

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*

UMI Number: U615954

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U615954

Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



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

F
9091



1203531

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).

The copyright of this Thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This Thesis may not be reproduced without the prior written consent of the author.

I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

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 than 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

Nomenclature	20
1 Introduction	21
1.1 Methodology	23
1.2 Key results and new approaches	24
1.3 Chapter 2	25
1.4 Chapter 3: How reliable are climate models?	25
1.5 Chapter 4: Intro to CPDN	26
1.6 Chapter 5: Heat Flux	27
1.7 Chapter 6: Initial Condition Ensembles in Climate Modelling	28
1.8 Chapter 7: Constraining New Results from the CPDN grand ensemble	28
1.9 Chapter 8: On the relevance of Model Means for Decision-Support	29
2 Uncertainty and the use of State-of-the-Art climate models in decision-support	31
2.1 Overview	31
2.2 The problem of climate prediction	32
2.3 Uncertainties and Ensembles	34
2.3.1 Forcing Uncertainty	34
2.3.2 Initial Condition Uncertainty	35
2.3.3 Model Uncertainty	36
2.3.4 Model Inadequacy	38
2.3.5 Ensembles	39
2.3.6 Initial Condition Ensembles	40
2.4 Model Evaluation and straw-men for climate	41
2.4.1 In-sample fit	42
2.4.2 Initial Condition Test	43

2.4.3	Model diversity Test	44
2.4.4	Utility of models that fail these tests	45
2.5	Decision support	47
2.6	Uncertainties in Adaptation and mitigation decisions	48
2.7	Communication of uncertainties	49
2.8	Conclusion	50
3	How reliable are the models used to make projections of future climate change?	51
3.1	Introduction	51
3.2	General Circulation Models	53
3.3	The IPCC Figures	54
3.4	Presentation of Model Output	55
3.5	Residual Analysis of Model Output	59
3.6	Exchangeability	61
3.6.1	Estimating temporal correlation using order statistics	64
3.6.2	Testing the exchangeability of GCM output	65
3.6.3	Discussion of Results	66
3.7	Recommended Presentation of Model Output	67
3.8	Conclusion	68
4	Introduction to the <i>climateprediction.net</i> experiment	83
4.1	Introduction	83
4.2	Climate Models	84
4.2.1	On statistical methods of climate prediction	84
4.2.2	Energy Balance	85
4.2.3	Feedbacks	87
4.2.4	GCMs and Grid Boxes	87
4.2.5	Parameterisation	88
4.2.6	Parameter values	88
4.2.7	Time steps	89
4.2.8	HadSM3	90
4.3	The CPDN experiment	92
4.3.1	CPDN Experimental Design	92
4.4	Data	95
4.4.1	Climate Sensitivity	95

4.4.2	Quality Control	101
4.4.3	Data Format	106
4.5	Conclusion	111
5	Investigating variations in heat flux adjustment in the CPDN ensemble	121
5.1	Introduction	121
5.1.1	HFA in the CPDN experiment	124
5.1.2	HFA Data Sets	125
5.1.3	Examples of HFA fields	127
5.2	HFA variability	129
5.2.1	The HFA bounding box	130
5.2.2	Variability with Initial Condition	131
5.2.3	Perturbed Physics Ensembles	138
5.2.4	Stabilisation of Global Mean HFA	146
5.3	Seasonality in the HFA	150
5.4	HFA and Climate Sensitivity	153
5.5	HFA and drift	157
5.6	Conclusion	163
6	ICEs and the Internal Variability of Climate Models	165
6.1	Overview	165
6.2	Introduction to ICEs	168
6.3	The internal variability of HadSM3	169
6.4	ICEs and Robust Model Response	172
6.4.1	Regional response in the 64 member HadSM3 ensemble	173
6.4.2	Comparison of two Perturbed Physics ICEs	177
6.5	ICEs in Transient Experiments	179
6.6	Conclusion	180
7	Constraining New Results from the CPDN grand ensemble	189
7.1	Introduction	189
7.2	The Data Set	194
7.3	Climate Sensitivity	195
7.4	Sub-global Behaviour	200
7.5	Constraining Model Simulations	207

7.5.1	Constraining using the Entrainment Coefficient	207
7.5.2	Constraining using HFA	211
7.5.3	Constraining using in-sample fit to observations	212
7.6	Conclusion	219
8	The relevance of global means for climate policy	228
8.1	Introduction	228
8.1.1	Data Sets Used	231
8.2	What does a 2 degree rise in GMST mean?	232
8.2.1	From global to super-continental length scales	233
8.2.2	Regional Impacts	238
8.2.3	Grid-scale Impacts	245
8.3	What is the difference between 2 and 3 degrees GMST?	248
8.3.1	Regional differences	248
8.3.2	Grid-scale differences	258
8.4	Linearity of Regional Response	263
8.5	Discussion	267
8.5.1	Mitigation	269
8.5.2	Adaptation and Impact Assessment	270
8.6	Conclusion	271
8.7	Additional	272
9	Conclusion	275
9.1	Overview	275
9.2	Uncertainties	275
9.2.1	Initial Condition Uncertainty	276
9.2.2	Model Uncertainty	276
9.2.3	Model Inadequacy	277
9.2.4	Constraining uncertainties and regional climate response	278
9.3	Implications	279
9.4	Further Work	281
9.4.1	Transient Experiments	281
9.4.2	Experimental design	283
9.5	Conclusion	285
A	Glossary	286

List of Figures

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 <i>et al.</i> (1999))). The multi-model mean is plotted in red. There is a difference of up to 3 degrees between simulations’ GMST. . . .	71
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. . . .	72
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.	73
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).	74
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).	75

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).	76
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).	77
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.	78
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.	79
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.	80
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.	81
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.	82
4.1	The global mean heat capacity in the doubled CO_2 phase is plotted over 1460 quality controlled simulations.	113
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.	114

4.3	The range in estimated CS for three different methods is plotted against the mean for 1460 quality controlled simulations.	115
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	116
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	117
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. .	118
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).	119
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.	119
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. . .	120

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%. .	128
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 into the ocean. The range of values shown in panel (c) can be as large as $200W/m^2$	132
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.	135
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 $40W/m^2$	136
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.	143
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.	144
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.	145

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.	146
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).	150
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 <i>first</i> 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.	151
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.	151
5.12	The HFA field for (a) DJF, (b) MAM, (c) JJA and (d) SON averaged over the Standard ICE.	153
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.	156
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.	156

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.	161
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. . . .	162
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).	162
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.	182
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.	183
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.	184
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.	185

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.	186
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.	187
6.7	The difference in 8 year mean temperature/precipitation change under a doubling of CO_2 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 from 3 to 5 degrees. Negative values denote areas with no overlap. .	188
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.	197
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.	198

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.	199
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.	203
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.	204
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. .	206
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). . .	222
7.8	The implied distribution of feedbacks for three different values of the Entrainment Coefficient	223
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).	224

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.	225
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.	226
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.	227
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.	236
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. . . .	237
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.	244

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. . . . 247

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. 257

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. 259

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 $50,000 km^2$ in area. 260

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). 261

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. 266

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. 274

Nomenclature

Roman Symbols

IC Initial Conditions

AR4 IPCC Fourth Assessment Report

CO₂ 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

¹Terms 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 (~ 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 $40W/m^2$ (~ 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 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 system¹, are known as *forcings*. Natural forcings include changes in solar

¹Internal, 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 Roberock & 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*

¹A 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 forecast¹ 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

¹The 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 Project¹, 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

¹I 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

¹Of 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 than 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**

¹Climate 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 in 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 produced 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 compiled 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 (~ 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.

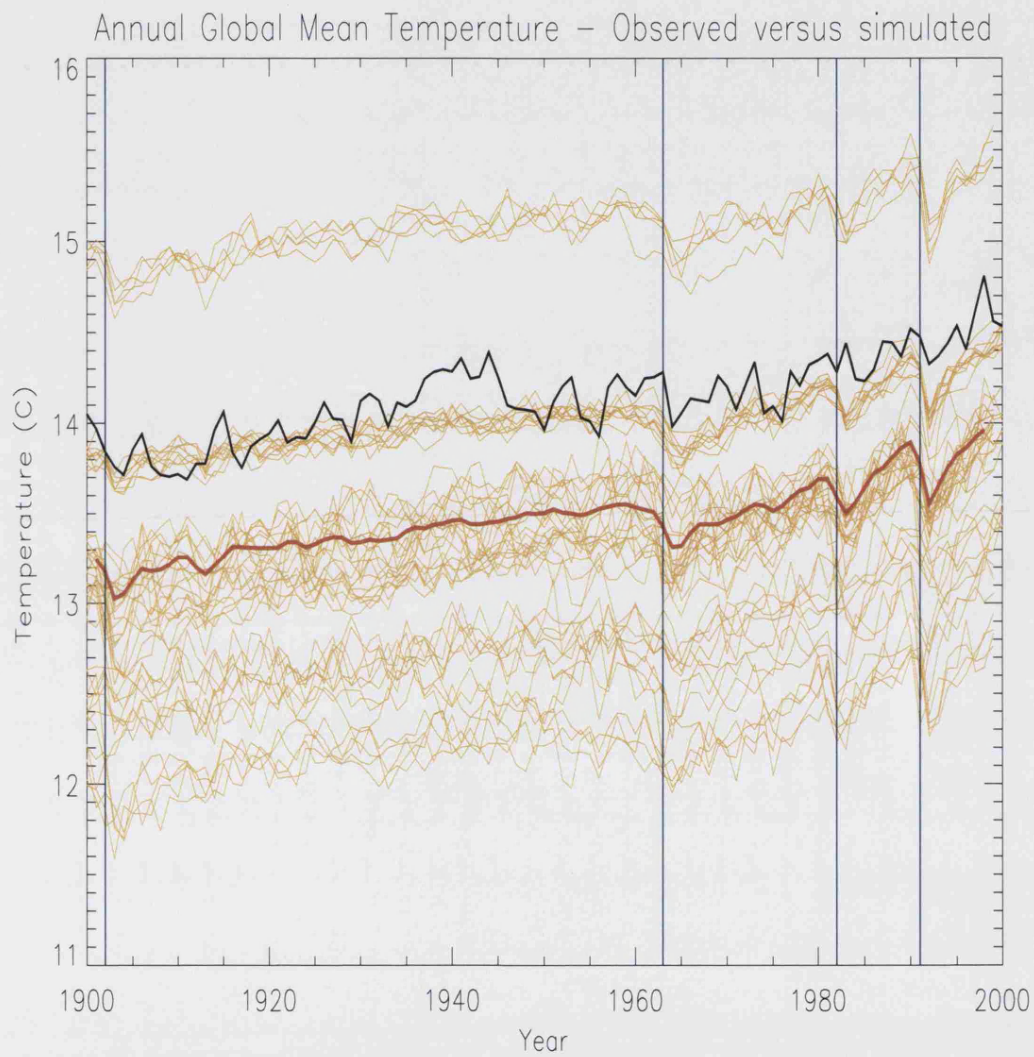


Figure 3.1: The absolute values of GMST from 47 simulations are plotted in yellow. The HadCRUT3 observations are plotted in black (the anomaly time series is offset using the 1961–1990 global mean (14.0 degrees Jones *et al.* (1999))). The multi-model mean is plotted in red. There is a difference of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

