in this Section for the first time that there are regional changes in temperature and precipitation consistent with both the 2 and 3 degrees sets.
For each of the two different sets of simulations considered (the 2 degree set and the 3 degrees set) there is a distribution of temperature and precipitation changes at each region, as shown in Section 8.2. Where these distributions overlap, there is a chance that a randomly selected simulation from the 2 degree set will be warmer (or wetter) than a simulation randomly selected from the 3 degree set. The size of this overlap places a limit on the value of GMST as a target for mitigation. Where this overlap is significantly greater than 0 GMST rise is a less effective tool for policy–makers.
The probability of an overlap between sets of simulations with GMST rise of 2 and 3 degrees Celsius in region r is calculated using a bootstrap Efron & Tibshirani (1994) method as follows. One simulation is selected at random from the 2 degree set and one at random from the 3 degree set and it is recorded whether the simulations from the 2 degree set is warmer than the simulation from the 3 degree set. This process is repeated 100000 times to estimate the probability of the overlap between the 2 degree set and the 3 degree set. Repeating this process 100 times in no case

gave a difference of greater than 1% in the estimates presented. This procedure was carried out for DJF and JJA temperature and precipitation (in the precipitation case the probability of a simulation from the 2 degree set being wetter than the 3 degree set was estimated).

The probability that a randomly selected simulation from the 2 degree set will be warmer(wetter) than a randomly selected simulations from the 3 degree set for each region and season is shown in Tables 8.4 and 8.5. A value of 0 (or 100) indicates that there is no overlap between these distributions. A value close to 50 indicates that there is no evidence to suggest that a simulation from the 3 degree set is likely to be warmer (wetter) than a simulation from the 2 degree set. Table 8.4 shows that the overlap between distribution varies by region e.g. in Greenland there is a 5% or less chance of a simulation randomly chosen from the 2 degree set being warmer than a simulation randomly chosen from the 3 degree set, whereas in South Asia there is roughly a 25% chance. Table 8.5 shows the probability of a randomly selected simulation from the 2 degree set being wetter than a randomly selected simulation from the 3 degree set is close to 50%. For some regions, such as Central Asia there is more chance that a 2 degree simulation will be wetter than a 3 degree in the DJF season but drier in the JJA season. For precipitation there is typically a significant probability of overlap – it is difficult to tell whether regional precipitation will be wetter for simulations from the 2 degree set than the 3 degree set. The presence of overlaps in these distributions puts limitations on our ability to assess the potential impact of climate change in terms of GMST rise.

It might be argued that the magnitude of overlap seen in Tables 8.4 and 8.5 can be reduced if only simulations that re-produce the current climate "accurately" are included and the range of behaviour might be reduced if a sub–set of simulations that perform well in–sample are analysed. In order to test this hypothesis, the probability of overlap was calculated for simulations with a low in–sample RMSE score in 7 different variables as calculated in Chapter 7, Section 7.5, calculated in

a similar way to Stainforth *et al.* (2005), by comparing model output in the 8 year mean field for a particular variable with observations. Such an ad hoc filter is not recommended as means to reduce model diversity as discussed in Chapter 7; results are shown here simply to demonstrate that this filter does not have a significant effect on results even if it were statistical meaningful.

For each variable, the probability of overlap was calculated for 10% of simulations within the lowest RMSE score. The probability of overlap was also found using the mean RMSE score over the 7 variables (calculated as the arithmetic mean of the RMSE in all variables). These probabilities are shown in Tables 8.6 and 8.7. There is no clear indication that constraining by in–sample RMSE will significantly alter the magnitude of overlap. Tables 8.6 and 8.7 show that the existence of an overlap is not sensitive to the accuracy of simulations in re-producing present day climate.

Region	% 2° > 3° (DJF)	% 2° > 3° (JJA)
Australia	5	6
Amazon Basin	8	14
Southern South America	4	10
Central America	14	12
Western North America	9	6
Central North America	20	23
Eastern North America	16	11
Alaska	15	4
Greenland	5	3
Mediterranean Basin	6	10
Northern Europe	11	5
West Africa	4	9
East Africa	9	10
Southern Africa	5	4
Sahara Region	5	2
South East Asia	6	14
East Asia region	16	16
South Asia region	25	23
Central Asia region	9	11
Tibet	13	11
North Asia region	11	9
Antarctica	13	9

Table 8.4: TEMPERATURE. The probability (in %) that a randomly selected simulation from the 2 degree set showing more warming, for regional DJF temperature, than a randomly selected simulation from the 3 degree set.

Region	% 2° > 3° (DJF)	% 2° > 3° (JJA)
Australia	36	63
Amazon Basin	39	75
Southern South America	37	56
Central America	64	67
Western N.America	37	46
Central N.America	41	73
Eastern N.America	45	66
Alaska	29	23
Greenland	11	33
Mediterranean Basin	46	71
Northern Europe	28	63
West Africa	65	55
East Africa	61	52
Southern Africa	59	47
Sahara Region	59	63
South East Asia	41	59
East Asia region	67	52
South Asia region	68	40
Central Asia region	51	38
Tibet	56	22
North Asia region	17	21
Antarctica	12	7

Table 8.5: PRECIPITATION. The probability (in %) that a randomly selected simulation from the 2 degree set rise simulates a greater increase in regional DJF precipitation, than a randomly selected simulation from the 3 degree set.

Region	NONE	Temp	SSP	Precip	Cloud	SS HF	SL HF	HF LS	Mean
Australia	5	5	3	3	6	6	5	6	5
Amazon Basin	8	10	12	12	6	8	8	9	9
Southern S.America	4	4	3	4	3	4	4	6	6
Central America	14	18	16	16	15	14	14	19	16
Western N.America	9	10	8	8	8	9	7	8	8
Central N.America	20	31	19	26	19	27	24	23	29
Eastern N.America	16	21	13	16	16	17	16	17	18
Alaska	15	11	17	10	15	13	15	10	17
Greenland	5	4	5	4	5	4	5	4	6
Mediterranean Basin	6	6	9	5	7	6	5	10	6
Northern Europe	11	12	8	10	13	11	12	14	12
West Africa	4	5	6	6	5	5	7	6	5
East Africa	9	14	13	13	8	13	14	12	11
Southern Africa	5	7	6	9	6	10	11	7	8
Sahara Region	5	6	10	8	5	7	8	7	6
South East Asia	6	11	12	10	8	12	11	10	8
East Asia region	16	21	20	17	14	17	17	18	18
South Asia region	25	35	33	38	27	37	35	34	30
Central Asia region	9	9	11	8	10	9	10	9	7
Tibet	13	18	20	15	14	18	17	13	13
North Asia region	11	10	9	11	14	9	10	13	10
Antarctica 13	21	18	16	11	19	18	12	17	

Table 8.6: Probability (%) of a hotter region for simulations from the 2 degree set than the 3 degree set. Values are shown where both sets are constrained by in-sample RMSE in each of 7 different variables, and an aggregate over these 7. The constraining variables shown are Temperature, Sea Surface Pressure, Total Precipitation Rate, Total Cloud Amount, Sea Sensible Heat Flux, Sea Latent Heat Flux, Heat flux latent surface and the Mean Score over these 7 variables. Results are shown for the DJF season.

Region	NONE	Temp	SSP	Precip	Cloud	SS HF	SL HF	HF LS	Mean
Australia	36	42	38	39	39	40	37	42	38
Amazon Basin	39	37	37	36	40	38	42	37	40
Southern S.America	37	34	34	32	31	34	33	38	32
Central America	64	60	67	57	60	63	64	60	65
Western N.America	37	37	31	35	38	37	41	31	33
Central N.America	41	43	41	42	34	42	44	42	41
Eastern N.America	45	38	49	47	49	39	41	41	37
Alaska	29	29	33	23	29	30	30	32	33
Greenland	11	11	11	9	11	9	11	11	12
Mediterranean Basin	46	41	50	46	46	45	46	49	46
Northern Europe	28	32	28	27	28	32	31	29	28
West Africa	65	66	66	63	62	62	65	66	66
East Africa	61	61	58	62	58	60	61	61	64
Southern Africa	59	68	58	54	60	62	56	62	64
Sahara Region	59	56	63	62	53	60	63	56	62
South East Asia	41	36	41	38	39	36	37	42	41
East Asia region	67	72	71	70	68	73	71	72	75
South Asia region	68	72	70	76	67	77	74	72	70
Central Asia region	51	55	50	53	49	52	58	48	54
Tibet	56	58	56	61	53	58	62	52	58
North Asia region	17	19	20	18	18	15	15	18	14
Antarctica	12	13	11	8 0	12	10	10	19	12

Table 8.7: Probability (%) of a wetter region for simulations from the 2 degree set than the 3 degree set. Values are shown where both sets are constrained by in-sample RMSE in each of 7 different variables, and an aggregate over these 7. The constraining variables shown are Temperature, Sea Surface Pressure, Total Precipitation Rate, Total Cloud Amount, Sea Sensible Heat Flux, Sea Latent Heat Flux, Heat flux latent surface and the Mean Score over these 7 variables. Results are shown for the DJF season.

It has been shown in this Section that there are regional changes in temperature and precipitation consistent with both the 2 degree and 3 degree sets. Next, distributions of regional temperature and precipitation response are looked at in more detail. The magnitude of regional change is now shown not to be robust in some regions in either temperature or precipitation.

In Figure 8.5 the distribution of simulations from the 2 and 3 degree sets are compared for the three regions used in Figure 8.3. The distribution of simulations from the 2 degree set is shown in blue and the 3 degree set in red. There is some overlap for all regions, seasons and variables shown. The distributions of temperature are more distinct (the overlap in JJA temperature for Australia, Central North America and Northern Europe are 6%, 23% and 5% respectively compared to 63%, 73% and 63% for precipitation), with higher regional warming seen in all regions in the 3 degree set. In Australia, the extra one degree of GMST rise leads to a shift in the distribution of about 1 degree to the right, whereas in Central North America and Northern Europe, the difference between the distributions is a shift of almost 2 degrees. Such shifts in distribution are best interpreted alongside the associated uncertainties and their overlaps. It is interesting to note that in the Northern European region, warming is more extreme in the DJF season than in the JJA season, whilst in Central North American warming appears more extreme in the JJA season. Such seasonal differences can be important to decision makers e.g. it is thought that warmer summers will increase mortality rates though heat deaths, whereas mild winters reduce mortality rates though a reduction in cold–related deaths Parry *et al.* (2007). Thus the winter warming seen in Northern Europe might be considered less worrying than the summer warming simulated in the Central North American region. These results further suggests that studies on the seasonal impact of climate change based on one region would not translate to other regions.

It is apparent from Figure 8.5 that regions such as Central North America or North-

ern Europe must be prepared to adapt to more extreme regional warming than is expected on the global scale. This is a key result for decision–makers since it suggests that significant adaptation planning might be still necessary in the presence of more modest levels of global warming.

The distributions for precipitation are less distinguishable than for temperature. For Australia (both seasons) and the DJF season in other regions, the two distributions in Figure 8.5 almost indistinguishable (36% and 63% overlap in the DJF and JJA seasons respectively). It would be difficult, based on evidence from the HadSM3 model, to tell what difference a 3 hotter world would look like compared to a 2 degree hotter world. For DJF precipitation in Northern Europe, all simulations in the 3 degree set show an increase in precipitation. Although this result might not be robust when compared to other models or larger ensembles, the CPDN PPE provides evidence to support the assumption of an increase in precipitation (anywhere up to a 50% increase) for this region and season. The assumption of an increase in precipitation would not appear to be justifiable for say, JJA precipitation over Australia since the ensemble is not consistent on the sign of the change. Thus, it may not be necessary to develop different adaptation decisions reliant on regional precipitation change depending on a forecast changing between 2 and 3 degrees of warming – in either case we can not say whether precipitation will increase or decrease, and the distributions of regional behaviour are similar based on the simulations examined here.

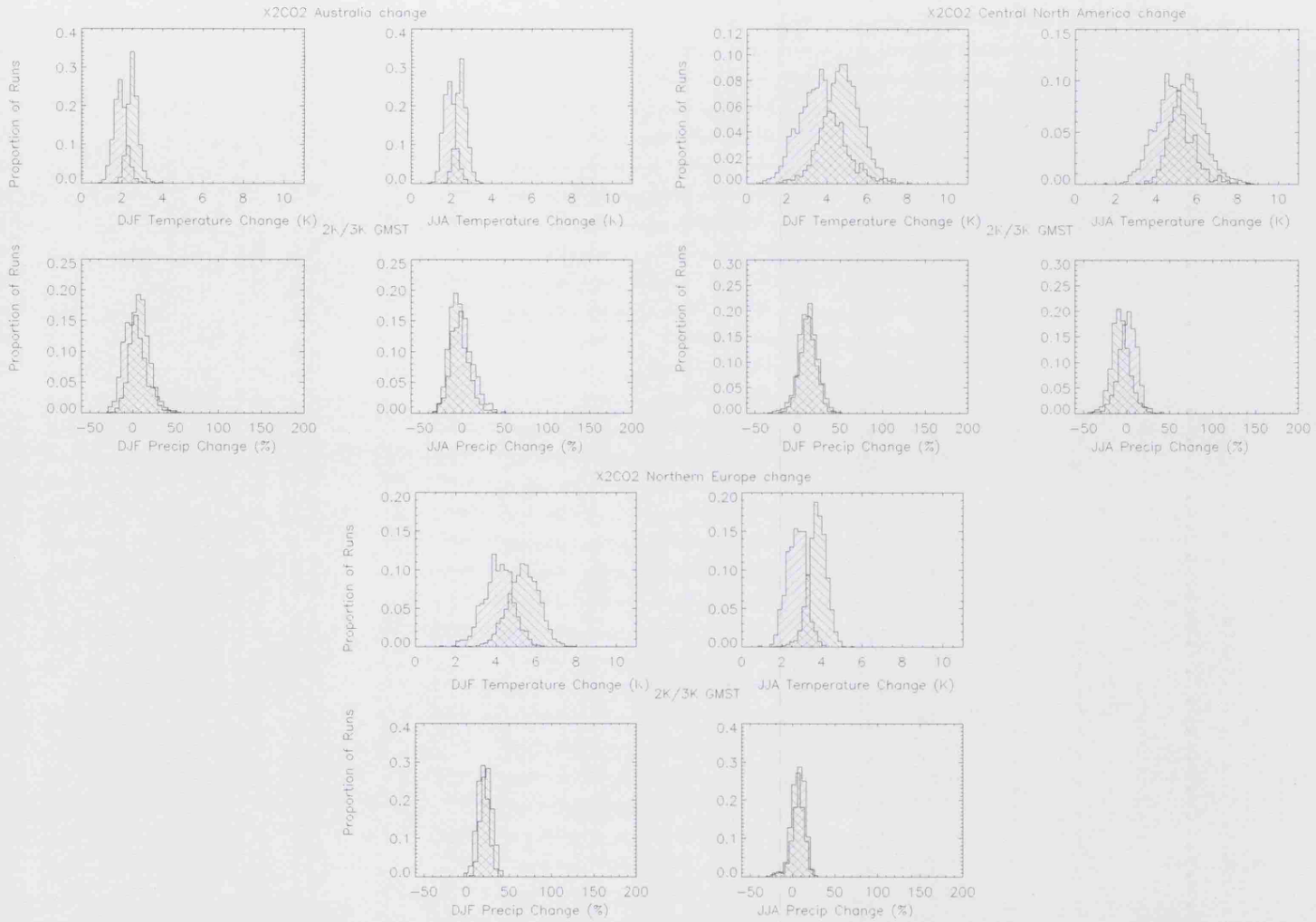


Figure 8.5: The distribution of 8 year DJF (JJA) temperature (precipitation) is shown for the 2 degrees set (blue) and the 3 degrees set (red). There is often a large overlap between the two ensembles, especially for precipitation. This shows that the 2 degree set and the 3 degree set are not always robustly distinguishable.

8.3.2 Grid–scale differences

Figure 8.7 shows histograms, similar to those shown on regional scales, for three selected grid–boxes; specifically those corresponding to the location of the cities of London (UK), Boulder (Colorado, USA) and Jakarta (Indonesia) (see Figure 8.6) although the grid–boxes are much larger than the cities themselves. The scale on the x–axis differs from Figure 8.5 in light of the increased range in precipitation seen at the grid–box level. Changes in temperature are again robustly positive for these three grid–boxes, with the 3 degree set typically showing greater warming than the 2 degree set, with a probability of overlap in DJF temperature of 12%, 22% and 6% for DJF temperature for the London, Boulder and Jakarta grid–boxes respectively. The upper end of warming in the boulder grid–box is over 8 degrees for the 3 degree set. Precipitation change is highly variable in the London and Boulder grid–boxes e.g. JJA precipitation change in Boulder ranges from a greater than 50% reduction to more than 100% increase for both sets of simulations. The Jakarta grid–box shows massively different responses in precipitation – from almost 100% decrease (almost no rainfall at all) to over 200% increase. Such variability is likely due to Jakarta lying in between rival patterns of change in the region, thus the extreme uncertainty does not suggest changes on a larger scale are unintelligible. The 3 degree set simulates precipitation to be wetter in summer and drier in winter, although there is much overlap in these distributions e.g. an overlap of 39%, 48% and 35% in DJF precipitation for the London, Boulder and Jakarta grid–boxes respectively. Note that Boulder shows a much wider range of values for temperature than Jakarta, but a much tighter distribution in precipitation. An area may show robust results in one variable and not another.

It is misleading to use such a wide range of model behaviour to drive downscaling models or impact assessment in areas such as Indonesia on the basis of these data. The most reasonable assessment of such information is that in many areas, on a grid–box level, the HadSM3 model versions used here do not allow us to make

statements about changing precipitation. In some areas and variables, models can rule nothing out.

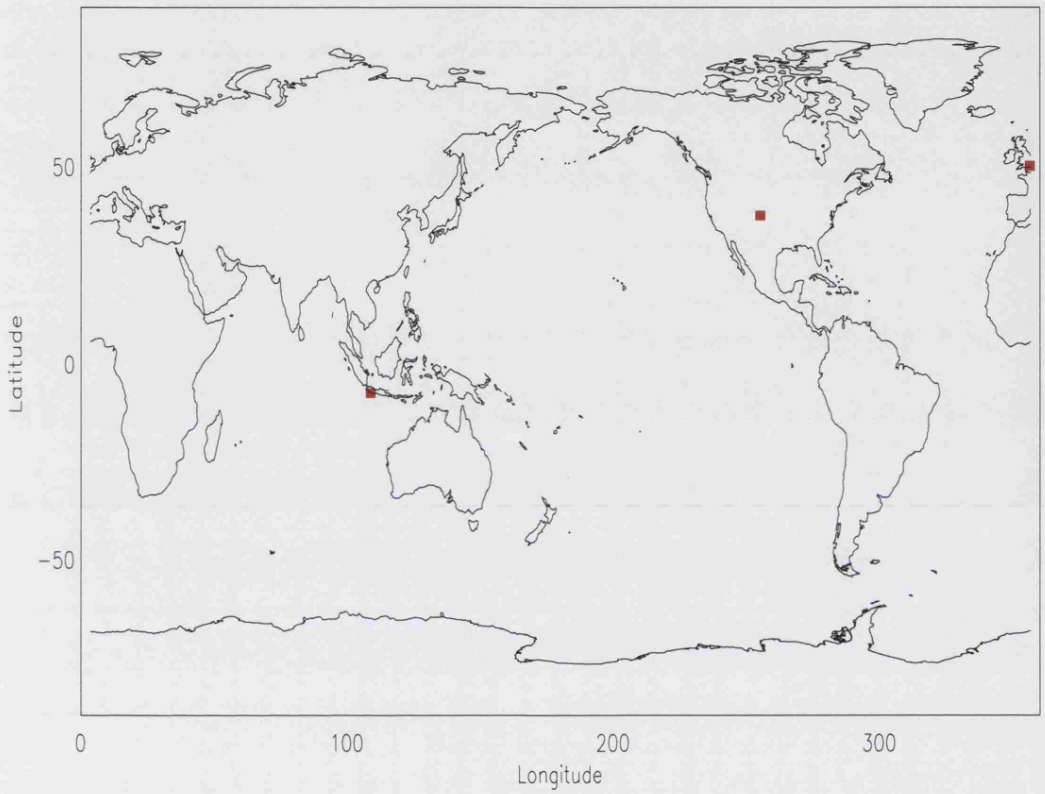


Figure 8.6: The three grid–boxes selected to look at local impacts are shown. These grid boxes are called London, Boulder and Jakarta since they contain those cities. It should be noted that the grid–boxes are much larger than the cities they contain.

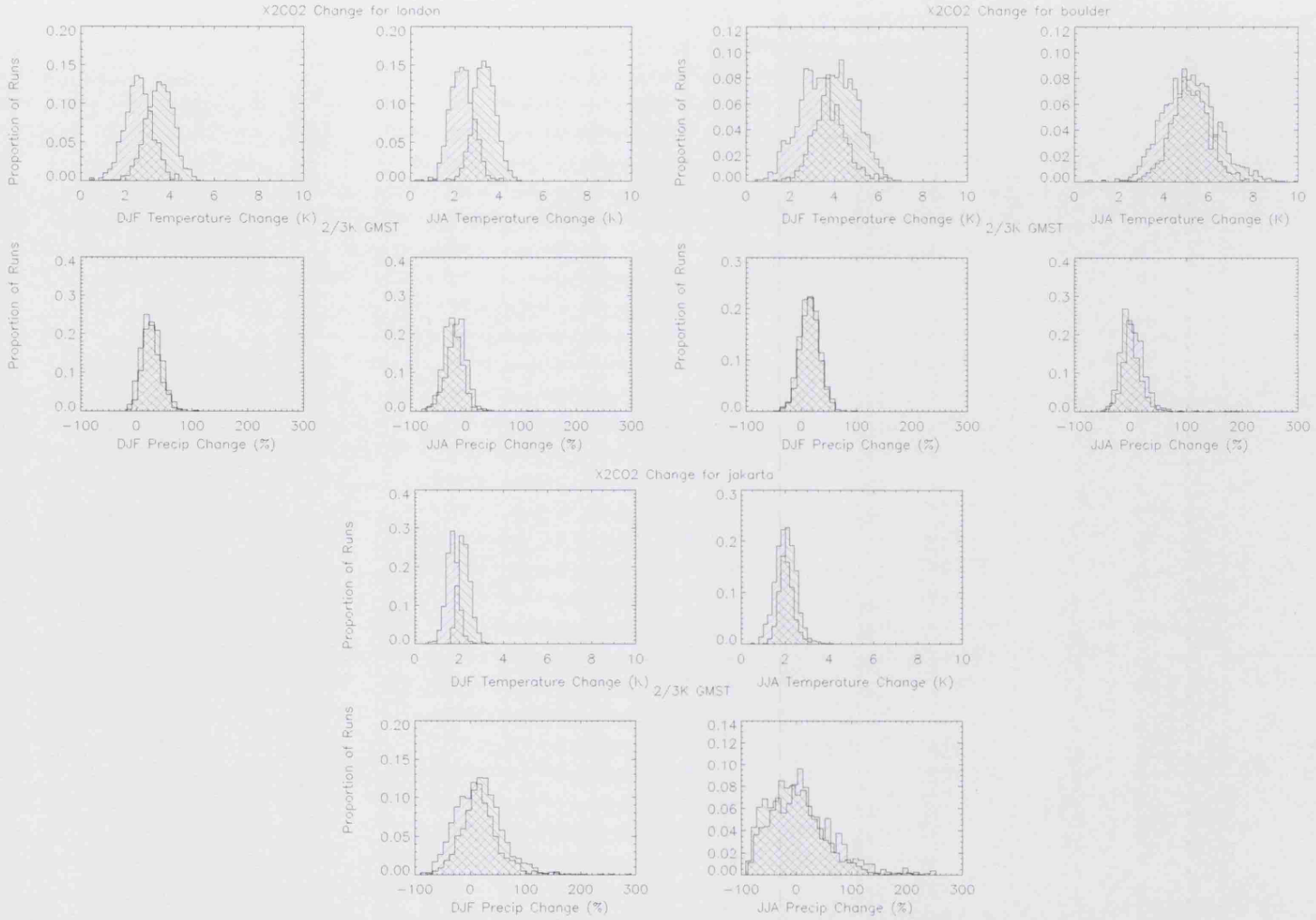


Figure 8.7: The distribution of 8 year DJF/JJA temperature/precipitation is shown for the 2 degree set (blue) and the 3 degree set (red). The grid–boxes that contain London, Boulder and Jakarta are shown. Note that the cities themselves are much smaller than the grid–boxes, which are typically 50,000km^2 in area.

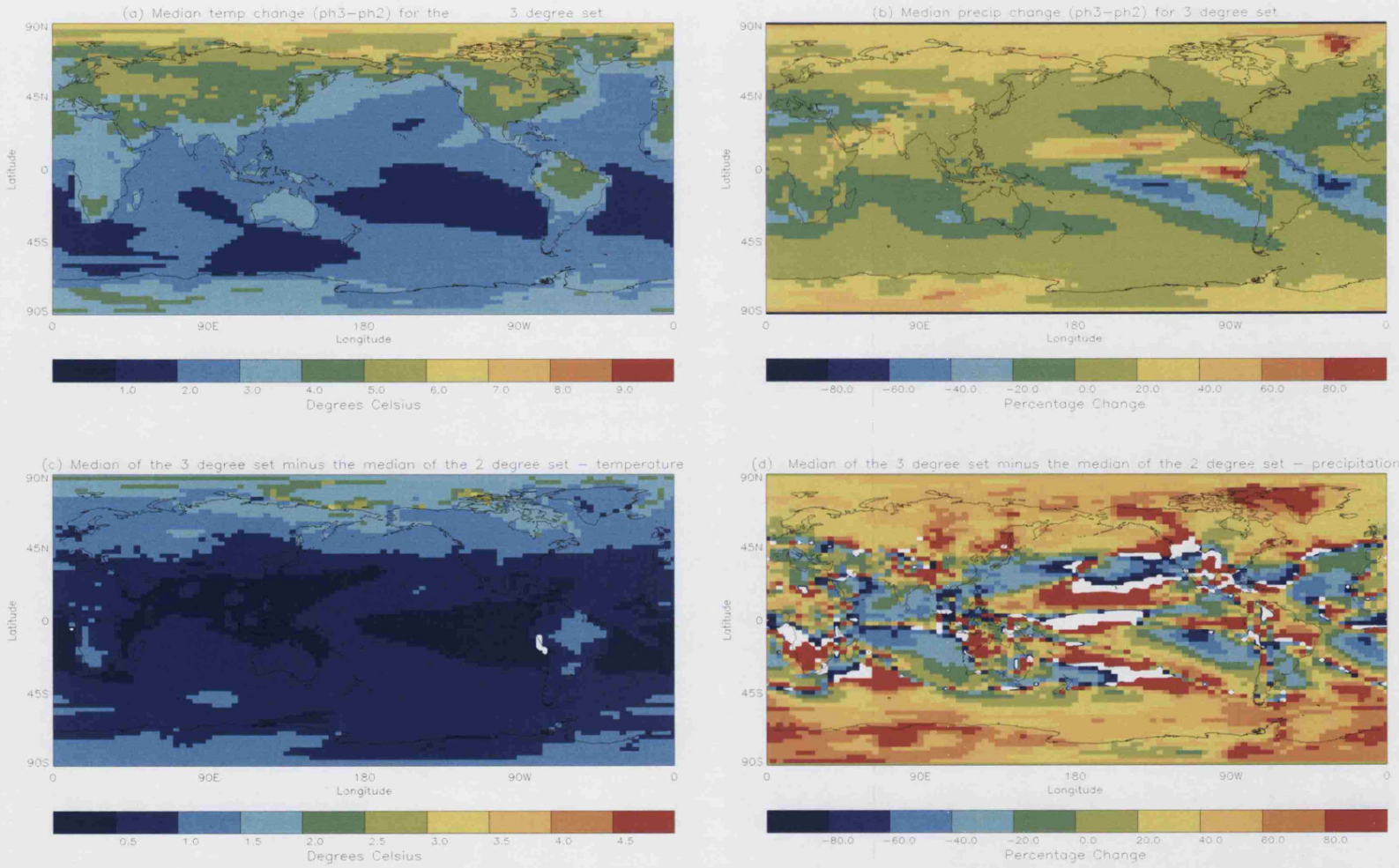


Figure 8.8: The median change in temperature (degrees Celsius) is shown in panel (a) and precipitation (mm per day) is shown in panel (b) over the 3 degree set. Also shown is the difference between this and the median temperature rise in temperature for the 2 degree set in panel (c) and precipitation in panel (d).