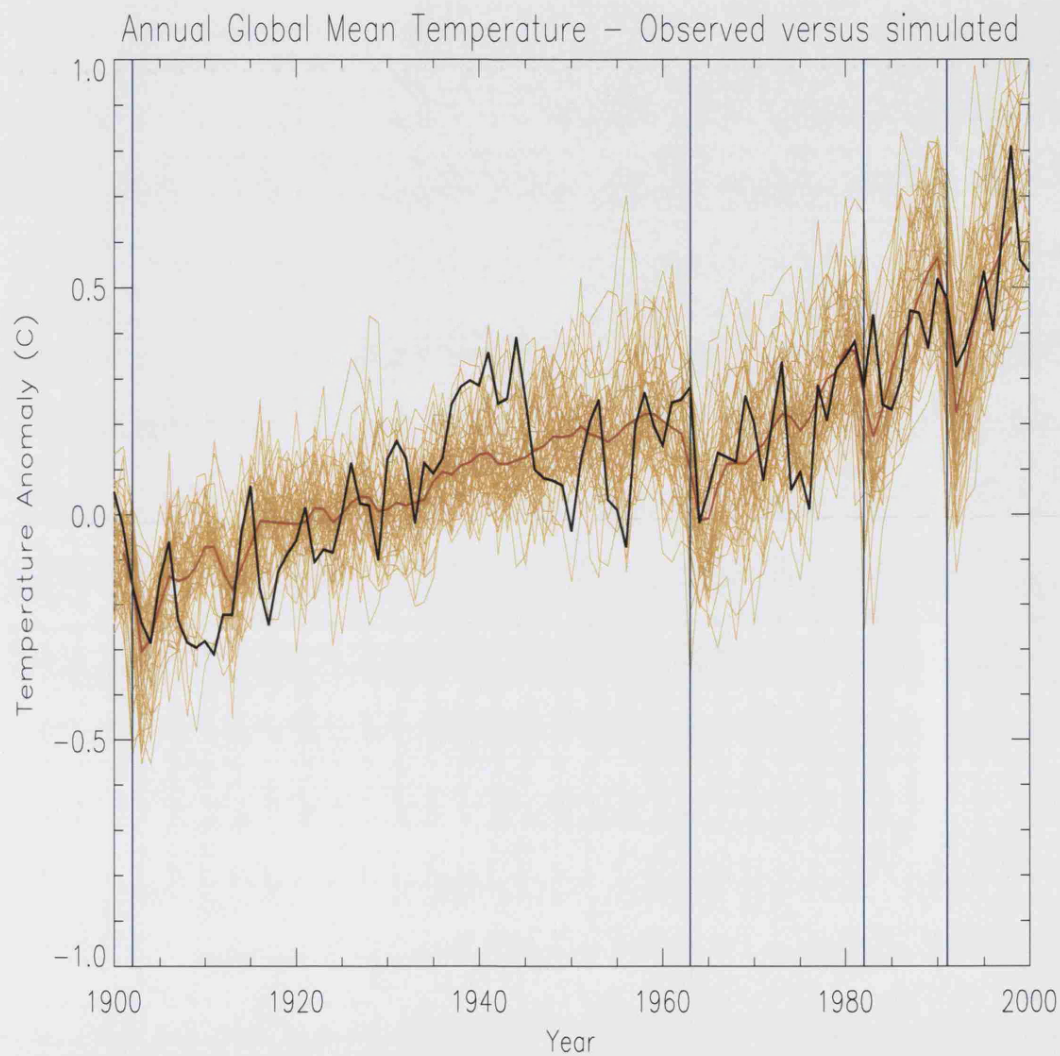


Figure 3.2: (Reproduction of IPCC Figure 8.1) Comparison of 47 simulations from 11 structurally distinct GCMs (yellow) used in the AR4 to HadCRUT3 observations (black). The multi-model mean is plotted in red. Each model simulation is “centred” by taking anomalies relative to 1901–1950. Blue lines show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

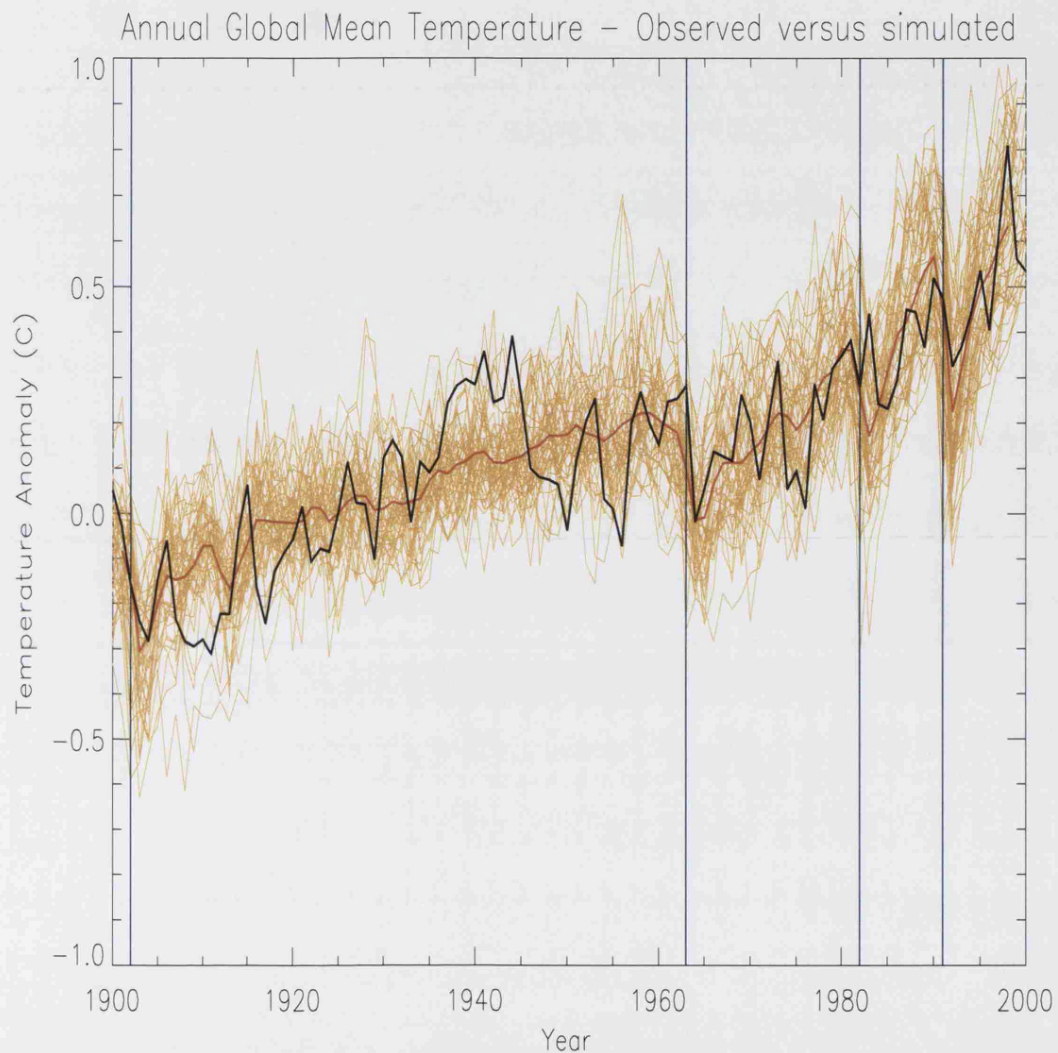


Figure 3.3: Comparison of 47 simulations from 11 structurally distinct GCMs (yellow) used in the AR4 to HadCRUT3 observations (black). The multi-model mean is shown in red. In this plot the model is centred using the mean 1901-1950 anomaly for each model (averaged over IC members). There is slightly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