Having examined the variability within the 2 and 3 degree sets, Figure 8.8 shows the median change in GMST (panel (a)) and total precipitation rate (panel (b)) for the 2 degree set simulations and the difference between these simulations and the 2 degree set in panels (c) and (d). As expected from previous analyses, the regional distribution of warming shown in panel (a) is for oceans to warm the least (typically by 1–3 degrees), and the Arctic region the most (typically by 6–7 degrees Celsius). In general, inland regions warm by more than coastal regions. These patterns of regional temperature change for simulations with a fixed GMST change are consistent with the regional changes seen in the standard HadSM3 ensemble in Chapter 6 and across the CPDN grand ensemble in Chapter 7. Panel (c) shows the difference between the median warming for the 3 degree simulations minus the 2 degree simulations. There are few areas that show a negative value (negative values indicate the median in the 2 degree set is greater than the median in the 3 degree set) and considerable areas show a difference of less than 0.5 degrees, particularly in the ocean. The median of the 3 degree set is up to 3 degrees warmer than the 2 degree set in the Arctic where the most rapid warming occurs.

Panels (b) and (d) in Figure 8.8 show the same statistics, but for precipitation. It is notable that, unlike temperature, the median change in precipitation (panel (b)) under a doubling of CO_2 concentrations is negative over large areas. This variability in local response is most noticeable around the tropics; the equatorial pacific and the Indian sub–continent show an increase of precipitation by more than 1 mm per day, whereas the East coast of Brazil shows a fall of more than 1 mm per day. The kind of regional variety shown here is lost in a global mean, with many of the features in this field cancelling each other out. This, together with basic counting statistics, explains how a tight distribution of global mean precipitation is consistent with large uncertainties in local changes. An important conclusion from this analysis is that global mean values destroy impact–relevant information on regional and local scales. One example of this is that at the global mean level

a misleadingly small change in precipitation (about 0–5% increase) hides changes of 50% or more at the regional level and over 200% at the grid–box scale. These changes are vital for adaptation planning and suggest that global mean metrics are of limited use to decision–makers.

8.4 Linearity of Regional Response

It has been proposed that regional climate response to forcings scales linearly with GMST change e.g. "the geographical pattern of the temperature, precipitation or other response is assumed to be independent of the forcing, the amplitude of this fixed pattern being proportional to the global mean change" Ruosteenoja *et al.* (2007), and "Seasonal temperature and precipitation have been shown to scale approximately linearly with the magnitude of global warming when analysing ensemble average change signals from multiple models" Diffenhaugh *et al.* (2007) and references therein. Other references to the linearity of regional climate change in proportion to global mean temperature change are made in Giorgi (2008); Solomon *et al.* (2007a). Whilst the CPDN experiment does not allow a consideration of the assumption of linearity to different magnitudes of forcing since the same doubled CO_2 forcing scenario is used in all simulations, a similar question can be investigated – do regional responses scale linearly with GMST in the CPDN PPE? If this question can be answered in the positive it implies that it is possible to infer patterns of regional change based on GMST alone – this would save experimental resources (there is less need to store regional data or to explore the patterns of change within simulations with different values of GMST change) and would greatly increase the utility of global mean statistics as a basis for decision–makers. If there is evidence that this question should be answered in the negative it suggests that regional behaviour can be non–linear relative to global changes and that these regional responses can not be inferred directly from global means. Since the CPDN grand ensemble can be divided into sub–sets of simulations with similar levels of

GMST change, the robustness of regional response can be looked at across different model versions and values of GMST rise.

Regional response factors are defined in this Section, following Giorgi (2008), as the proportional change in regional climate response in seasonal temperature and precipitation with respect to GMST change. The regions used relate to the Giorgi regions Giorgi & Francisco (2000), as well as tropics, extra–tropics, Northern and Southern hemispheres. These regional response factors, R_i, for data set i and region R are calculated as follows: define the 8 year mean global mean temperature change from the doubled CO_2 phase to the control phase as Δ_G and the regional change (in temperature, or precipitation) as Δ_R. R_i is estimated by the ensemble mean regional change proportional to the ensemble mean global mean temperature change. This regional factor is calculated here for four different sets of data with different values of GMST change – the standard HadSM3 model ICE and three sets of simulations with global mean temperature changes of 2±0.1, 3±0.1, 4±0.1 degree with 64, 402, 2441 and 795 available simulations respectively. The regional response for set i is defined as:

$$R_i = \frac{\Delta_{R,i}}{\Delta_{G,i}} \tag{8.1}$$

where $\Delta_{R,i}$ denotes the ensemble mean change in region R for data set i for $i = 1, 2, 3, 4$ and $\Delta_{G,i}$ denotes ensemble mean GMST change. These regional response factors are calculated for 8 year mean DJF/JJA temperature (panels (a) and (b)) and DJF/JJA precipitation (panels (c) and (d)). Figure 8.9 shows these regional response factors for four sets of data, based on the 64 member standard HadSM3 model and sub–sets of simulations with 2, 3 and 4 degrees of GMST change, respectively.

Estimates from the standard HadSM3 ICE are shown in black, from simulations with 2 degrees GMST change in blue, 3 degrees GMST in green and 4 degrees in red. Vertical black bars are shown for each region and set of simulations, denoting

Region Code	Region
aust	Australia
amsa	Amazon Basin
ams	Southern South America
amc	Central America
amnw	Western North America
amnc	Central North America
amne	Eastern North America
amal	Alaska
grnl	Greenland
eum	Mediterranean Basin
eun	Northern Europe
afw	West Africa
afe	East Africa
afs	Southern Africa
afsh	Sahara Region
asse	South East Asia
ase	East Asia
asso	South Asia
asc	Central Asia
astb	Tibet
asn	North Asia
ant	Antarctica
hmn	North Hemisphere
hms	South Hemisphere
trop	Tropics
hmnt	North Hemisphere Extra-tropics
hmst	South Hemisphere Extra-tropics

Table 8.8: Region codes and names for the 27 areas used to calculate regional response factors.

± 2 standard deviations in the estimate of the mean regional response. Regions are denoted using the table of region codes shown in Table 8.8.

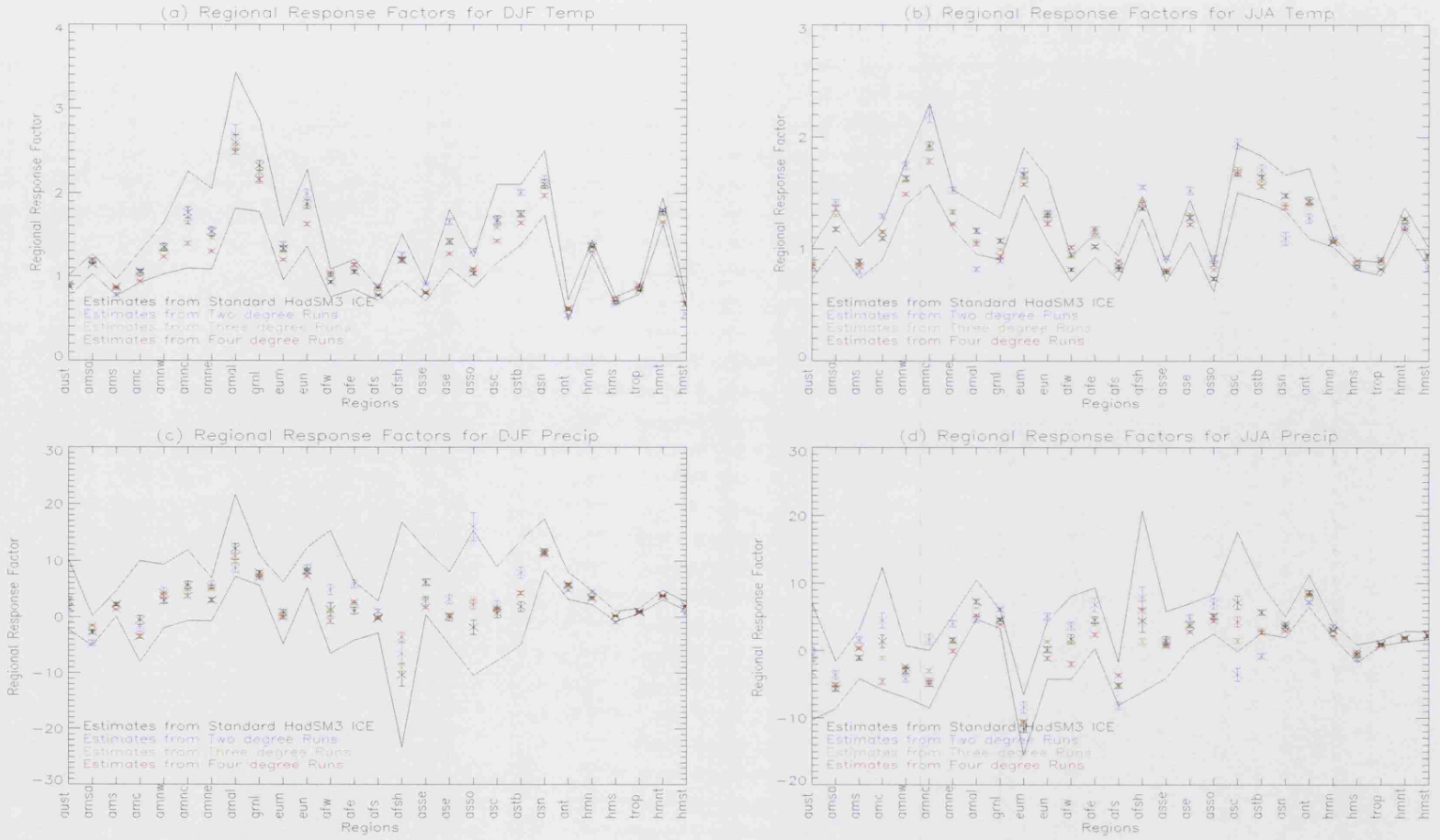


Figure 8.9: Regional response factors for 27 regions shown in Table 8.8 in four different variables. DJF temperature response factors are shown in panel (a), JJA temperature in panel (b), DJF precipitation in panel (c) and JJA precipitation in panel (d). Estimates from the 2 degree set of simulations are shown in blue, the 3 degree set in green and the four degree set in red. Estimates from the individual simulations from the standard HadSM3 ICE are shown in black. The two black lines show the minimum and maximum regional response in simulations from the standard HadSM3 ICE. Vertical bars show 2 standard deviations in estimates of the mean regional response from each set.

Statistically, simulations with different values of GMST rise produce different estimates of the mean regional response rates (in many regions the estimates differ by over 2 standard deviations). The statistical significance is largely due to the low standard deviation of estimates in the mean since large sample sizes are used to form these estimates. It should be noted that even apparently small differences in estimates of regional response factors will multiply when estimating regional response for higher GMST change e.g. a difference of 1 in regional response factor estimates between *amnc* (Central Northern America) temperature response (panel (a)) relates to a 5 degree Celsius differences in estimated regional temperature under 5 degrees of global warming. For the larger areas representing the tropics, extra–tropics and hemispheric means the response factors are more consistent than for smaller areas. In precipitation (panels (c) and (d) of Figure 8.9)the regional response factors estimated from individual simulations in the standard HadSM3 model do not show a consistent sign of response (the upper and lower black lines often cover the lines $Y = 0$). Furthermore, estimates of regional response from the different sets do not always agree on the sign of regional response. Regional precipitation change can not be robustly estimated where different sets disagree on the sign of regional response factors.

The results in this Section show evidence that the assumption of linear regional response to global mean changes is not relevant for precipitation response and there are significant uncertainties in the estimation of regional temperature response in many regions. These results show that GMST can not be assumed to be related in a linear way to regional responses in temperature and precipitation, although non–linearities are larger in some regions than in others.

8.5 Discussion

Even when model GMST is known, significant regional "uncertainties" remain, particularly for precipitation. In all the regions (of area comparable to or larger than

many nations) presented it is not clear whether precipitation will increase or decrease for the 2 degree set.

The difference between the 2 and 3 degree sets is not clear for precipitation – in some regions and seasons the distribution shifts further from 0 (wetter or drier than the 2 degree ensemble), in others it shifts towards 0 (less significant change than for the 2 degree ensemble). This may suggest a non–linear "dose–response" to rising temperatures in the model. For example, in Figure 8.5, JJA precipitation in Northern Europe increases under the 2 degree set, then shifts back towards 0 for the 3 degree set. This result may be peculiar to this model, or a robust feature over state–of–the–art climate models in general. It is difficult to compare such distributions to other models since there are no other suitably large ensembles amenable to the approach used in this Chapter[1]. Nonetheless, it is clear that the distributions shown in this Chapter should not be over interpreted and may contain model–specific features or biases.

GMST is commonly used in policy discussion as the key index for climate change Parry *et al.* (2007); Stern (2006). The use of GMST change as an index for climate change impacts is, at best, highly restricted and can even provide misleading information[2]. New approaches to communicating the results of large ensemble experiments would help motivate more robust impact assessment and aid better policy formulation.

The results presented in this Chapter also pose questions relevant to those interested in constraining model output. Whilst constraining GMST changes might be useful for reducing certain uncertainties, significant regional uncertainties remain. Attempts to constrain GMST are thus more useful for mitigation strategies that operate typically on global length scales (such as targets for global carbon emissions) than regional or local adaptation decisions.

[1]To form a distribution for a narrow range of GMST either requires a very large ensemble of simulations or a specially designed experiment.

[2]In some model regions, a 2 degree increase in GMST might lead to over 6 degree increases in temperature or more. Thus, limiting GMST to a certain level may not restrict regional warming to sustainable levels.

The data presented in this Chapter has been in terms of an 8 year seasonal mean. Finer temporal scales would provide more relevant for many impact studies in order to evaluate an increased risk of flooding, droughts, heat waves etc. Such data is unavailable for the CPDN data set but, were any future large ensemble experiments to record such information[1], the methods presented in this Chapter would be useful for assessing how robust certain impact studies might be. It is only with large ensembles of climate models that the tails of distributions can be evaluated. Small ensembles do not allow robust statements regarding the tails of distributions.
The implications of the results presented in this Chapter are now discussed in terms of mitigation and adaptation decisions.

8.5.1 Mitigation

There is much uncertainty in how GMST will respond to increasing concentrations of GHGs Murphy *et al.* (2004); Stainforth *et al.* (2005). GMST and CS are often used as an index of this change. According to the IPCC AR4 Solomon *et al.* (2007a), the likely GMST rise over the next 100 years will be between 1.5 and 4.5 degrees Celsius (range of best estimates from 6 different emissions scenarios), with a best guess of 3 degrees. Studies suggesting that increases in GMST will have serious social and economic costs Parry *et al.* (2007); Stern (2006) have prompted discussion as to what is an acceptable level of warming for a given cost of mitigation measures. The extent to which mitigation policy relating to global mean temperature will limit regional changes has been shown to be unclear in this Chapter. There are uncertainties in determining how policy will relate to emissions, how emissions will relate to GHG concentrations, how quickly GHG will warm the Earth and how GMST relates to climate impacts. It is this last link that is dealt with in this Chapter e.g. if GMST were to rise by 2 degrees Celsius, what would this mean

[1]Studies, such as those contributing to the IPCC Fourth Assessment Report, do record such data but the ensembles typically consist of at most 9 members, too few to carry out the analysis presented in this Chapter.

for temperature and precipitation changes in different regions? Under the scenario of a 2 degree GMST rise, regional impacts and uncertainties therein have been analysed using data from the CPDN experiment. The results from this Chapter show that mitigation of regional risks is difficult given the rapidity of simulated warming in some regions (typically over twice as fast as GMST rise in Northern high latitudes) and the significant uncertainties in regional response. For a GMST rise of 2 degrees Celsius, there can remain significant risks of exceeding 6 degrees or more of regional temperature change. It should be noted that this Chapter has only considered GMST rises of up to 4 degrees Celsius, whilst the CPDN PPE does not rule out values of CS of over 10 degrees Celsius. Regional responses from high CS simulations are not studied here since high CS simulations have not reached an approximate equilibrium by the end of the experimental phase[1].

8.5.2 Adaptation and Impact Assessment

It has been shown in this Chapter that knowing GMST change can not be robustly translated into regional changes in temperature and precipitation. In many regions, in key variables, there are large uncertainties for a given GMST change, hence it would be misleading to suggest that certain regional impacts will occur at a particular GMST rise. For example, Table 8.1 shows that for a 2 degree GMST rise, the range of Greenland temperature change is from 1.98 to 6.78 degrees Celsius in the winter season. Such differences in temperature could lead to very different futures for the Greenland ice sheet and highlights the need to consider the regional variability associated with GMST rise. Such events can occur at a variety of different GMST increases although risk of regional impacts can change with GMST. More generally, local decisions based on GMST would need to be robust to the diversity of behaviour seen here, even assuming the HadSM3 model used to generate these

[1] 100% of simulations in the 2 degree set have 8 year GMST rise of more than 90% of CS. This percentage drops to 70% and 15% for the 3 and 4 degree sets and 0% for simulations with over 5 degrees of GMST rise, as explained in Section 4.4.1 and shown in Figure 4.6

distributions is perfect.

It has been shown in this Chapter is that different regions can require very different adaptation strategies. HadSM3 suggests that some regions are likely to face far more extreme climate changes that others e.g. Figure 8.5 has shown that simulated warming of up to 10 degrees in Central Northern America is consistent with 5 degrees of global warming, whilst Australians would expect half this level of warming. It is vital, when planning adaptation strategies, to consider the regional and seasonal differences in climate change. An important result in this Chapter is that a wide range of regional responses can be consistent with the same variable, season and GMST change. This suggests that adaptation strategies should be flexible in order to account for a wide range of regional changes in climate even when GMST is assumed to be known.

Presenting a distribution of simulations, as in this Chapter, also provides decision makers with some (albeit rudimentary) idea of model diversity; it is important to consider not only expected changes but the the chance of extreme levels of regional warming (e.g. 6 degrees or greater).

It has been shown that GMST can not be used to robustly infer regional changes.

8.6 Conclusion

Data from the CPDN experiment has been used here to look at regional model diversity for a given level of global warming. The large range of regional behaviour present suggests that GMST has limited utility as an index for adaptation policy and impact assessment. Using an ensemble of simulations as a distribution, rather than simply taking a mean, and looking at changes on the length scales relevant to impact studies provides a more relevant framework for making robust statements of climate change.

New results presented in this Chapter are:

- Regional changes can differ significantly (over 6 degrees Celsius for some re-

gions in 8 year mean seasonal temperature) for simulations with the same global mean temperature change.

- The spatial scales of model diversity have been quantified using the CPDN PPE. The magnitude of regional uncertainties for a given GMST change has been shown to vary with length scale and variable. The distribution of regional change has been used to present uncertainties in sub–global response on a variety of length scales - global, hemispheric, tropical and extra-tropical, regional and local. Uncertainties in precipitation are large – the sign of change is uncertain on length scales as large as many nations in most regions looked at.

- The distributions of regional change have been contrasted between the 2 and 3 degree sets. The magnitude of overlap between these distribution is large; in some regions and variables this overlap is over 20% for temperature and close to 50% in precipitation, indicating that it is not possible to robustly ascribe regional responses based on GMST.

- Even if GMST is constrained to within 0.2 degrees Celsius, significant regional uncertainties would remain. It follows from this result that global mean constraints are of limited use and that methods based on the patterns of change might be preferable.

8.7 Additional

Data for this Chapter has been taken from the CPDN experiment, as described in Chapter 4, Section 4.4.3. Data was available on regional and grid–box scale as 8 year seasonal means for each experimental phase. Analysis is based on these 8 year mean temperature and precipitation fields and climate changes are calculated by taking the difference between the last 8 years of the doubled CO_2 and control phases. This approach assumes that the model simulations are roughly in equilibrium during

these years.

Figure 8.10 shows the time series of annual GMST for all three phases for the 2 and 3 degree sets. The distribution of these sets is shown in the left panels. The quality controlled ensembles are stable during the calibration phase, showing a little more variability in the control phase, then a marked increase in GMST at the start of phase 3 at the point of CO_2 doubling. Whilst all simulations may not have converged to equilibrium by year 38, it appears that most of the transient warming has already taken place by year 38 in the final phase. It is assumed in this Chapter that the last 8 year mean of phases 2 and 3 provide an indicative estimate of regional climate changes.

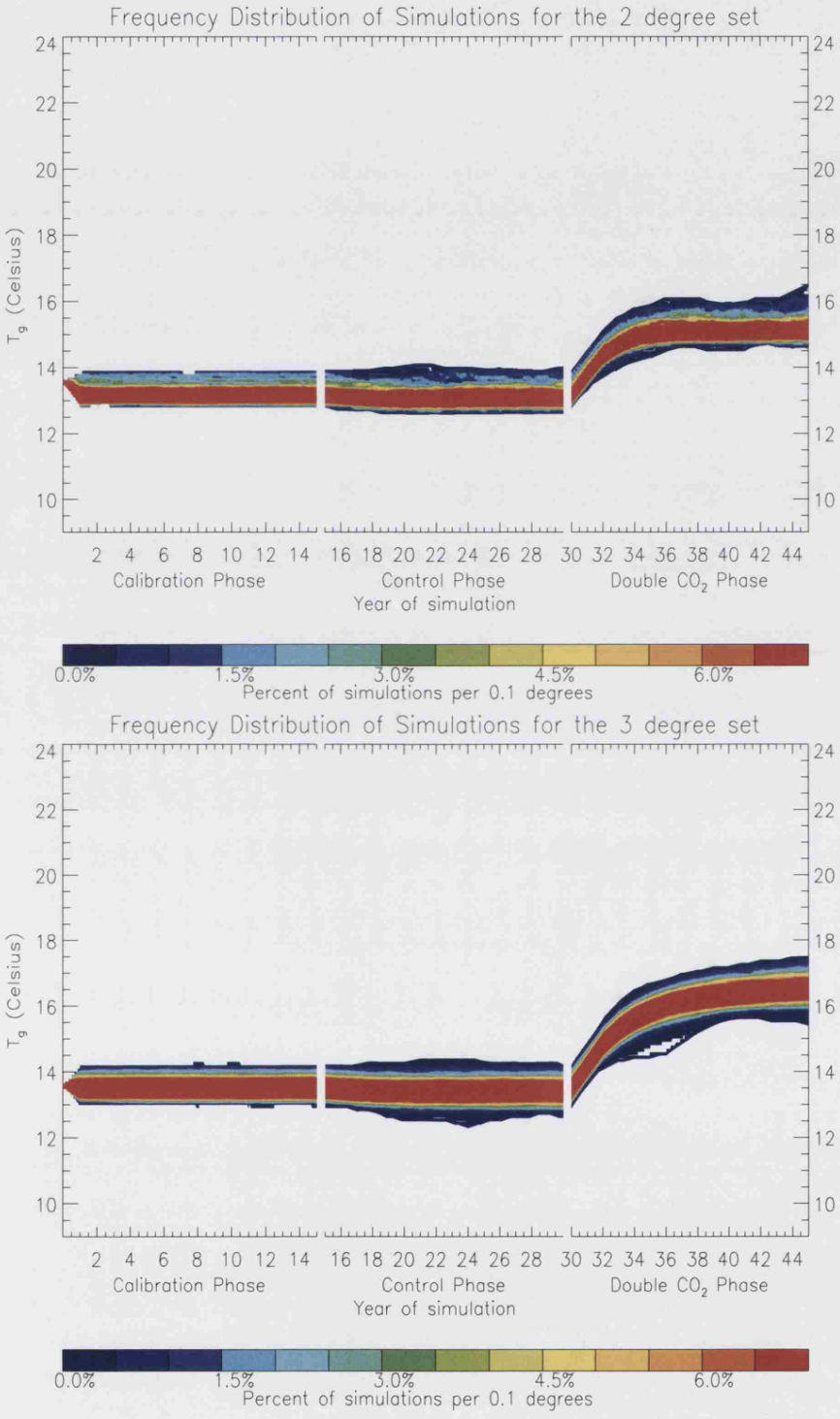


Figure 8.10: Time series for the 2 and 3 degree sets. Transient warming occurs at the point of CO_2 doubling (year 30) and stabilises by the end of the final phase. Data on regional climate changes was available for the final 8 years of each phase.

Chapter 9

Conclusion

9.1 Overview

This Chapter summaries the results presented in this Thesis and interprets them as implications for climate science and decision makers. Possible ways for climate science to move towards providing more decision–relevant information are proposed and discussed.

Section 9.2 summarises results concerning the different types of uncertainty present in climate predictions. Important results from Chapters 6, 7 and 8 are explained in light of three different categories of uncertainty. The implications of the results presented in this Thesis are explained in Section 9.3. Section 9.4 gives suggestions for further work that leads directly on from this Thesis.

9.2 Uncertainties

The four main categories of uncertainty are presented in Chapter 2; forcing uncertainty, initial condition uncertainty, model uncertainty and model inadequacy. Forcing uncertainty is not dealt with in this Thesis. The three remaining uncertainties are listed below, together with the insights that arise from this Thesis:

9.2.1 Initial Condition Uncertainty

Initial Condition Uncertainty (ICU) (Chapter 6) has previously been little studied due to pressures on computational resources and the widespread assumption that the magnitude of ICU is small on climatic time and length scales. This assumption has been shown to be significant on some length and time scales in Chapter 6.

Whilst the question of the division of computational resources must depend critically on the goals of each particular experiment, it can not be assumed a priori that ICU is small. It has been shown that the magnitude of ICU can be large on regional length scales in key variables. Differences on a grid–box level of up to 10 degrees Celsius are shown in 8 year mean DJF temperature change due to doubling of CO_2, arising from perturbation of ICs alone in Section 6.3. Furthermore, it is argued that ICEs are important tools in the analysis of the distribution of climate, providing the ability to evaluate the internal variability of models and differentiate more robustly the differences between models and scenarios (Section 6.4).