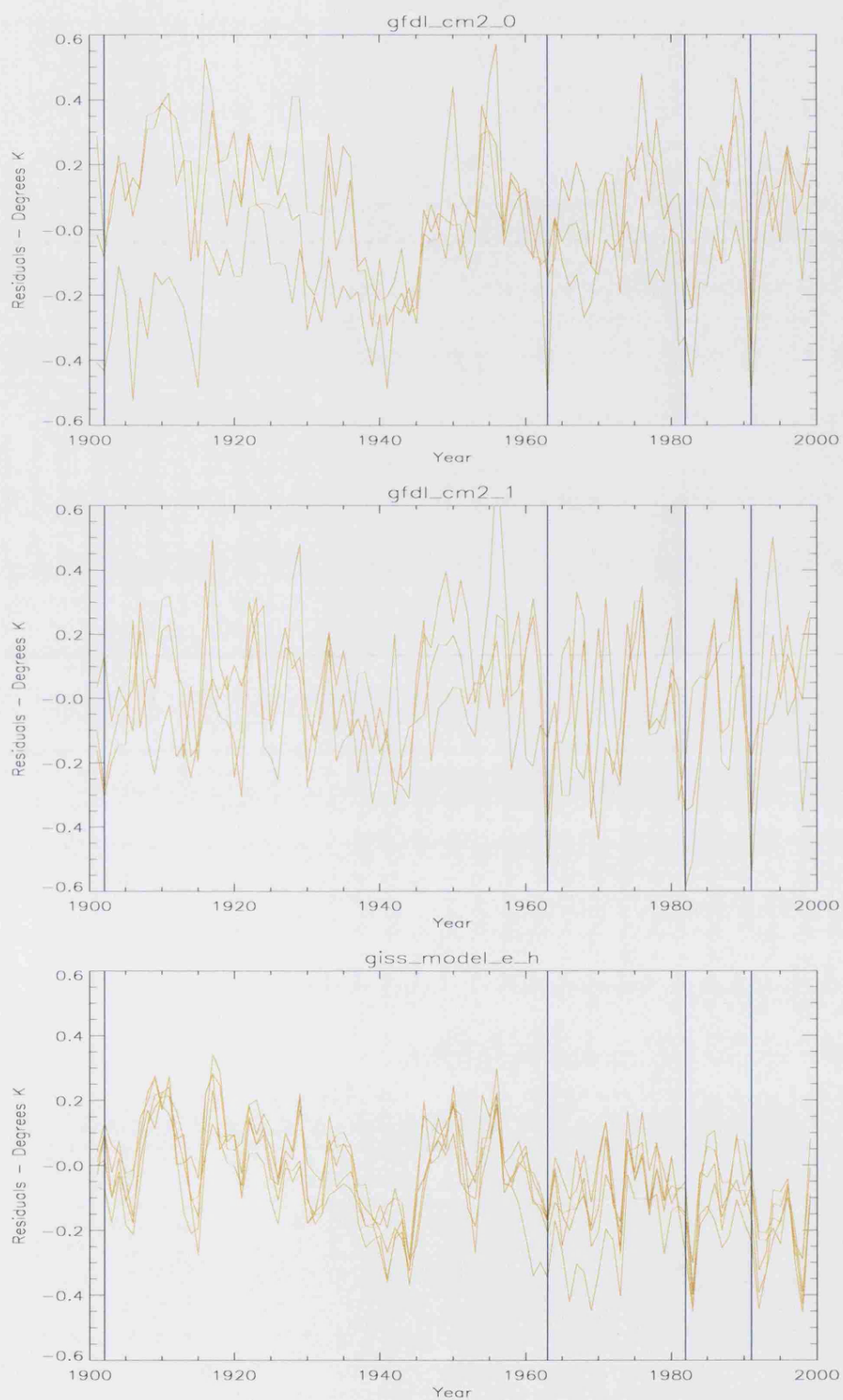


Figure 3.4: The residuals for 3 different GCMs are shown as a time series. Residuals for each simulation are found by subtracting the HadCRUT3 observations from each simulation (and adjusting for any differences in baseline 1901–1950 GMST).

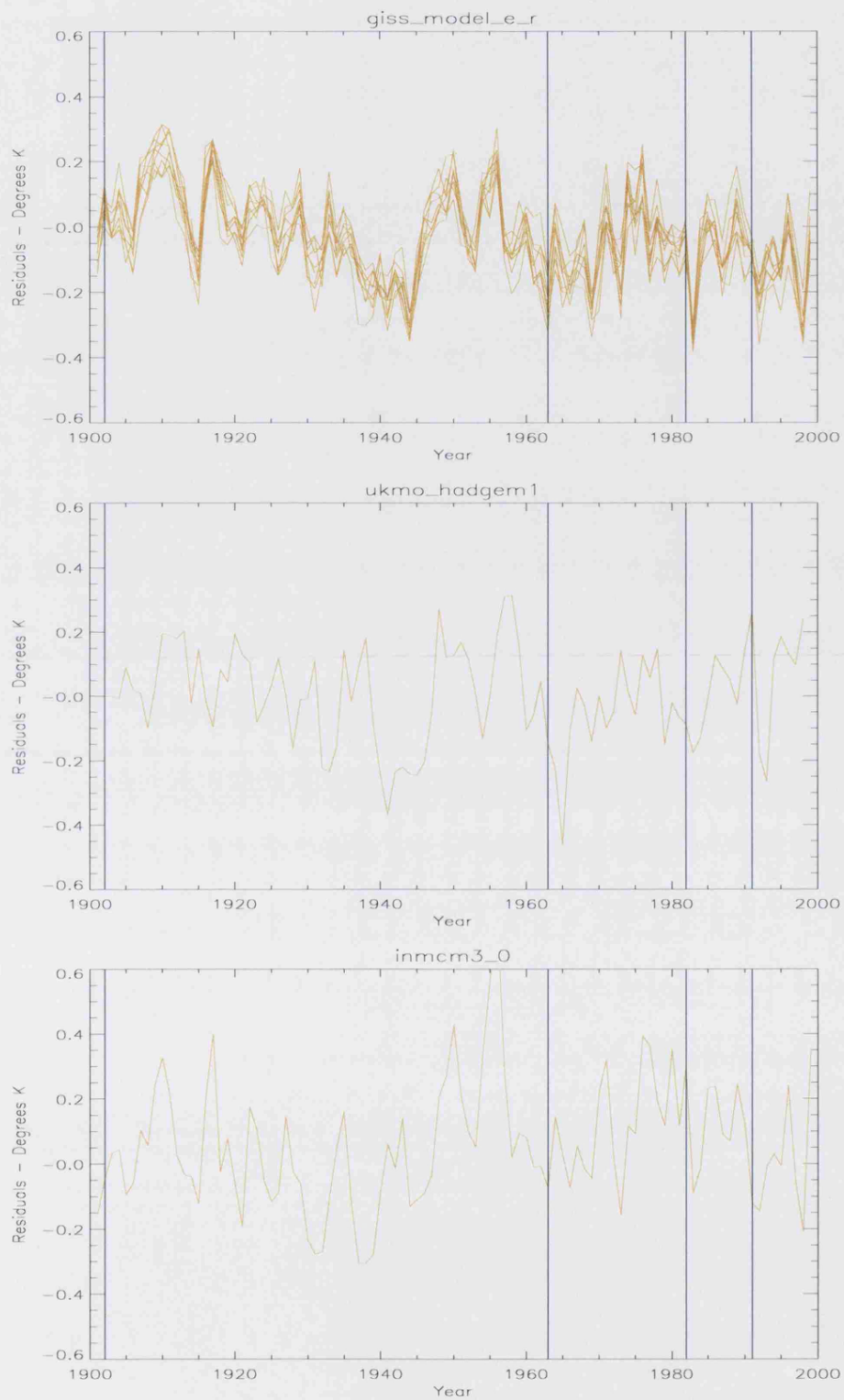


Figure 3.5: The residuals for 3 different GCMs are shown as a time series. Residuals for each simulation are found by subtracting the HadCRUT3 observations from each simulation (and adjusting for any differences in baseline 1901–1950 GMST).

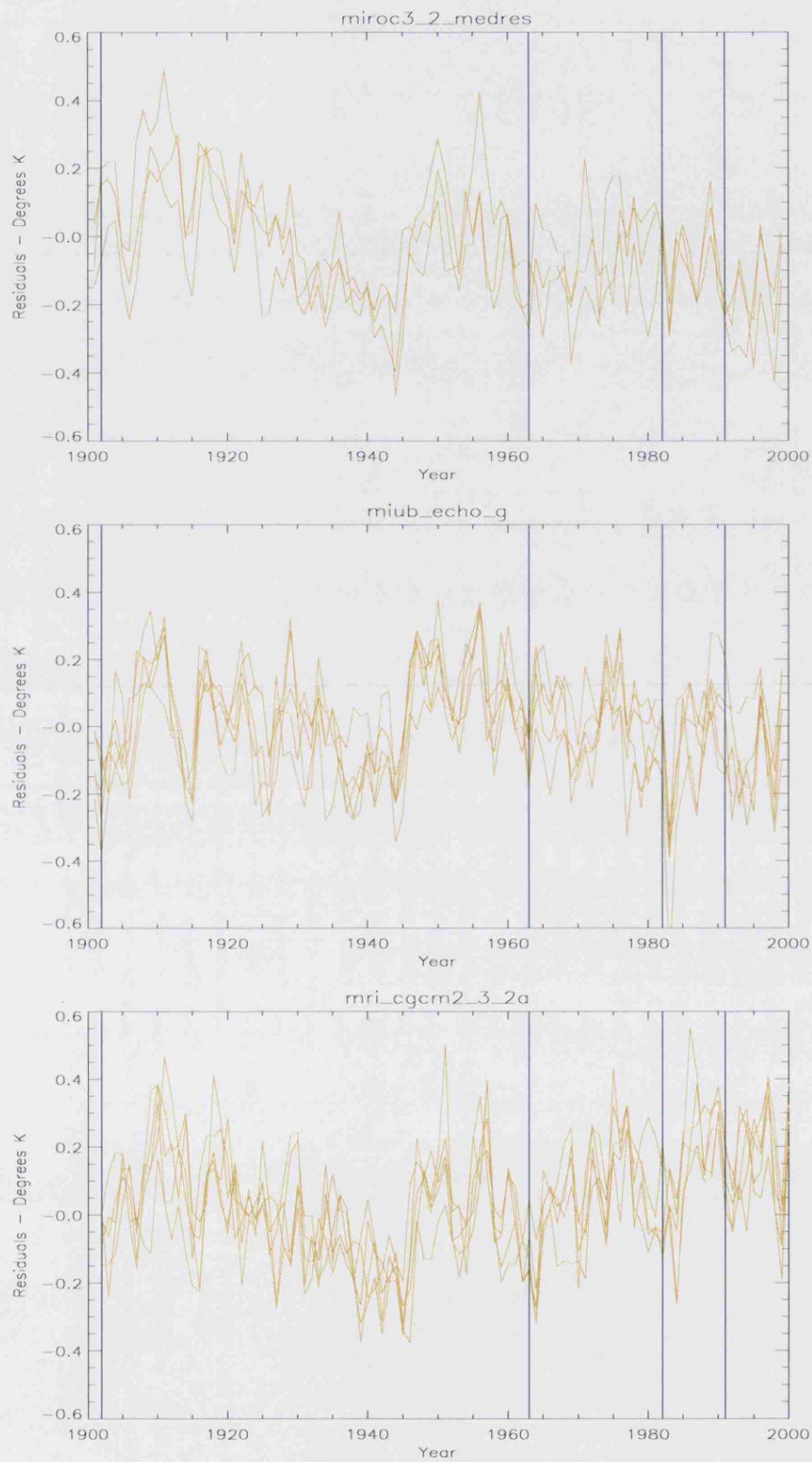


Figure 3.6: The residuals for 3 different GCMs are shown as a time series. Residuals for each simulation are found by subtracting the HadCRUT3 observations from each simulation (and adjusting for any differences in baseline 1901–1950 GMST).