9.2.2 Model Uncertainty

Model uncertainty is a result of the existence of a diversity of climate models. There are different ways to estimate model uncertainty; it is possible to use a multi–model ensemble or to perturb parameters in a single structural model to understand inter–model differences. Efforts to estimate model uncertainty are hampered by the availability of different structural models and their lack of independence. It is not possible to fully explore model uncertainty since there is no definable class of models that can be exhaustively sampled.

The parameter perturbation approach is used in Chapter 7 using data from the CPDN experiment. Using a set of parametrically "de–tuned" Bender (2008); Stocker (2004) models might give a more realistic view of predictive uncertainties than taking a model with a single set of parameters, tuned on past observations Allen *et al.* (2006). The choice of data set was further motivated by the availability of large

amounts of climate model data – 45644 simulations have been presented in this Thesis. Access to a large ensemble of climate models simulations allows for a more thorough analysis of model uncertainties than has previously been carried out. For comparison, the IPCC Fourth Assessment Report Solomon *et al.* (2007a), the largest attempt to synthesise research on climate change, brought together a total of only 58 simulations from 14 different modelling centres for an analysis of 20th Century climate; similar sized ensembles were also used to simulate future climate changes. The analysis of model uncertainties in the CPDN experiment yielded uncertainties of an unprecedented magnitude. In particular, the range of values for estimated Climate Sensitivity (CS) ranged from 0.9 to 16.4 degrees Celsius, as presented in Chapter 7. Chapter 7 also examines the problem of constraining model uncertainty in light of statistical good practice. It is shown that this problem is non–trivial and the results of constraining can depend on subjective choices such as which variable to use when evaluating simulations' in–sample performance. The results of Chapter 7 suggest that important aspects of model behaviour, such as CS, are not robust to changes in parameter values. The parametrically de–tuned HadSM3 model can provide very different estimates of CS and regional climate changes, suggesting that the value of CS in the standard HadSM3 model is misleadingly robust due to the tuning of parameter values.

9.2.3 Model Inadequacy

Model inadequacy is a result of the fact we do not have access to a perfect climate model, even if such a model might exist. All climate models are necessarily inadequate in some way due **1)** a lack of accurate and complete representation of physical processes and **2)** low resolution of model features and a lack of scientific understanding. The inadequacy of models is difficult to evaluate given the lack of out–of–sample climate data; other methods must be used to estimate the potential effect of model inadequacy on the utility of predictions. The quality of in–sample

fit and the diversity of model predictions provide lower bounds on out–of–sample model performance. A model not evaluated on out–of–sample data and then used to extrapolate can not be established as adequate, although a model can be shown to be inadequate by failing in–sample tests. Thus, climate predictions can only ever be conditional and provisional in nature.

9.2.4 Constraining uncertainties and regional climate response

The large range of behaviour in the CPDN ensemble is examined in Chapter 7. Chapter 7 discusses three methods that can be used to constrain this range of behaviour demonstrates some of the difficulties in constraining mode diversity in a way consistent with statistical good practice. The range and distribution of estimated CS is dependent on choices such as which method to use and which variables are looked at (Section 7.5).

Leading on from the results of Chapter 7, Chapter 8 looks at sets of simulations within a narrow range of GMST change. By choosing simulations with a specific value of GMST it is possible to examine the regional variations that are present when GMST change (or CS) is tightly constrained at a global level (in this case, to within 0.2 degrees Celsius). It is shown in Chapter 8 that there are still significant regional uncertainties present for simulations with similar GMST changes. This result holds deep implications for policy–making and model–based decision support (Section 8.5). Whilst the sign of temperature change is everywhere positive, the magnitude of regional temperature change can vary by over 4 degrees Celsius for simulations with very similar GMST change. In almost every region, the sign of seasonal precipitation change is uncertain. These results imply a limit to our ability to answer the question "What does a 2 degree world look like?", and therefore the relevance of setting targets in terms of GMST. Furthermore, it is not clear how to apply pattern–scaling techniques in light of uncertainties in the direction of regional

response.

9.3 Implications

The results of this Thesis have implications for the future of climate modelling and for the manner in which climate models might be used to inform policy. In particular, it has been shown in Chapter 3 that the uncertainties involved in climate prediction are large and have not been fully explored or well communicated in the past.

The analysis of climate model data carried out in Chapters 3, 5, 6, 7 and 8 suggests that:

1. Climate scientists need to improve their communication of uncertainties. It is important for decision makers to be aware of the relative confidence they should place in different aspects of climate science. This can only take place with a transparent disclosure of uncertainties on the behalf of climate scientists.

2. The internal variability of climate models can be large, especially on small spatial scales. ICEs can be used to evaluate such uncertainties and provide an estimate of the robustness of model response. Large ICEs are increasingly necessary as climate modelling experiments move from equilibrium to transient simulations where a single, long, simulations can not be used to reliably estimate a model's internal variability.

3. Attempts to reduce uncertainties should be based on spatial patterns of change and not just global mean statistics, such as GMST or CS. It has been shown Allen (1999); Hegerl *et al.* (1997) that it is preferable for the purposes of the detection and attribution of climate change to look at the spatial patterns of change rather than global means. The analysis of spatial climate changes might be particularly useful in reducing the regional uncertainties that remain

when constraining climate change using global means. The large regional uncertainties presented in Chapter 8 suggest that spatial constraining of the pattern of climate change Forest *et al.* (2007) is an important avenue of future research.

4. Downscaling methods that interpret climate output with impacts models should take into account the diversity of model outputs. Downscaling and integrated modelling can only provide robust information once the range of climate simulations is taken into account. Since uncertainties will cascade as more layers of modelling are included, local impact assessment should at least include the range of behaviour simulated by climate models. Studies that treat changes in climate as known can not offer robust results since they do not reflect the inherent uncertainties in climate predictions. It is highly recommended that the sensitivity of impact studies to the assumptions about climate change is included.

5. Policy might better aim to be robust to uncertainties either in the type of decisions made or by remaining flexible. Decision making in the presence of uncertainty is a significant field of research in itself Jordaan (2005); Tversky & Kahneman (1974) and will not be dealt with in any detail here. Nonetheless, due to the inherently provisional nature of climate science it is necessary for decisions to remain flexible, taking into account the relative costs and benefits of action and in–action Stern (2006). In general, over–confident actions based on incomplete information are likely to be sub–optimal; it is important for climate science to provide all the information relevant to the problem, including a full treatment of uncertainty.

6. Monitoring climate is important, especially in ascertaining which data gives the best indication of future climate changes. The signal of climate change can be stronger in certain regions, variables and seasons. For example, the

HadSM3 model suggests that winter temperature change will be most significant over land in the high Northern latitudes. If this result is robust across structural models, this may suggest that the signal of climate change could be more robustly and quickly distinguished from noise that using a global mean.

Section 9.4 presents further work that leads on from this Thesis.

9.4 Further Work

This Section highlights some areas of further work that lead on from this Thesis. Three main directions of research are suggested here; **1)** Extension of results here to different structural models and to transient experiments, **2)** Design of climate modelling experiments to provide optimal use of resources and improve relevance to decision makers and **3)** In general, a more focused collaboration with decision-makers in order to investigate how climate models can be of most practical use. By working with the end–users of climate model output climate scientists can improve the relevance and communication of results and ensure their work has be most long–term impact on important decisions. Such a collaboration between the providers and users of climate science would be helped by the methods presented in this Thesis for evaluating uncertainties and an honest interpretation of the robustness of model output.

9.4.1 Transient Experiments

A key extension of this work is to check how robust key results are when looked at under different model structures and in the case of transient simulations. It is expected that some of the results presented in this Thesis are features of the HadSM3 model and that other models will behave differently. Since there are no other PPEs of similar size to the *climateprediction.net* experiment, many of the methods used in this Thesis are not currently possible for other structural models. For example, it

is not possible to look at the magnitude of ICU without large ICEs. Small ensemble size is problematic for a complete evaluation of model uncertainties.

An alternative CPDN experiment that uses transient forcings could be a more fruitful source of data. This new experiment, that uses the HadCM3 model with transient increases in GHGs to simulate climate from 1920–2080 using a perturbed physics grand ensemble. At the time of writing over 20,000 simulations had been completed in this transient experiment. The transient experiment allows further evaluation methods to be used since model time series can be compared to observations. In the CPDN equilibrium experiment, it was not possible to compare time series due to the experimental design. Comparing model output to observations will allow for improved understanding of the dynamics of the HadCM3 model. Of particular relevance is the concept of shadowing Daron & Stainforth (2008); Gilmour (1999); Judd (2008b); Sauer *et al.* (1997). Whilst different definitions exist, a model can be said to shadow noisy observations if it stays close to the noisy observations for a period amount of time. The ability of a model to shadow to within at least a useful degree of precision (this is know as ϕ–shadowing Smith (2001)) is an important guide to its ability to match the dynamics of the underlying system. Other methods can be used, based on likelihood and probabilistic measures of model skill. Measures of model performance based on the similarity between model behaviour and the observed dynamics of the Earth's climate system hold potential to inform questions on the value of model predictions. A deeper understanding of systematic model errors can also help focus model development.

Another area of research that is likely to become very useful to decision makers is an evaluation of the short term (yearly to decadal) performance of state–of–the–art climate models. There are proposals to produce more forecasts on time scales of up to 10 years using current state–of–the–art climate models Keenlyside *et al.* (2008); Troccoli & Palmer (2007). Since predictions use information in the current state of climate and its internal variability, there is reason to believe they might

prove more accurate than simulations made without this information. The results will likely depend on how well observational analyses can be translated into model space. Yearly to decadal predictions have the advantage that they can be checked against observations in the not too distant future, unlike forecasts for 2080 and beyond or equilibrium experiments. Hindcast performance and the extent of model diversity can give an idea of the likely robustness of these forecasts.

9.4.2 Experimental design

Climate modelling experiments have focused on understanding the climate system from a scientific viewpoint. Attention is now turning to using climate models to inform decisions on climate impacts and mitigation policy. The relevance of climate science to such decisions might be improved in a number of ways by focusing models on pragmatic decision support rather than scientific understanding[1]. The precise form model improvement should take will not find a unique solution since decision–makers' needs vary. Potential paths towards providing more decision–relevant model output are listed below:

1. **Models** : Developing models with emphasis on the processes of interest to decision makers. This could include turning points in the climate system, such as ice–cap melting or shutdown of the Thermohaline Circulation, adding hurricanes or increasing resolution.

   In some cases, climate models are not able to provide decision makers with any useful information. This is particularly true on small length scales. In such cases, statistical models could be useful e.g. providing an estimate of the local change in precipitation for the next 5 years based on a statistical fit to local data, although statistical models will likely fail when the system undergoes a dynamical change, especially due to turning points in the climate

[1]Of course, scientific understanding must precede any attempt to use models to inform decisions. What is advocated here is rather that the use of models to inform decisions requires new ways of designing experiments and interpreting results

system.

2. **Design** : Experimental design could be influenced by the requirements of decision makers e.g. if a climate modelling experiment aims to inform impact assessment, perhaps running a large set of simulations, covering a range of structural models and parameter values, for the next 10–20 years would be of most direct use for decision support. In this case, different emission scenarios will likely have little effect Cox & Stephenson (2007), leaving experimental resources to be diverted to running larger ensembles or using a model with higher grid–box resolution.

3. **Ensembles** : Running ICEs of climate models allows for a quantification of the models' internal variability and an estimate of the scale of potential model error. The typical ensemble size of 1–9 members commonly used at present is often insufficient to evaluate a model's internal variability, which can be critical from a decision–makers' point of view.

4. **Data** : Due to constraints on computational resources it is often not practical to store all the output from a climate modelling experiment. Usually only a small subset of output is stored. Climate modelling experiments could be made more relevant to decision makers by storing specific variables of interest e.g. extreme values, local impact statistics, regional time series etc. It should be considered which variables are of most interest to scientists, for model development and for decision makers. The use of metadata is an important component of storing data. With limited storage capacity and time to analyse data, it is important to find "decision–sufficient" statistics that contain all (or as much as possible) the relevant information in the data set at large. Such statistics would be sufficient in the sense of decision–makers not benefitting from any further statistics of the data set.

9.5 Conclusion

The problem of climate prediction and a framework for evaluating climate models has been presented in this Thesis. A systematic categorisation of uncertainties has been introduced and explored using the largest set of climate simulations to date. Original work has been highlighted at the end of each Chapter. The main new results are that **1)** The presentation of model output in high–profile reports such as the IPCC AR4 obscures important information regarding uncertainties in climate model simulations, **2)** The effect of perturbing both initial conditions can be large on length scales relevant for decision–makers and that the internal variability can account for more model diversity than has been previously acknowledged, **3)** the effect of parameter perturbation on global and regional climate response can be huge. The de–tuned HadSM3 model versions generated by the CPDN experiment show that climate models can be highly sensitive to the values of uncertain parameter values, **4)** the range of model behaviour is not easily reducible in a physically and statistically consistent way and **5)** regional responses can not be robustly inferred from global mean changes. Global mean temperature is a much less decision–relevant statistic than was usually taken to be the case.

Whilst no result has been found that casts doubt on the link between rising GHGs cause and simulated warming (in fact, in a broad sense, the CPDN experiment adds evidence to support this relationship), it has been shown that the potential predictive skill of climate models is limited. It is important that uncertainties are understood as fully as possible and are communicated effectively to decision-makers. Not to do so will likely lead to sub-optimal policy, wasted resources and a loss of confidence in climate science.

Appendix A

Glossary

Adaptation Action taken to adapt to the impacts of climate change e.g. building flood defences.

Albedo The proportion of solar radiation reflected from the Earth's surface. After Baede (2007).

Analysis In a meteorolgical sense, an estimate of the state of the climate system.

Anthropogenic Man–made or caused by man. In this thesis, "anthropogenic" relates to the effect that the human race has on the Earth's climate.

Attractor The equilibrium set of states to which a dynamical system evolves to after a long time. Attractors are only well-defined in the equilibrium for certain types of mathematical system; they do not exist in real world systems Orrell (2007).

Bootstrap A non–parametric method of re–sampling used to estimate properties of an estimator, such as confidence intervals.

Bounding Box The range of values spanned by an ensemble in each dimension. The bounding box is defined by the minimum and maximum points of an ensemble in each dimension Judd *et al.* (2007); Weisheimer *et al.* (2004).

Carbon Dioxide (CO_2) A Greenhouse gas occurring naturally and as a result of burning fossil fuels such as oil, gas and coal. Carbon dioxide is often used as a reference against which other greenhouse gases are measured. After Baede (2007).

Chaos A mathematical term relating to deterministic systems that exhibit properties of sensitive dependence on initial conditions, recurrence and aperiodicity. Whilst neither the climate system or GCMs are chaotic in a strict sense, they are often both treated as such since they display chaotic-like properties.

Climate The distribution of weather over long time scales (30 years is used by the World Meteorological Organisation).

Climatology The long–run distribution of historical observations. Can also refer to the study of climate in general.

Climate Change An alteration of the climate distribution. Climate change can occur due to a number of sources both natural and anthropogenic. In the context of this thesis, the potential effect of increasing GHGs on the climate system is studied.

Climate Model A mathematical representation of the Earth's climate.

climateprediction.net (CPDN) A distributed computing experiment, launched to the public on 12th September 2003. A grand ensemble of parametrically perturbed versions of the HadSM3 model are run on home PCs in order to evaluate uncertainties in climate predictions.

Climate Sensitivity (CS) A number of definitions exist in the literature. In this thesis, climate sensitivity is defined via the following thought experiment: a model is run using pre–industrial concentrations of CO_2 for a long–time, so that its mean global temperature can be calculated. CO_2 concentrations are then doubled and the model run for a long time. The model then reaches a

new equilibrium global mean temperature. The difference between the doubled CO_2 global mean temperature and the pre–industrial CO_2 global mean temperature is the model's climate sensitivity.

Democracy Plot A plot showing the number (or percentage) of ensemble members that simulate an increase (or decrease) in a particular variable e.g. precipitation change. Democracy plots are useful for showing to what extent an ensemble agrees on the sign of the change.

Dynamical System A mathematical description of a deterministic process that develops over time. Here, dynamical systems refers to a set of differential equations that display sensitive dependence on initial conditions.

Ensemble A set of simulations run over the same time period. There are a number of different types of ensemble that are used for quantifying different quantities. See Initial Condition Ensembles, Perturbed Physics Ensembles and Grand Ensembles for details of these types of ensemble.

Equilibrium Experiment A climate modelling experiment in which models are used to estimate the equilibrium (long–run) response to forcing. In some cases equilibrium is not reached by the end of the experiment and must be estimated.

Feedbacks Interactive mechanisms that cause a non–linear response. An initial process causes another which in turn affect the original process, or the system itself. Feedbacks can be either positive (exagerrating the base effect) or negative (dampening the effect).

Forcing An agent that forces a system into a new set of dynamics. The IPCC uses the term "radiative forcing" to understand the effect of forcing agents: "The radiative forcing of the surface-troposphere system due to the perturbation in or the introduction of an agent (say, a change in greenhouse gas concentra-

tions) is the change in net (down minus up) irradiance (solar plus long-wave; in Wm-2) at the tropopause AFTER allowing for stratospheric temperatures to readjust to radiative equilibrium, but with surface and tropospheric temperatures and state held fixed at the unperturbed values." Baede (2007).

Flux adjustment An artificial adjustment made of the model for some physical variable such as heat, water or salt. A flux adjustment is usually calculated such that the model produces more realistic dynamics and behaves in a similar way to the Earth's climate.

Heat Flux Adjustment (HFA) A flux of heat into the ocean from the atmosphere so that a slab ocean displays Sea Surface Temperature and heat transport similar to that of a dynamic ocean. This is discussed in more detail in [chapterhflux].

General Circulation Model (GCM) Despite some variability in the literature, state of the art climate models are referred to here as General Circulation Models. Also called Global Climate Models.

Global Mean (Climate Model) An average taken over the entire model state space. GCMs operate in a 3 dimensional grid system on discrete time steps. The global mean can be calculated as the volume-weighted mean over the entire Earth [this is discussed in more detail in Chapter 4. Since observations are far more heterogeneous and sporadic, they are often adjusted and filtered through a model in order to produce an estimate of the global mean. This is also known as the analysis.

Global Mean Surface Temperature (GMST) The global mean value of surface temperature calculated over a certain period, usually annual or decadal, using observations.

Grand Ensemble A set of model simulations containing two layers of perturba-

tion. At the higher level, either different structural models or the same model with different parameter values is used. At the lower level, each of these models is run using a set of different initial conditions. A grand ensemble is a collection of initial condition ensembles run using models with different dynamics.

Greenhouse Gas (GHG) An atmospheric gas that absorb and emit radiation at specific wavelengths. This property cases the *greenhouse effect.* Greenhouse gases include methane, water vapour and carbon dioxide. In some modelling studies, the total effect of GHGs is simplified to the equivalent concentration of CO_2. The climateprediction.net experiment investigates the effect of doubling CO_2 concentrations from 275 parts per million to 550 parts per million.

Grid box A set of 3 dimensional points that make up a climate model.

Initial Condition The starting state of a model simulation at time zero. In the case of a GCM, this will be of very high dimension.

Initial Condition Ensemble (ICE) An ensemble of model simulations in which each run's initial condition is changed slightly. In the case of the climateprediction.net experiment, the ocean temperature in one grid box is perturbed by a small amount.

In–sample Data that was used to build or estimate parameters in a model. This contrasts with *out–of–sample.*

Local In this thesis, referring to length scales roughly equal to a model grid box i.e. $\sim$40,000 km^2.

Mitigation Action taken to prevent or limit climate change e.g. reducing carbon emissions.

Mixing time The time taken for two *initial condition ensembles* to become statistically indistinguishable.

Model version A copy of a standard structural model with parameter perturbation. In the case of the *climateprediction.net* experiment, the HadSM3 model structure is used throughout sometimes with its standard (original) parameter values but mostly with various parameters perturbed. The HadSM3 model with a non–standard set of parameter values is called a HadSM3 model version.

Multi–model Ensemble An ensemble in which more than one model is used..

Nonlinear Not linear. In general, nonlinear systems do not have the property of additivity ($f(x+y) = f(x) + f(y)$) or homogeneity ($f(ax) = af(x)$).

Numerical Weather Prediction (NWP) Computational modelling of the Earth's weather system. In this thesis, NWP refers to weather prediction on time scales up to two weeks.

Observation Meteorological measurements of the Earth's surface. Raw data is post–processed, usually using models, to obtain a uniform grid of values across the Earth.

Out–of–sample Data that was not available at any stage in the building or development of a model.

Parameter Multiple definitions exist. In climate modelling, a parameter is any non–mathematical constant (such as π or e) that defines the model and as such is not state–dependent. Parameter values define the model and its behaviour, unlike initial conditions that do not affect the model's behaviour or dynamics. Parameters do not change with time.

Parameterisation A representation of physical processes in a climate model not explicitly simulated in the model. Parameterisations are often statistical, arising as a result of finite model resolution.

Perturbed Physics Ensemble (PPE) An ensemble in which simulations are run using the same structural model but using more than one set of parameter values.

Probability The classical definition treats probability as the relative frequency of occurrence of total number of events. Bayesians treat probabilities in a subjective fashion. In this thesis, neither approach is adopted explicit, rather it is assumed that a probability is any quantity that meets the Kolmogorov axioms of probability and is intended to be used as such.

Probabilistic An expression of uncertainty that resembles a probability but that does not meet all the necessary requirements. Probabilistic statements can be qualitative e.g. "It is almost certain that rising CO_2 concentrations will contribute to global warming" or quantitative e.g. expressing the likelihood of events as odds that do not sum to one (break the Second Kolmogorov Axiom of probability).

Regional In this thesis, refers to an area relating to one of the regions as defined in Giorgi & Mearns (2000).

Scenario An emissions scenario used as a basis for forcings in a climate projection. Also see SRES.

Sea Surface Temperature (SST) The temperature close to the surface of a large body of water.

Skillful In the context of forecasting, a forecast is said to be skillful if it can beat some simple straw–men that are tantamount to guessing.

Special Report on Emissions Scenarios Emissions scenarios first developed by Nakicenovic *et al.* (2000) to be used as a basis for climate projections. These scenarios are based on story lines that follow particular demographic, societal, economic and technical changes.

Straw–man A naïve test for quality of information. Straw–men for forecast skill can include simple statistical techniques such as persistence or climatology.

Trajectory The time–ordered path of states taken by a simulation of a dynamical system.

Transient Experiment A climate modelling experiment run using time–varying forcings, usually as a simulation of Earth's climate system.

Verification The target values for a model. Not necessarily verifying the model itself Oreskes *et al.* (1994) but targets that we would like the model to get close to, in some sense. In perfect model experiments verifications are the values from the a model simulation itself. In the case of climate models, that are imperfect, observations of the Earth's climate can be used as verifications.

References

ALLEN, M. (1999). Do it yourself climate prediction. *Nature*, **401**, 642. 92, 189, 279

ALLEN, M. & FRAME, D. (2007). Call off the quest. *Science*. 46

ALLEN, M. & STAINFORTH, D. (2002). Towards objective probabilistic climate forecasting. *Nature*. 23, 37, 45, 46, 92, 169

ALLEN, M., FRAME, D., KETTLEBOROUGH, J. & STAINFORTH, D. (2006). *Predictability in Weather and Climate*, chap. Model Error in weather and climate forecasting. Cambridge University Press. 208, 276

ANDRONOVA, N. & SCHLESINGER, M. (2001). Objective estimation of the probability density function for climate sensitivity. *Journal of Geophysical Research*. 50, 195

ANNAN, J. & HARGREAVES, J. (2006). Using multiple observationally-based constraints to estimate climate sensitivity. *Goephysical Research Letters*. 23, 24, 95, 195, 219, 231

ANNAN, J. & HARGREAVES, J. (2007). Efficient estimation and ensemble generation in climate modelling. *Philosophical Transactions of the Royal Society A*. 40, 231

ANNAN, J., HARGREAVES, J., OHGAITO, R., ABE-OUCHI, A. & EMORI, S. (2006). Efficiently constraining climate sensitivity with paleoclimate simulations. *SOLA*. 108, 195, 207

ARNOLD, V. & AVEZ, A. (1968). *Ergodic Problems of Classical Mechanics*. W.A. Benjamin, New York. 166

ARRHENIUS, S. (1896). On the influence of carbonic acid in the air upon the temperature of the ground. *Philosophical Magazine and Journal of Science*. 21, 95

Association of British Insurers (2005) (2005). *Financial risks of climate change*, Association of British Insurers, climate Risk Management in Colloboration with Metroeconomica. 21

BAEDE, A. (2007). Glossary of Terms used in the IPCC Fourth Assessment Report. Tech. rep., Intergovernmental Panel on Climate Change, annex to Working Group 1. 286, 287, 289

BENDER, F.A.M. (2008). A note on the effect of GCM tuning on climate sensivitiy. *Environmental Research Letters*. 52, 58, 276

BINTER, R., SMITH, L. & CLARKE, L. (2009). Ensemble analog–based conditional climatologies. Tech. rep., Centre for the Analysis of Time Series. 84

BOLTZMANN, L. (1884). Ableitung des stefanischen gesetzes, betreffend die abhängigkeit der wärmestrahlung von der temperatur aus der electromagnetischen licttheorie. *Annalen der Physik und Chemie*, **22**, 291–294. 55, 85

BRÖCKER, J. (2005). Scoring probabilistic forecasts, Available online at cats.lse.ac.uk. 42

BRÖCKER, J. & SMITH, L. (2007a). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, **22**. 36

BRÖCKER, J. & SMITH, L. (2007b). Scoring probabilistic forecasts: On the importance of being proper. *Weather and Forecasting*. 42

BUDYKO, M. (1958). *The heat balance of the earth's surface*. U.S. Dept. of Commerce, Weather Bureau. 84

CESS, R.D., POTTER, G., BLANCHET, J., BOER, G.J., GHAN, S.J., KIEHL, J.T., H., L.T., LI, Z.X., LIANG, X.Z.L., MITCHELL, J.F.B., MORCRETTE, J.J., RANDALL, D.A., RICHES, M.R., ROECKNER, E., SCHLESE, U., SLINGO, A., TAYLOR, K.E., WASHINGTON, W.M., WETHERALD, R.T. & YAGAI, I. (1989). Interpretation of cloud–climate feedback as produced by 14 atmospheric general circulation models. *Science*. 155

CHANGNON, S., RAVENSCROFT, R., PILKEY, O., MATTINGLY, S., WALAKER, D., FELLOWS, J., PENDLETON, J., BRUNNER, R., STEWART, T., CHAPMAN, C., GAUTEIR, D., HERRICK, C., HOOKE, W., JAMIESON, D. & METLAY, D. (2000). *Prediction: Science, Decision Making and the Future of Nature*. Island Press. 154

CHATFIELD, C. (2002). Confessions of a pragmatic statistician. *The Statistician*, **51, Part 1**, 1–20. 38

CHRISTENSEN, C., AINA, T. & STAINFORTH, D. (2005). The challenge of volunteer computing with lengthy climate modelling simulations. *Proceedings of the 1st IEEE Conference on e-Science and Grid Computing, Melbourne, Australia, 5–8 Dec, 2005*. 90, 92