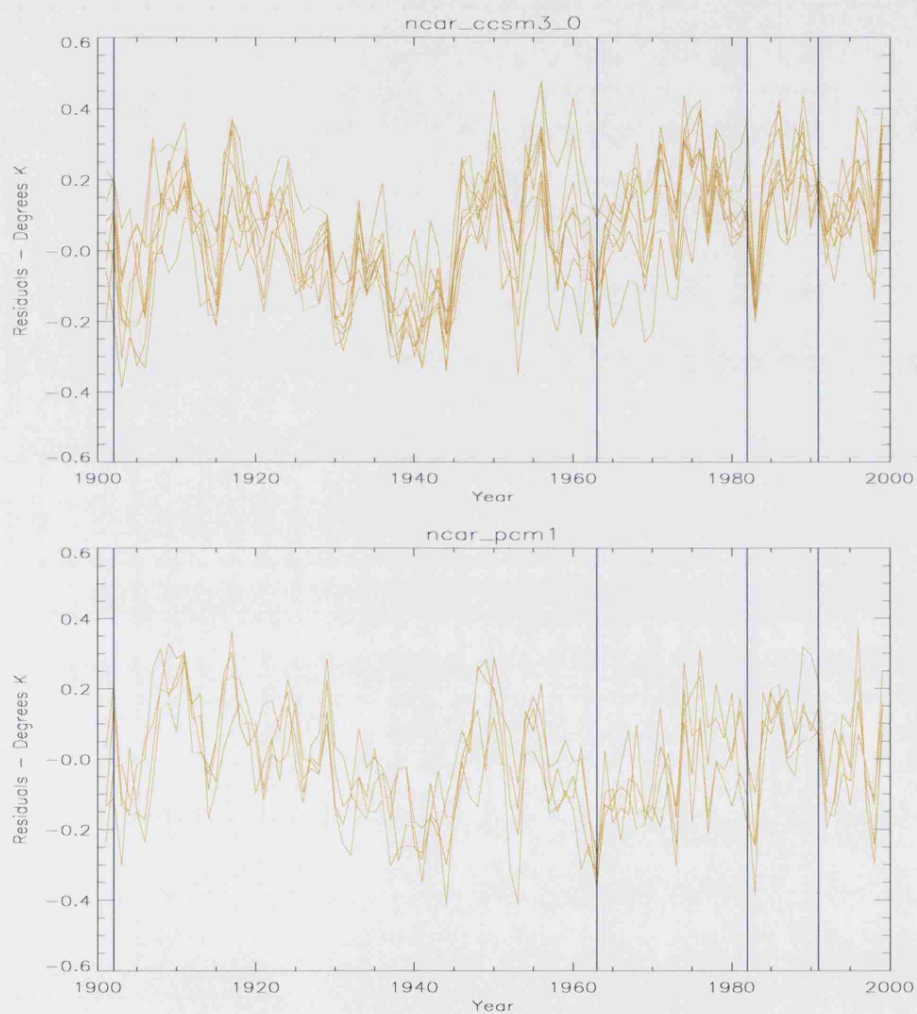


Figure 3.7: The residuals for 2 different GCMs are shown as a time series. Residuals for each simulation are found by subtracting the HadCRUT3 observations from each simulation (and adjusting for any differences in baseline 1901–1950 GMST).

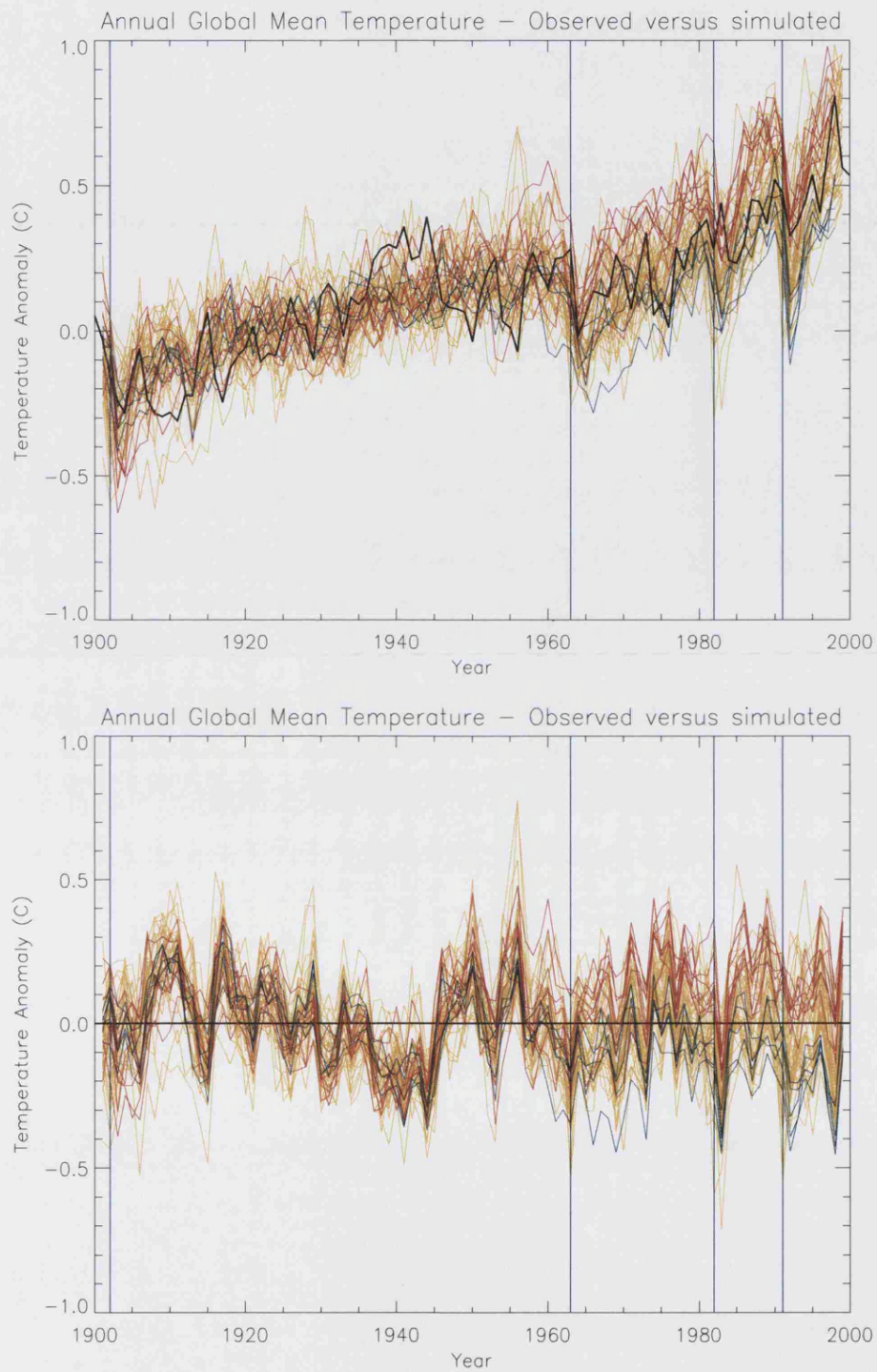


Figure 3.8: The time series of 47 GCM simulations is plotted against observations as 1901–1950 anomalies (top) and as residuals (bottom). The NCAR PCM1 GCM is highlighted in red and the GISS-h model in blue. The highlighted models overlap in the first half of the 20th Century but diverge from 1960 onwards.