COLLINS, M. & ALLEN, M. (2002). Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. *Journal of Climate*, **15**, 3104–3109. 179

COLLINS, M. & KNIGHT, S. (2007). Ensmbles and probabilities: a new era in the prediction of climate change. *Phil. Trans. R. Soc. A*, **365**, 1957–1970. 39

COLLINS, M., BOOTH, B., HARRIS, G., MURPHY, J., SEXTON, D. & WEBB, M. (2006). Towards quantifying uncertainty in transient climate change. *Climate Dynamics*, **27**, 127–147. 108, 121, 122, 130, 211

COLMAN, R. (2003). A comparison of climate feedbacks in general circulation models. *Climate Dynamics*, **20**, 865–873. 197

COMMISSION, E. (2007). Limiting global climate change to 2 degrees celsius - the way ahead for 2020 and beyond. 229

COVEY, C., ACHUTARAO, K., CUBASCH, U., JONES, P., LAMBERT, S., MANN, M., PHILLIPS, T. & TAYLOR, K. (2003). An overview of results from the coupled model intercomparison project. *Global and Planetary Change*. 26, 54, 199, 215

COX, P. & STEPHENSON, D. (2007). A Changing Climate for Prediction. *Science*, **317**, 207–208. 179, 284

COX, P., BETTS, R., JONES, C., SPALL, S. & TOTTERDALL, I. (2000). Acceleration of Global Warming Due to Carbon-Cycle Feedbacks in a Coupled Climate Model. *Nature*. 90

CUBASCH, U., MEEHL, G., BOER, G., STOUFFER, R., DIX, M., NODA, A., SENIOR, C. & RAPER, K., S.C.B.AND YAP (2001). *Projections of Future Climate Change*, chap. In: Climate Change 2001: The Scientific Basis: Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (J.T. Houghton, Y. Ding, D.J. Griggs, M. Noguer, P.J. Van der Linden, X. Dai, K. Maskell and C.A. Johnson (Eds.). Cambridge University Press (Cambridge, New York). 35

CUELLAR SANCHEZ, M. (2006). *Perspectives and Advances in Parameter Estimation of Nonlinear Models*. Ph.D. thesis, London School of Economics, University of London. 37

DAI, A. (2006). Precipitation characteristics in eighteen coupled climate models. *Journal of Climate*. 38

DARON, J. & STAINFORTH, D. (2008). Shadowing techniques for the evaluation and interpretation ofclimate models. *Geophysical Research Abstracts*. 282

DESSAI, S. & HULME, M. (2004). Does climate need probabilities? *Climate Policy*, **4**. 24

DIFFENHAUGH, N., GIORGI, F., RAYMOND, L. & BI, X. (2007). Indicators of 21st century socioclimatic exposure. *PNAS*, **104**, 20195–210198. 263

EFRON, B. & TIBSHIRANI, R. (1994). *An Introduction to the Bootstrap*. Cambridge University Press. 175, 248

FOREST, C., ALLEN, M., STONE, P. & SOKOLOV, A. (2000). Constraining uncertainties in climate models using climate change detection methods. *Geophysical Research Letters*. 190

FOREST, C., STONE, P., SOKOLOV, A., ALLEN, M. & WEBSTER, M. (2002). Quantifying uncertainties in climate system preoperties with the use of recent climate observations. *Science*. 195

FOREST, M., C.E. ALLEN, SOKOLOV, A. & STONE, P. (2007). Constraining climate model properties using optimal fingerprint detection methods. *Climate Dynamics*. 190, 207, 231, 280

FOURIER, J. (1824). Remarques generales sur les temperatures du globe terrestre et des espaces planetaires. *Annales de Chimie et de Physique*, **27**, 136–67. 86

FRAME, D., FAULL, N., JOSHI, M. & ALLEN, M. (2005). Constraining climate forecasts: The role of prior assumptions. *Philosophical Transactions of the Royal Society A*. 23, 37, 169, 195, 208, 219

FRAME, D., BOOTH, B., KETTLEBOROUGH, J., STAINFORTH, D., GREGORY, J., COLLINS, M. & ALLEN, M. (2007). Probabilistic climate forecasts and inductive problems. *Geophysical Research Letters*. 23, 37, 42

GALAMBOS, J. (1995). *Advanced Probability Theory*. Probability: Pure and Applied, CRC Press, second edition edn. 61

GILMOUR, I. (1999). *Nonlinear model evaluation: ι shadowing, probabilistic prediction and weather forecasting*. Ph.D. thesis, Universtiy of Oxford. 282

GIORGI, F. (2008). A Simple Equation for Regional Climate Change and Associated Uncertainty. *Journal of Climate*. 263, 264

GIORGI, F. & FRANCISCO, R. (2000). Evaluating uncertainties in the prediction of regional climate change. *Geophys. Res. Lett.*, **27**, 1295–1298. 34, 172, 264

GIORGI, F. & MEARNS (2000). Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the reliability ensemble averaging (REA) method . *J. Clim*, **15**, 1141–1158. 106, 232, 292

GIORGI, F. & MEARNS, L. (2003). Probability of regional climate change based on ther reliability ensemble averaging (REA technique). *Geophysical Resrach Letters*. 23

GOLDSTEIN, M. & ROUGIER, J. (2006). Reified bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*. 24, 47

GORDON, C., COOPER, C., SENIOR, C., BANKS, H. & GREGORY, J. (2000). The simulation of SST, sea ice extents and ocean heat transports in a version of the hadley centre coupled model without flux adjustments. *Climate Dynamics*, **16:147–168**. 84, 90

GREGORY, J., STOUFFER, R., RAPER, S., STOTT, P. & RAYNER, N. (2002).

An observationally based estimate of climate sensitivity. *Journal of Climate*. 98, 195

HANSEN, J., FUNG, I., LACIS, A., RIND, D., LEBEDEFF, S., RUEDY, R. & RUSSELL, G. (1988). Global climate changes as forecast by goddard insitute for space studies three–dimensional model. *Journal of Geophsyical Research*, 9341–9364. 41

HANSEN, J., SATO, M., RUEDY, R., LO, K., LEA, W. & MEDINA-ELIZADE, M. (2006). Global temperature change. *Proc. Nat. Acad. Sci.*, 14288–14293. 41

HEGERL, G., HASSELMANN, K., CUBASCH, U., MITCHELL, J., ROECKNER, E., VOSS, R. & WASKEWITZ, J. (1997). Multi–fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas–plus aerosol and solar forced climate change. *Climate Dynamics*, **13**, 613–634. 279

HEGERL, G.E.A. (2006). Climate sensitivity constrained by temperature reconstructions over the past seven centuries. *Nature*. 108, 190, 207, 231

HEINLEIN (1973). *Time Enough for Love*. G.P. Putnam's Sons, 605pp. 165

HEWITT, C. & MITCHELL, J. (1997). Radiative forcing and response of a gcm to ice age boundary conditions: cloud feedback and climate sensitivity. *Climate Dynamics*, **13**, 821–834. 91

HODGES, J. (1991). Six (or so) things you can do with a bad model. *Operations Research*. 46

HOUGHTON, J., MEIRA FILHO, L., CALLENDER, B., HARRIS, N., KATTENBERG, A. & MASKELL, K., eds. (1995). *Climate Change 1995: The Science of Climate Change: Contribution of Working Group I to the Second Assessment of the Intergovernmental Panel on Climate Change*. Cambridge University Press. 97

HOUGHTON, J., DING, Y., GRIGGS, D., NOGUER, M., VAN DER LINDEN, P. & XIAOSU, D., eds. (2001). *Climate Change 2001: The Scientific Basis*. Cambridge University Press, 994pp. 97, 122, 123, 153, 154, 168

HUME, D. (1748). *An enquiry concerning human understanding*. Oxford Philosophical texts. 39, 42

JENKINS, G., PERRY, M. & PRIOR, M. (2007). The climate of the united kingdom and recent trends. Tech. rep., Met Office Hadley Centre. 47

JOHNS, T., CARNELL, R., CROSSLEY, J., GREGORY, J., MITCHELL, J., SENIOR, C., TETT, S. & WOOD, R. (1997). The second Hadley Centre coupled atmosphere-ocean GCM: model description, spinup and validation. *Climate Dynamics*, **13**, 103–134. 122

JOHNS, T.C., DURMAN, C.F., BANKS, H.T., ROBERTS, M.J., MCLAREN, A.J., RIDLEY, J.K., SENIOR, C.A., WILLIAMS, K.D., JONES, A., RICKARD, G.J., CUSACK, S., INGRAM, I.M., CRUCIFIX, M., SEXTON, D.M.H., JOSHI, M.M., DONG, B.W., SPENCER, H., HILL, R.S.R., GREGORY, J.M., KEEN, A.B., PARDAENS, A.K., LOWE, J.A., BODAS-SALCEDO, A., STARK, S., & SEARL, Y. (2006). The new Hadley Centre climate model HadGEM1: Evaluation of coupled simulations. *Journal of Climate*. 84

JOHNSON, D. (1997). General coldness of climate models and the second law: Implications for modeling the earth system. *Journal of Climate*. 52

JONES, P., NEW, M., PARKER, D., MARTIN, S. & RIGOR, I. (1999). Surface air temperature and its changes over the past 150 years. *Reviews Geophsyics*. 9, 71

JORDAAN, I. (2005). *Decisions under Uncertainty*. Cambridge University Press. 280

JUDD, K. (2008a). Non–probabilistic odds and forecasting with imperfect models, submitted. 24, 47

JUDD, K. (2008b). Shadowing pseudo–orbits and gradient descent noise reduction. *Journal of Nonlinear Science*, **18**, 57–74. 282

JUDD, K. & SMITH, L. (2001a). Indistinguishable states 1: The perfect model scenario,. *Physica D*, **151**. 32, 35, 42

JUDD, K. & SMITH, L. (2001b). Indistinguishable states 2: The imperfect model scenario. *Physica D*, **196**. 42

JUDD, K., SMITH, L. & WEISHEIMER, A. (2007). How good is an ensemble at capturing truth? *Quarterly Journal of the Royal Meteorologial Society*, **133**, 1309–1325. 60, 130, 286

KARL, T. & TRENBERTH, K. (2003). Modern global climate change. *Science*, 1719–1723. 38

KEENLYSIDE, N., LATIF, M., JUNGCLAUS, J., KORNBLUEH, L. & ROECKNER, E. (2008). Advancing decadal–scale climate prediction in the north atlantic sector. *Nature*. 282

KENNEDY, M. & O'HAGAN, A. (2001a). Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B*. 38

KENNEDY, M. & O'HAGAN, A. (2001b). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society B*. 23, 37

KIEHL, J., SHIELDS, C., HACK, J. & COLLINS, W. (2006). The climate sensitivity of the community climate system model version 3 (CCSM3). *Journal of Climate*, **19**, 2584–2596. 62

KNIGHT, C., KNIGHT, S., MASSEY, N., AINA, T., CHRISTENSEN, C., FRAME, D., KETTLEBOROUGH, J., MARTIN, A., PASCOE, S., SANDERSON, B., STAINFORTH, D. & ALLEN, M. (2007). Association of parameter, software and hardware variation with large scale behaviour across 57,000 climate models. *Proceed-*

ings of the National Academy of Sciences. 41, 98, 99, 102, 108, 111, 158, 167, 191, 194, 208

KNUTSON, T. (2008). Fms slab ocean model technical documentation. Tech. rep., Geophysical Fliud Dynamics Laboratory, National Oceanic and Atmospheric Administration. 121

KNUTTI, R. (2008a). Should we believe model predictions of future climate change? *Submitted to the 2008 Visions of the Future Triennial Issue of Philosophical Transactions A.* 190

KNUTTI, R. (2008b). Why are climate models reproducing the observed global surface warming so well? *Submitted to Geophysical Research Letters.* 53

KNUTTI, R., MEEHL, G., ALLEN, M. & STAINFORTH, D. (2006). Constraining climate sensitivity from the seasonal cycle in surface temperature. *Journal of Climate.* 108, 194, 207, 209, 212, 219, 231

KORPELA, E., WERTHIMER, D., ANDERSON, D., COBB, J. & LEBOFSKY, M. (2001). Set@home – massively distributed computing for SETI. *Computing in Science & Engineering*, **3**, 78–83. 92

KRUSKAL, W. & WALLIS, W. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47**, 583–621. 61, 65

LEWIS, S. (2003). Modelling the Martian Atmosphere. *Astronomy and Geophsyics.* 84

LIU, H.L., SASSI, F. & GARCIA, R. (2008). Error growth in a whole atmosphere climate model. *Journal of the Atmospheric Sciences*, to appear. 169

LORENZ, E. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Sciences.* 32, 36

LYNAS, M. (2007). *Six Degrees: Our future on a hotter planet.* Fourth Estate. 248

MANABE, R., S. ADN WETHERALD (1975). The effects of doubling co_2 concentrations on the climate of a general circulation model. *Journal of the Atmospheric Sciences*. 84, 95

MANABE, S. & BRYAN, K. (1969). Climate calculations with a combined ocean-atmosphere model. *Journal of the Atmospheric Sciences*. 84

MANN, H. & WHITNEY, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*. 65, 141

MAY, R. (1976). Simple mathematical models with very complicated dynamics. *Nature*. 39

MCGUFFIE, K. & HENDERSON-SELLERS, A. (2006). *A Climate Modelling Primer*. John Wiley and Sons, Ltd. 22, 88

MEEHL, G., COVEY, C., MCAVANEY, B., LATIF, M. & STOUFFER, R. (2005). Overview of the coupled model intercomparison project. *Bulletin of the American Meteorological Society*, 89–93. 54