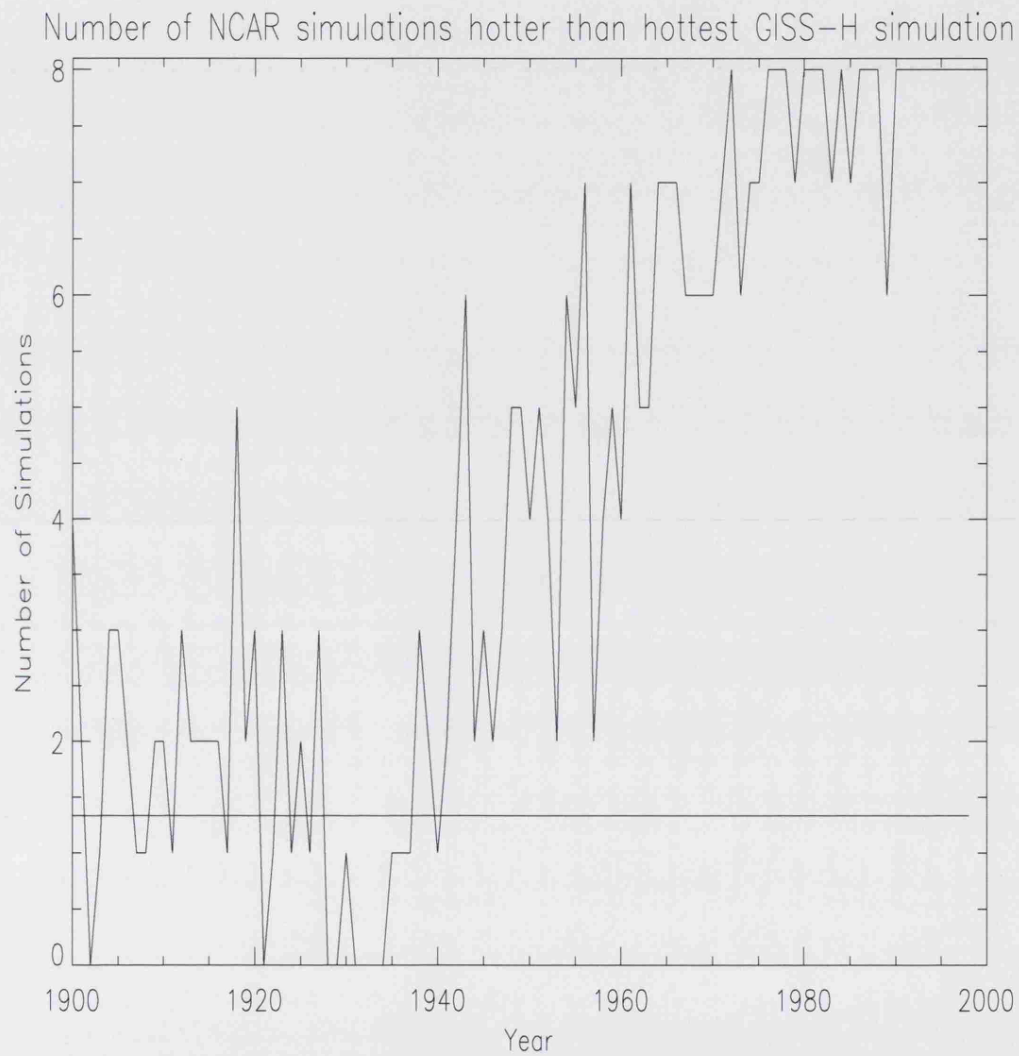


Figure 3.9: The number of NCAR PCM1 simulations that are hotter than the hottest GISS-h simulation over the 20th Century. The horizontal line shows the number of simulations we would expect to be hotter, on average, at each time point if the models were sampling from the same distribution. The horizontal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

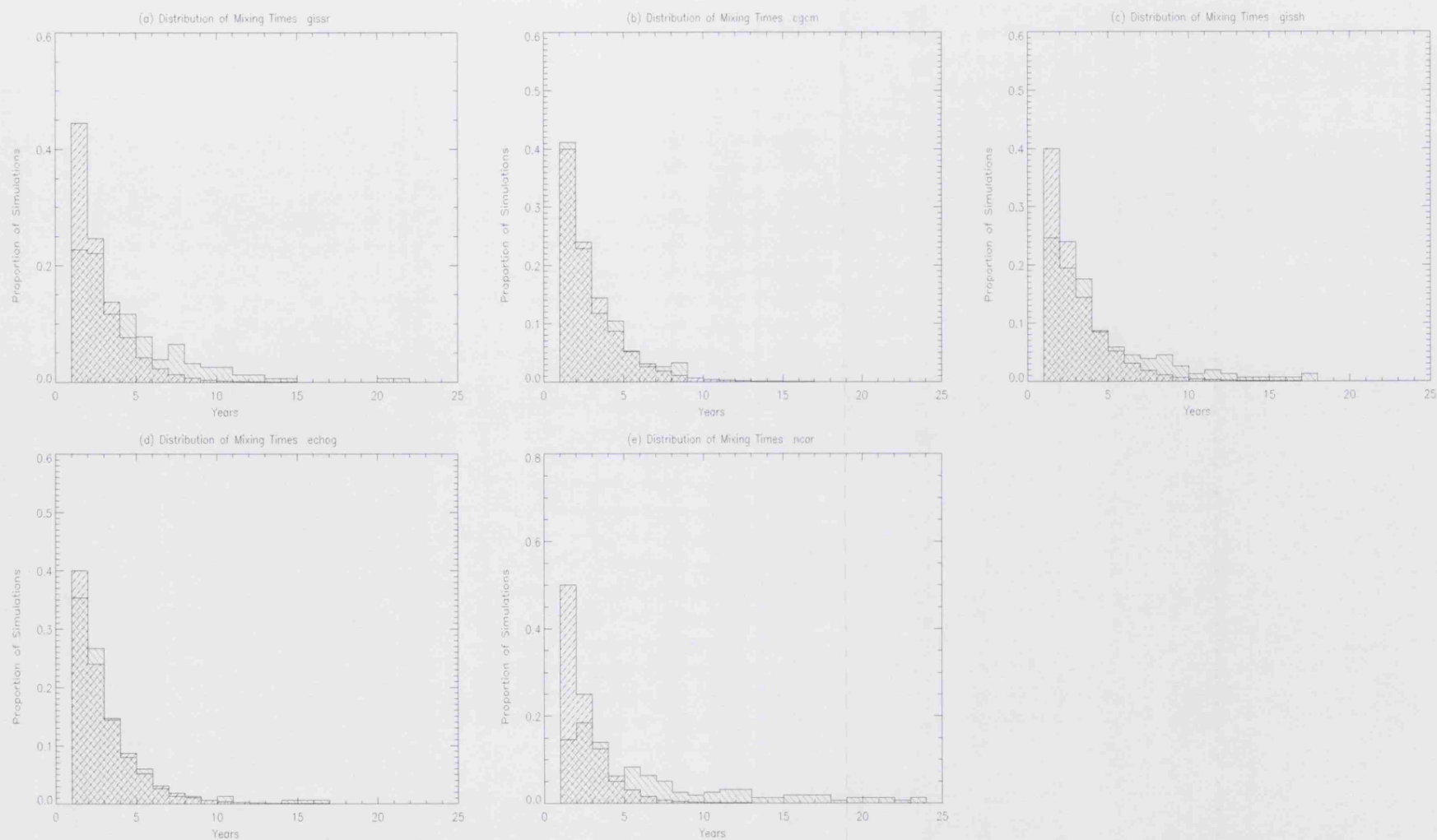


Figure 3.10: The distribution of mixing times for 5 GCMs - *mri-cgcm2-3-2a*, *miub-echo-g*, *giss-echo-e-h*, *giss-echo-e-r* and *ncar-ccsm3* with 5, 5, 5, 9 and 8 simulations respectively.

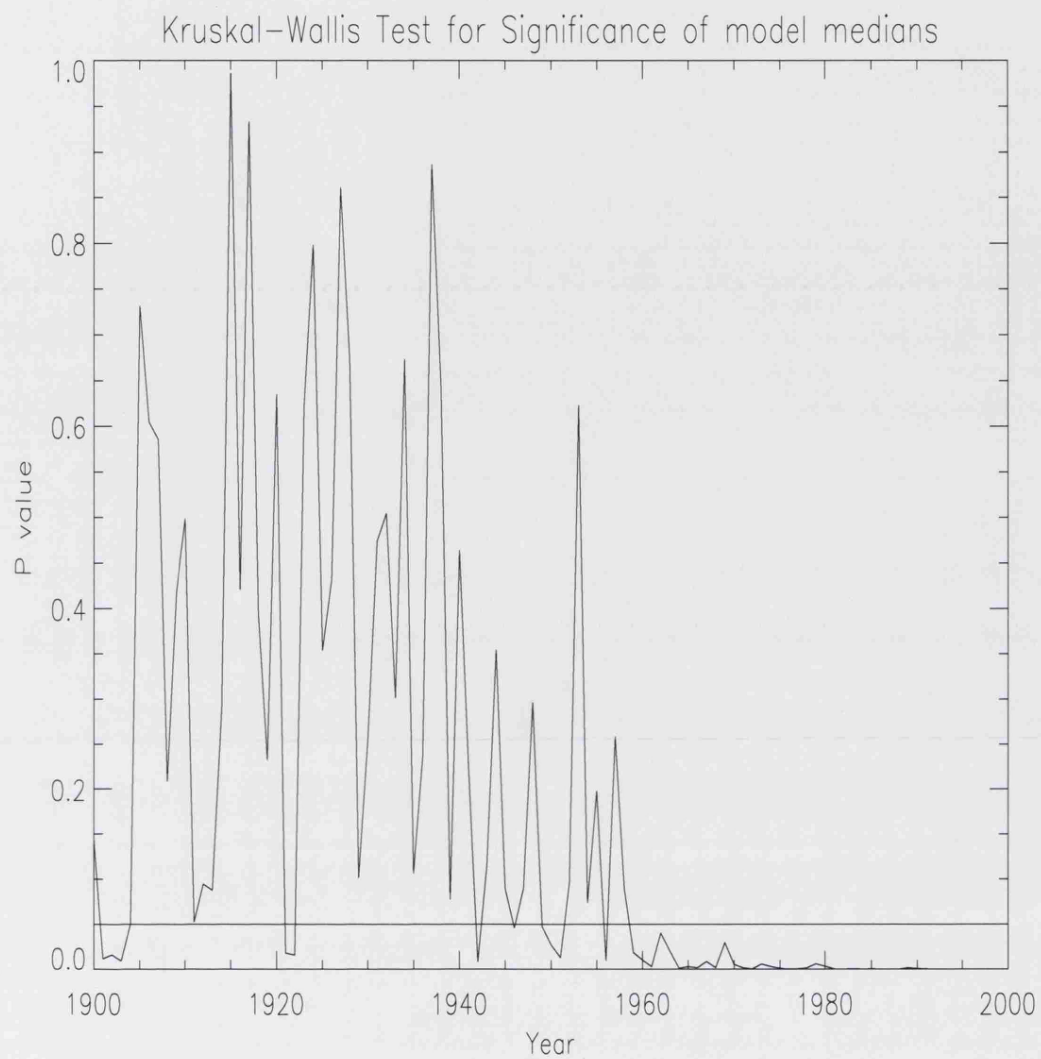


Figure 3.11: The p -values for the Kruskal-Wallis test are shown for the 20th Century. Low values suggest evidence against the null hypothesis that all five models have the same median.

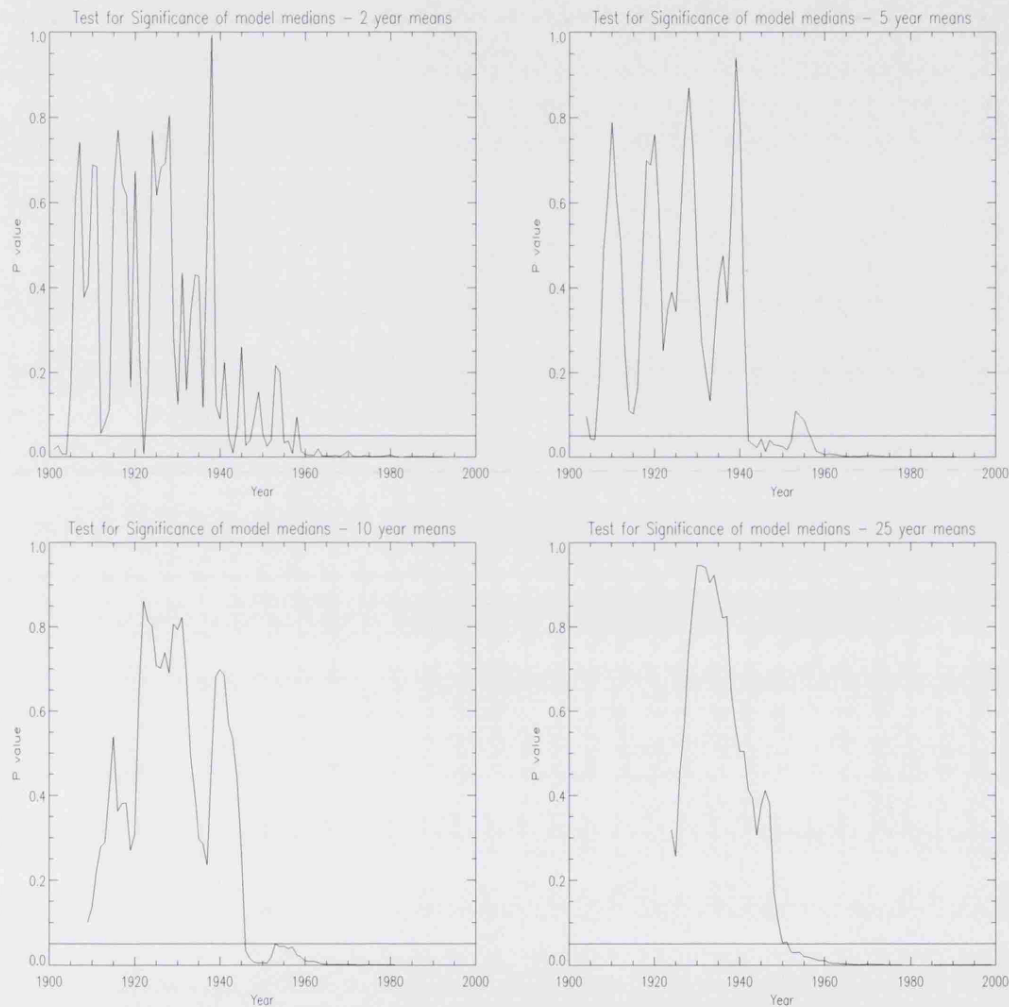


Figure 3.12: The p -values for the Kruskal-Wallis test are shown for the 20th Century for multi-year running medians of 2, 5, 10 and 25 year means respectively. In all cases, the test is non-significant during the first half of the 20th Century, the becomes significant towards the end.

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 system¹. 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,

¹It 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:

$$4\pi R^2 \epsilon \sigma T^4 \quad (4.3)$$

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,

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

Which reduces to:

$$S(1 - \alpha) = KT^4 \quad (4.5)$$

where K is a constant, or

$$T \propto S^{1/4} \quad (4.6)$$

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

¹Actually 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 $\sim 10,000 \text{ km}^2$ 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

¹A 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 (~ 40 Mb) 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 ($\sim 15\text{Mb}$ 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 oceans¹.

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).

$$\frac{dT}{dt} = \frac{F_{x2co2}}{hc} \exp\left(\frac{\lambda}{hc}\right) \quad (4.7)$$

Where $\frac{dT}{dt}$ is the change in GMST, T , with respect to time, t , in years. CS is estimated by $\frac{F_{x2co2}}{\lambda}$.

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 out 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 due to 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:

$$A = (area51_3 - area52_3) - (area51_1 - area52_1) \quad (4.8)$$

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:

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

Where r = the radius of the Earth ($\sim 6371\text{km}$), $nlons$ =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 vortices 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.

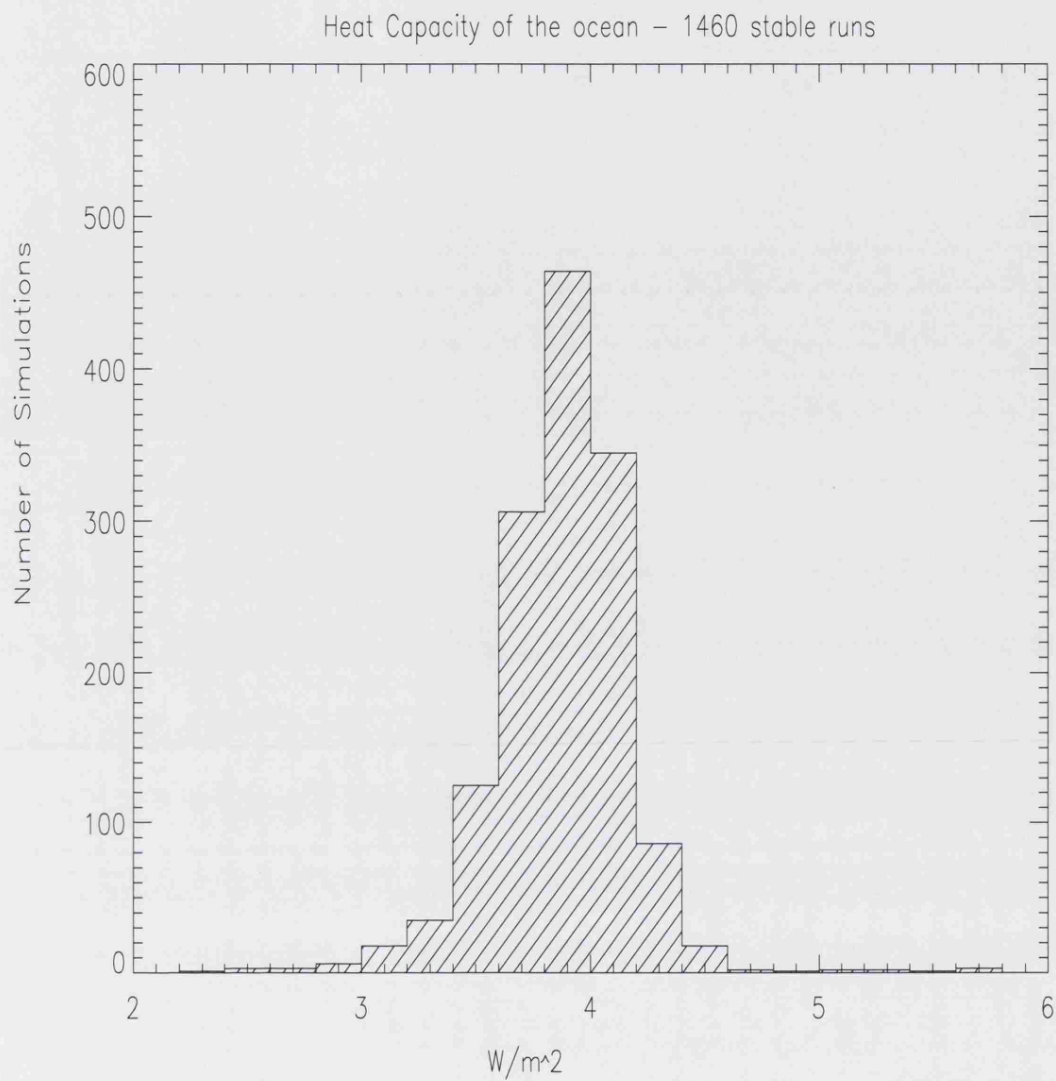


Figure 4.1: The global mean heat capacity in the doubled CO_2 phase is plotted over 1460 quality controlled simulations.