MIN, S.K., LEGUTKE, S., HENSE, A. & KWON, W.T. (2005). Internal variability in a 1000-yr control simulations with the coupled climate model ECHO-G-I. near-surface temperature, preciptation and mean sea level pressure. *Tellus*. 168, 192

MORAN, P. (1950). Notes on continuous stochastic phenomena. *Biometrika*. 137

MORGAN, M. & KEITH, D. (1995). Subjective judgements by climate experts. *Environmental policy analysis*, **29**. 195

MURPHY, J. (1995). Transient response of the Hadley Centre coupled ocean-atmosphere model to increasing carbon dioxide. part I: control climate and flux adjustment. *Journal of Climate*. 121, 122

MURPHY, J., SEXTON, D., BARNETT, D., JONES, G., WEBB, M., COLLINS, M. & STAINFORTH, D. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*. 39, 40, 58, 89, 108, 121, 191, 269

NAKICENOVIC, N., ALCAMO, J., DAVIS, G., DE VRIES, B., FENHANN, J., GAFFIN, S., GREGORY, K., GRBLER, A., JUNG, T., KRAM, T., LA ROVERE, E., MICHAELIS, L., MORI, S., MORITA, T., PEPPER, W., PITCHER, H., PRICE, L., RIAHI, K., ROEHRL, A., ROGNER, H., SANKOVSKI, A., SCHLESINGER, M., SHUKLA, P., SMITH, S., SWART, R., VAN ROOIJEN, S., VICTOR, N. & DAD, Z. (2000). Special Report on Emissions Scenarios: A Special Report of Working Group III of the Intergovernmental Panel on Climate Change. Tech. rep., Intergovernmental Panel on Climate Change. 35, 48, 180, 292

NEW, M., HULME, M. & JONES, P. (1999). Representing twentieth–century space–time climate variability. Part I: Development of a 1961–90 mean monthly terrestrial climatology. *Journal of Climate*, **12**, 829–856. 86, 93, 122, 124

ORESKES, N. (1998). Evaluation (not validation) of quantitative models. *Environmental Health Perspectives*. 23

ORESKES, N. (2004). The Scientific Consensus on Climate Change. *Science*. 21

ORESKES, N., SHRADER-FRECHETTE, K. & BELITZ, K. (1994). Verification, validation and confirmation of numerical models in the earth sciences. *Science*. 23, 38, 42, 52, 293

ORRELL, D. (2007). *The Future of Everything : The Science of Prediction*. Thunder's Mouth Press. 286

ORRELL, D., SMITH, L., BARKMEIJER, J. & PALMER, T. (2001). Model error in weather forecasting. *Nonlinear Processes in Geophsyics*. 32

PALMER, T., SHUTTS, G., HAGEDORN, R., DOBLAS-REYES, F., JUNG, T. & LEUTBECHER, M. (2005). Representing uncertainty in forecasts of weather and climate. *Annual Review of Earth and Planetary Sciences*, **33**, 163–193. 32, 89

PARRY, M., CANZIANI, O., PALUTIKOF, J., VAN DER LINDEN, P. & HANSEN, C., eds. (2007). *IPCC, 2007: Climate Change 2007: Impacts, Adaptation and Vunerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. 21, 229, 255, 268, 269

PHILLIPS, T., POTTER, G., WILLIAMSON, D., CEDERWALL, R., BOYLE, J., FIORINO, M., HNILO, J., OLSON, J., XIE, S. & YAO, J. (2004). Evaluating parameterizations in general circulations models: Climate simulation meets weather prediction. *Bulletin of the American Meteological Society*. 88

PIANI, C., FRAME, D., STAINFORTH, D. & ALLEN, M. (2005). Constraints on climate change from a multi-thousand member ensemble of simulation. *Geophysical Review Letters*, **32**. 93, 124, 190, 194, 207, 209, 212, 217

PITTOCK, A., JONES, R. & MITCHELL, C. (2001). Probabilities will help us plan for climate change. *Nature*. 23

POPE, V., GALLANI, M., ROWNTREE, P. & STRATTON, R. (2000). The impact of new parametrisations in the Hadley Centre climate model—HadAM3. *Climate Dynamics*, **16:123–146**. 84, 88, 90

PRESS, W.E.A. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press. 97, 140

RAISANEN, J. (2007). How reliable are climate models? *Tellus A*. 24

RAPER, S., GREGORY, J. & STOUFFER, R. (2001). The role of climate sensitivity and ocean heat uptake on AOGCM transient temperature response. *Journal of Climate*, **15**, 124–130. 62

REICHLER, T. & KIM, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*. 41, 56, 192

RINGER, M.A., MCAVENEY, B.J., ANDRONOVA, N., BUJA, L.E., ESCH, M., INGRAM, W.J., LI, B., QUAAS, J., ROECKNER, E., SENIOR, C.A., SODEN, B.J., VOLODIN, E.M., WEBB, M.J. & WILLIAMS, K.D. (2006). Global mean cloud feedbacks in idealized climate change experiments. *Geophsyical Research Letterse*. 155

ROBOCK, A. & OPPENHEIMER, C., eds. (2003). *Volcanism and the Earths Atmosphere, Geophysical Monograph 139*. American Geophysical Union, Washington, DC. 35

ROE, G.H. & BAKER, M. (2007). Why is climate sensitivity so unpredictable? *Science*, **318**, 629–632. 29, 95, 195, 197, 208, 209, 210, 221

RUOSTEENOJA, K., TUOMENVIRTA, H. & JYLHA, K. (2007). GCM–based regional temperature and precipitation change estimates for europe under four SRES scenarios applying a super–ensemble patter–scaling method. *Climatic Change*. 263

RUSSELL, B. (1946). *The problems of philosophy*. Oxford University Press. 39

SANDERSON, B., PIANI, C., INGRAM, W., STONE, D. & ALLEN, M. (2007). Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics GCM simulations. *Climate Dynamics*. 108, 122, 189, 191, 194, 208

SANDERSON, B., KNUTTI, R., AINA, T., CHRISTENSEN, C., FAULL, N., FRAME, D., INGRAM, W., PIANU, C., STAINFORTH, D., STONE, D. & ALLEN, M. (2008). Constraints on model reponse to greenhouse gas frocing and the role of subgrid-scale processes. *Journal of Climate*. 40, 89, 101, 190, 208, 213, 217

SANTER, B., BRUGGERMANN, W., CUBASCH, U., HASSELMANN, K., HOCK, H., MAIER-REIMER, E. & MIKOLAJEWICZ, U. (1994). Signal–to–noise analysis of time–dependent greenhouse warming experiments. *Climate Dynamics*. 122

SAUER, T., GREBOGI, C. & YORKE, J. (1997). How long do numerical chaotic solutions remain valid? *Physical Review Letters*. 282

SAUSEN, R., BARHEL, K. & HASSELMANN, K. (1988). Coupled ocean-atmosphere models with flux correction. *Climate Dynamics*. 121

SCHELLNHUBER, H., CRAMER, W., NAKICENOVIC, N., WIGLET, T. & YOHE, G., eds. (2005). *Avoiding Dangerous Climate Change*. Cambridge Universtiy Press, 406pp. 21

SCHMIDT, G.A., RUEDY, R., HANSEN, J., ALEINOV, I., BELL, N., BAUER, M., BAUER, S., CAIRNS, B., CANUTO, V., CHENG, Y., DELGENIO, A., FALUVEGI, G., FRIEND, A., HALL, T., HU, Y., KELLEY, M., KIANG, N., KOCH, D., LACIS, A., LERNER, J., LO, K.K., MILLER, R., NAZARENKO, L., OINAS, V., PERLWITZ, J., RIND, D., ROMANOU, A., RUSSELL, G., SATO, M., SHINDELL, D., STONE, P., SUN, S., TAUSNEV, N., THRESHER, D. & YAO, M.S. (2006). Present day atmospheric simulations using GISS Model E: Comparison to in-situ, satellite and reanalysis data. *Journal of Climate*, **19**, 153–192. 62

SCHWARTZ, S.E., CHARLSON, R.J. & RODHE, H. (2007). Quantifying climate change – too rosy a picture? *Nature Reports – Climate Change*. 195

SELLERS, W. (1969). A global climatic model based on the energy balance of the earth-atmosphere system. *Journal of Applied Meteorology*. 84

SHACKLEY, S., RISBEY, J., STONE, P. & WYNNE, B. (1999). Adjusting to policy expectations in climate change modeling. *Climatic Change*, **43**, 413–454. 121, 122

SMITH, L. (2001). *Nonlinear Dynamics and Statistics*, chap. Disentangling uncertainty and error: on the predictability of nonlinear systems. Birkhauser, mees, A. 42, 60, 282

SMITH, L. (2002). What might we learn from climate forecasts? *Proceedings of the National Academy of Sciences of the United States of America*, **99**. 23, 32, 34, 38, 190, 229

SMITH, L. (2007). *Chaos: A Very Short Introduction*. Oxford University Press. 32

SMITH, L., TREDGER, E. & STAINFORTH, D. (2008). On the relevance of model averages for science based policy. *Submitted to Geophysical Research Letters*. 52, 59, 175, 200

SOLOMON, S., QIN, D., MANNING, M., CHEN, Z., MARQUIS, M. & AVERY, K. (2007a). *IPCC, 2007: Climate Change 2007: The Physical Science Basis. Contribution of Working Group 1 to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 21, 22, 24, 33, 37, 39, 48, 49, 52, 53, 56, 57, 63, 86, 87, 95, 97, 99, 122, 123, 154, 168, 195, 196, 197, 201, 212, 228, 229, 248, 263, 269, 277

SOLOMON, S., QIN, D., MANNING, M., CHEN, Z., MARQUIS, M., AVERYT, K., TIGNOR, M. & MILLER, H.E. (2007b). IPCC, 2007: Summary for Policymakers. 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. Tech. rep., IPCC. 24

SPROTT, J. (2003). *Chaos and Time-Series Analysis*. Oxford University Press. 37, 39

STAINFORTH, D., KETTLEBOROUGH, J., MARTIN, A., SIMPSON, A., GILLIS, R., AKKAS, A., GAULT, R., COLLINS, M., GAVAGHAN, D. & ALLEN, M.

(2002). Climateprediction.net: design principles for public resources modelling research. *Proc. 14th IASTED conference on parallel and distributed computing systems*. 101, 189

STAINFORTH, D., AINA, T., CHRISTENSEN, C., FAULL, N., FRAME, D., KETTLEBOROUGH, J., KNIGHT, S., MARTIN, A., MURPHY, J., PIANI, C., SEXTON, D., SMITH, L., SPICER, R., THORPE, A. & ALLEN, M. (2005). Uncertainty in prediction of the climate response to rising levels of greenhouse gases. *Nature*, **433**. 26, 37, 39, 50, 69, 83, 91, 94, 95, 97, 99, 102, 103, 104, 105, 108, 112, 121, 126, 153, 157, 158, 189, 192, 194, 195, 212, 213, 215, 250, 269

STAINFORTH, D., ALLEN, M., TREDGER, E. & SMITH, L. (2007a). Confidence, uncertainty and decision–support relevance in climate predictions. *Philosophical Transactions of the Royal Society A :Mathematical, physical and engineering sciences*. 23, 32, 34, 36, 37, 39, 43, 166, 168, 194

STAINFORTH, D., DOWNING, T., WASHINGTON, R., LOPEZ, A. & NEW, M. (2007b). Issues in the interpretation of climate model ensembles to inform decisions. *Philosophical Transactions of the Royal Society A :Mathematical, physical and engineering sciences*. 21, 47, 52, 190, 194

STEFAN, J. (1879). Über die beziehung zwischen der wärmestrahlung und der temperatur. *Sitzungsberichte der mathematisch-naturwissenschaftlichen Classe der kaiserlichen Akademie der Wissenschaften*, **79**, 391–428. 55, 85

STERN, N. (2006). *The Economics of Climate Change: The Stern Review*. Cambridge University Press, Cambridge, UK and New York, NY, USA. 21, 228, 229, 248, 268, 269, 280

STOCKER, T. (2004). Climate change: Models change their tune. *Nature*. 52, 58, 276

STOTT, P. & KETTLEBOROUGH, J. (2002). Origins and estimates of uncertainty in predictions of twenty–first century temperature rise. *Nature*, **416**. 35

STOTT, P., STONE, D. & ALLEN, M. (2004). Human contribution to the European heatwave of 2003. *Nature*, **432**, 610–613. 180

STOTT, P., JONES, G., LOWE, J., THORNE, P., DURMAN, C., JOHNS, T. & THELEN, J.C. (2006). Transient climate simulations with the HadGEM1 climate model: Causes of past warming and future climate change. *Journal of Climate*, **19**, 2763–2782. 69

TALEB, N. (2008). *The Black Swan: The Impact of the Highly Improbable*. Random House. 39

TEBALDI, C. & KNUTTI, R. (2007). The use of multi–model ensemble in probabilistic climate projection. *Phil. Trans. Roy. Soc.*, **365**. 28, 39, 166, 168, 181

THORPE, A. (2005). Climate change : A challenging scientific problem. *Institute of Physics*. 22, 37

TOTH, Z., BRUCKNER, T., H-M, F., LEIMBACH, M. & PETSCHEL-HELD, G. (2003). Integrated assessment of long–term climate policies part 1 – model presentation. *Climatic Change*, 37–56. 229

TROCCOLI, A. & PALMER, T. (2007). Ensemble decadal prediction from analysed initial conditions. *Philosophical Transactions of the Royal Society A*, **365**, 2179–2191. 179, 282

TVERSKY, A. & KAHNEMAN, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*. 280

TYNDALL, J. (1861). On the absorption and radiation of heat by gases and vapours, and on the physical connexion of radiation, absorption, and conduction. *Philosophical Magazine*. 21