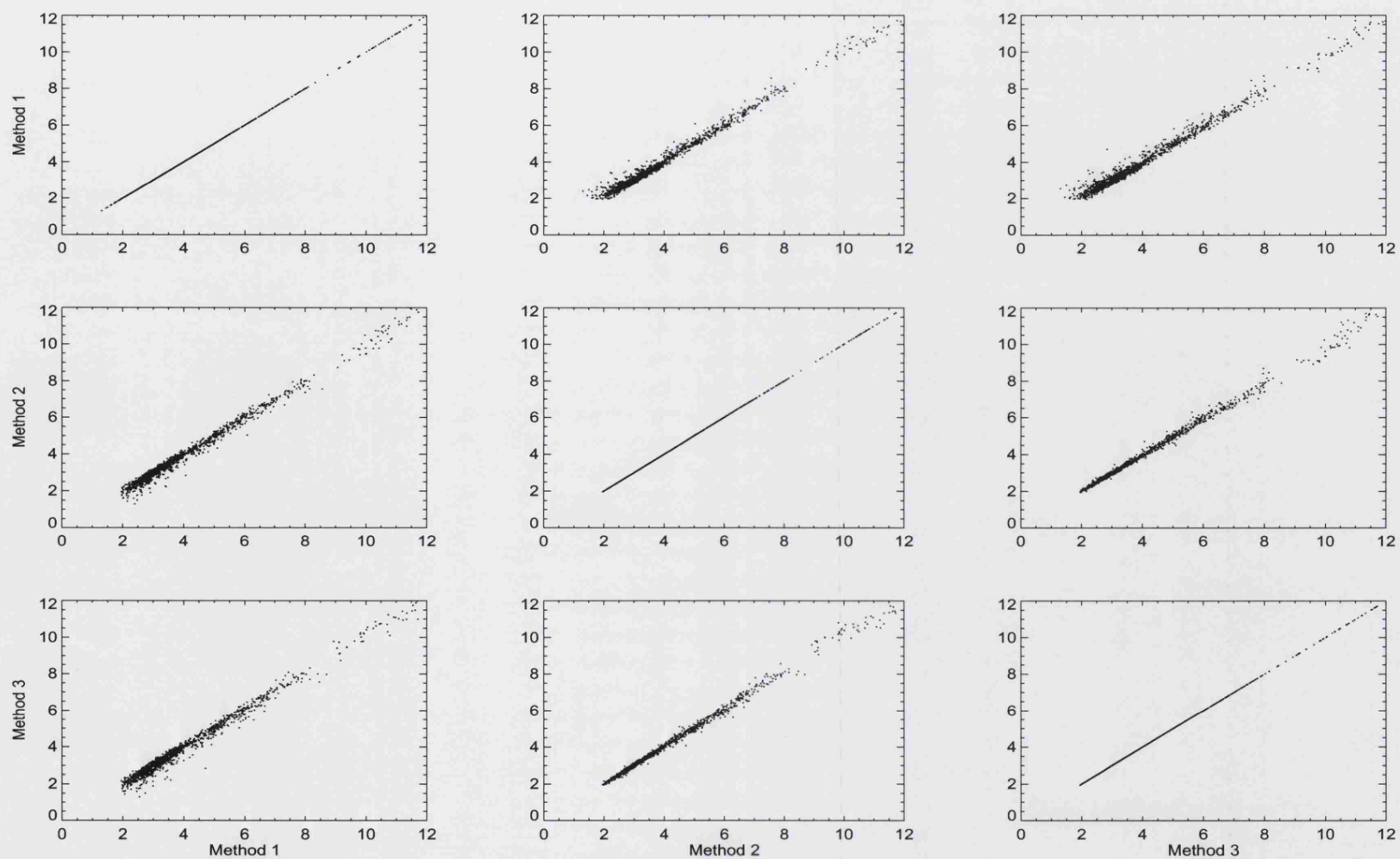


Figure 4.2: Estimates of CS are plotted against each other for three different methods over 1460 quality controlled simulations. There is a strong linear relationship between each of the methods.

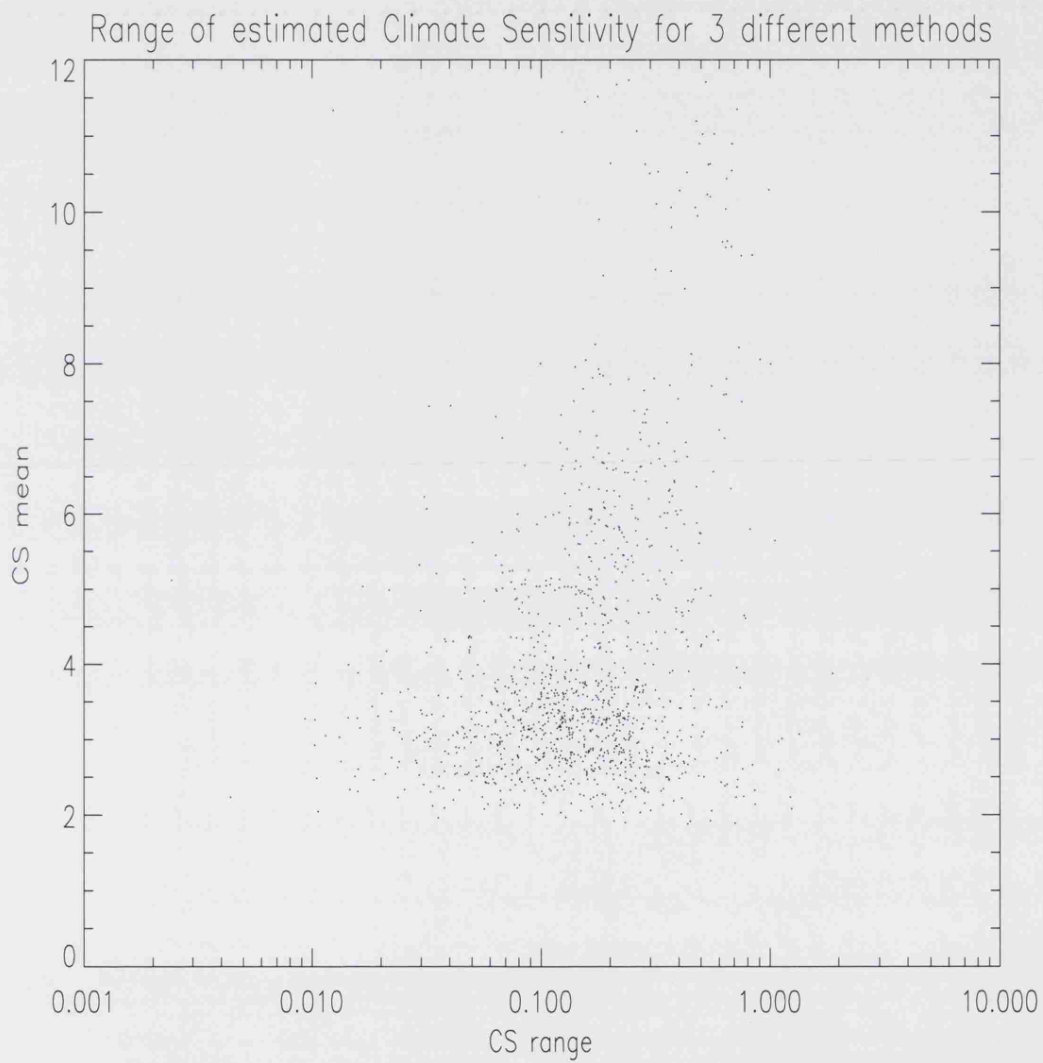


Figure 4.3: The range in estimated CS for three different methods is plotted against the mean for 1460 quality controlled simulations.

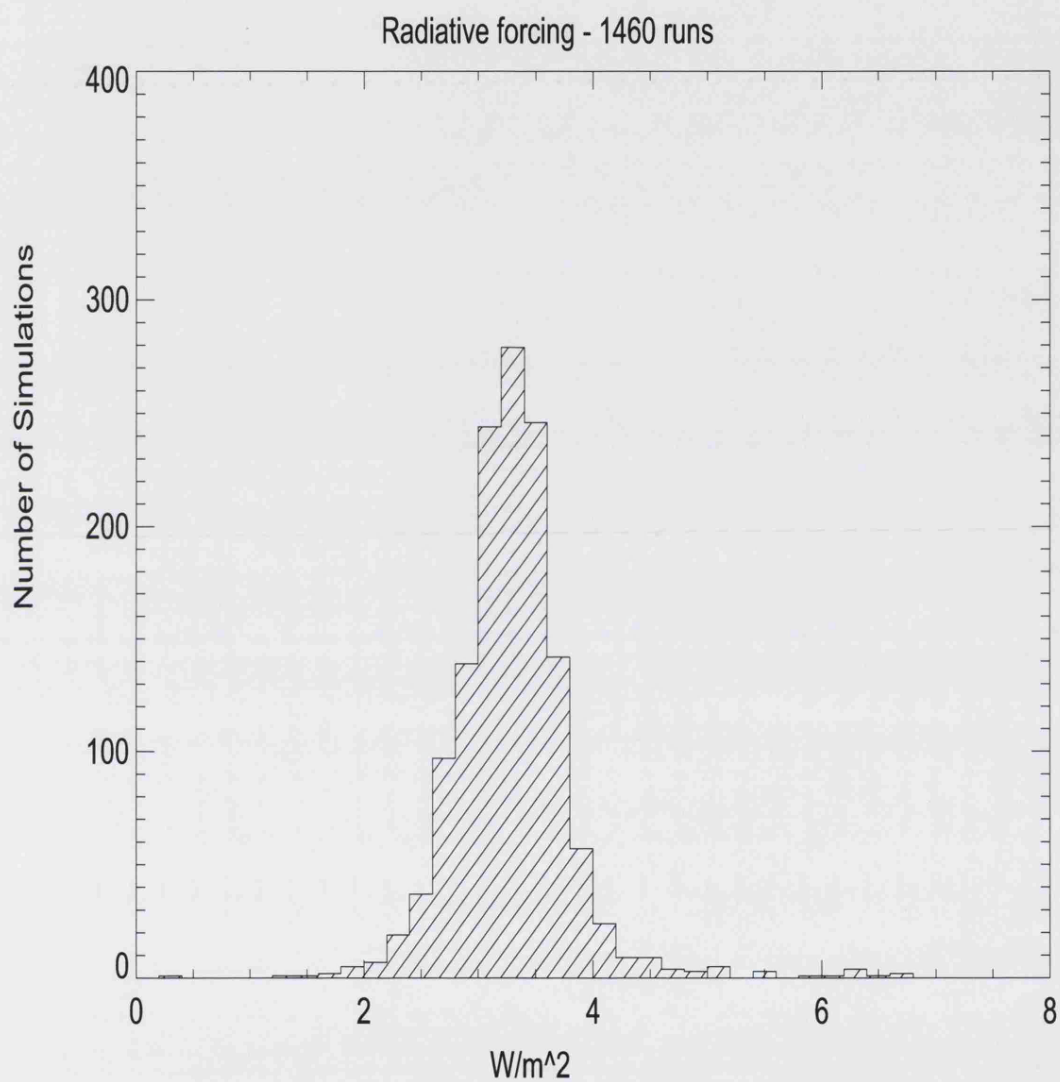


Figure 4.4: The distribution in top of atmosphere radiative flux imbalance in the doubled CO_2 phase is shown for 1460 quality controlled simulations. The method used to estimate this flux imbalance is the exponential fit of temperature change. There is a wide range of estimates for heat capacity, ranging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

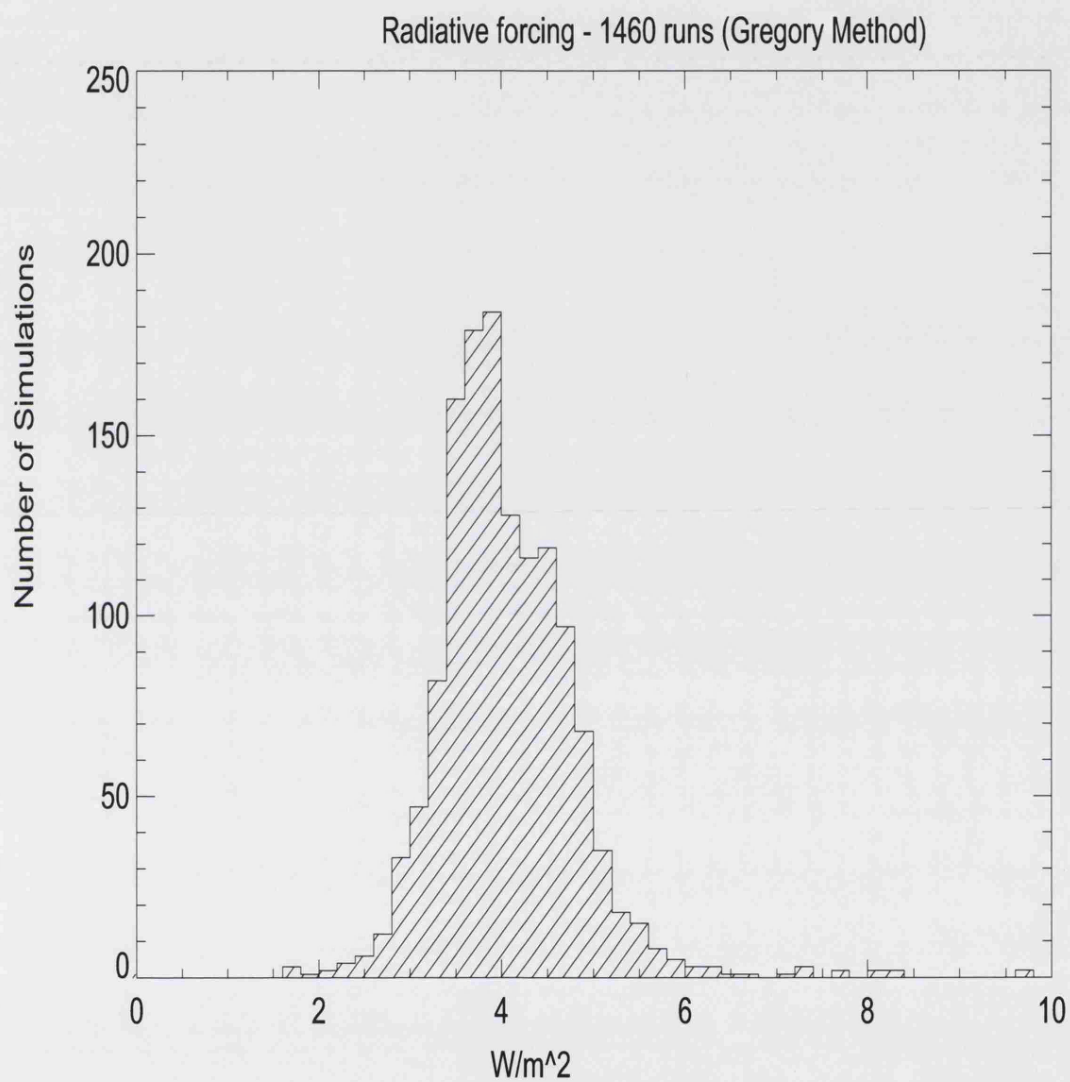


Figure 4.5: The top of atmosphere radiative flux imbalance in the doubled CO_2 phase is plotted over 1460 quality controlled simulations. The method used to estimate this flux imbalance is the Gregory plot method. There is a wide range of estimates for heat capacity, ranging from below 2 W/m^2 to over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

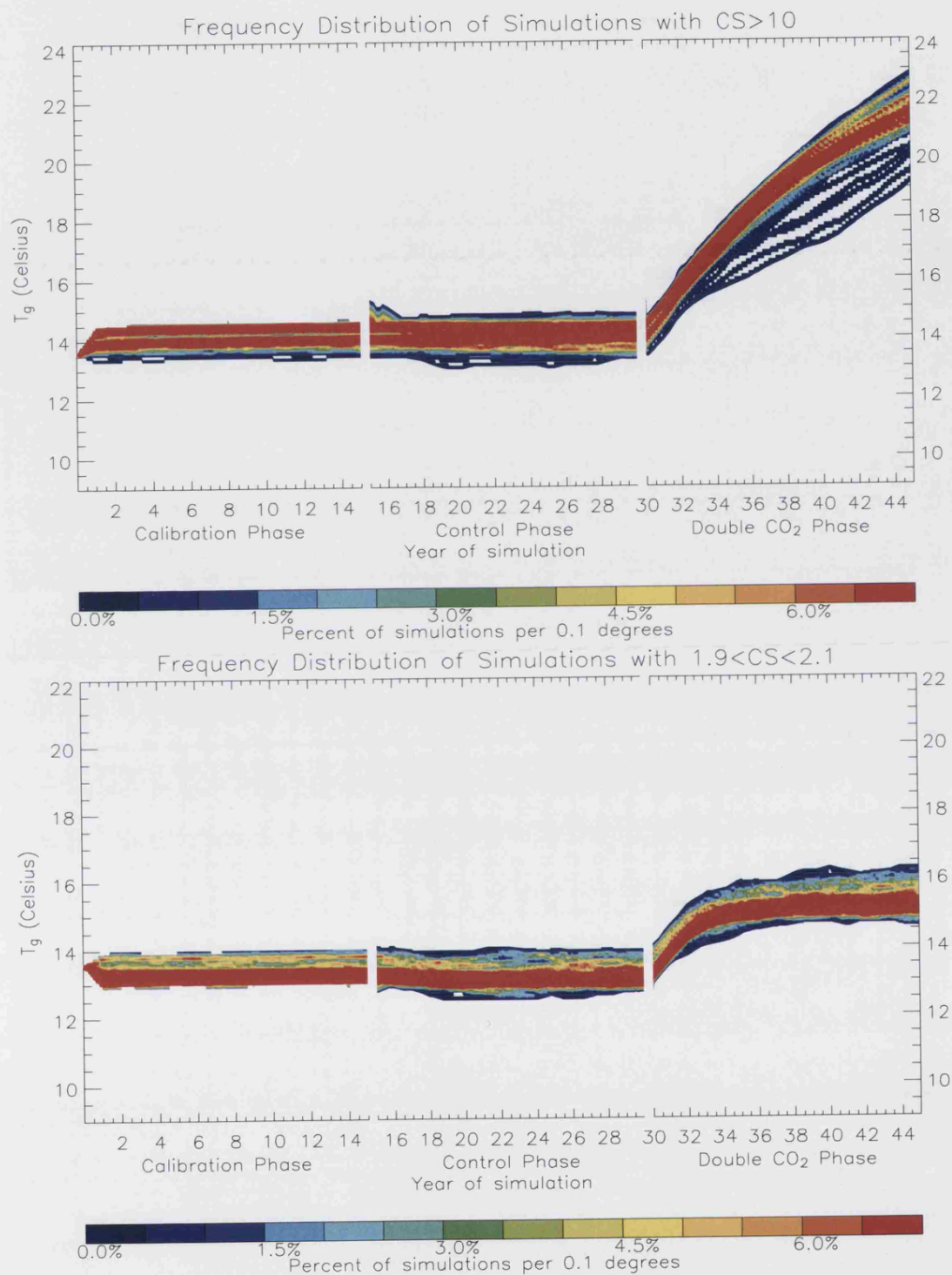


Figure 4.6: The GMST time series over the 3 experimental phases of the CPDN experiment for 350 simulations with estimated CS over 10 degrees Celsius (top). Also shown is the time series for 822 simulations with 2 degrees CS (bottom). Whereas the 2 degree simulations seem to have reached an equilibrium 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.

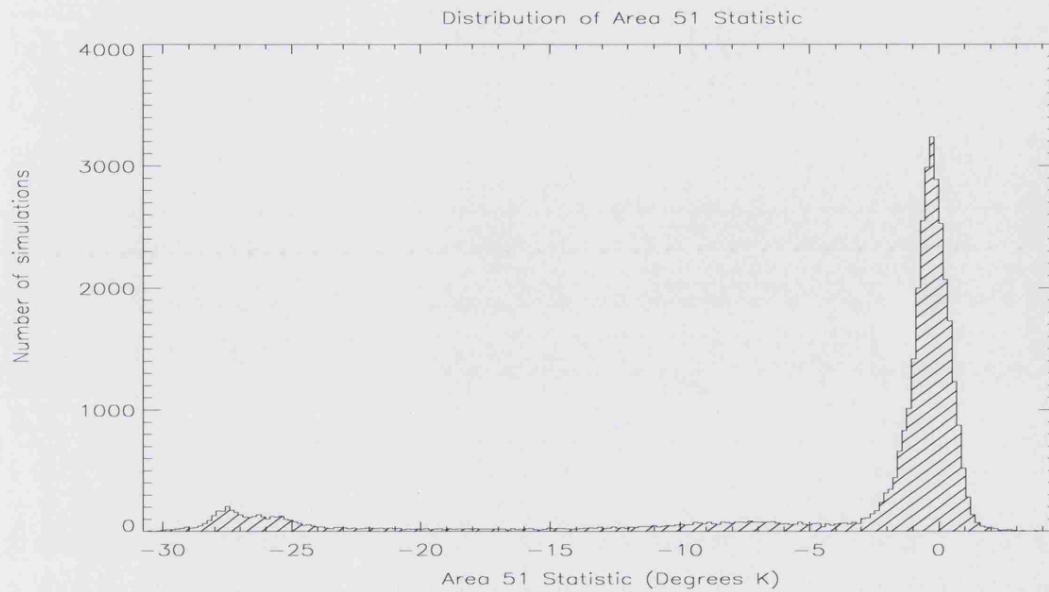


Figure 4.7: The distribution of values of the Area 51 anomaly for 45644 simulations before applying any quality control. The distribution is clearly bi-modal, representing simulations that do not exhibit a negative feedback (peak around 0), those that have (peak around -27) and a smaller number of intermediate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

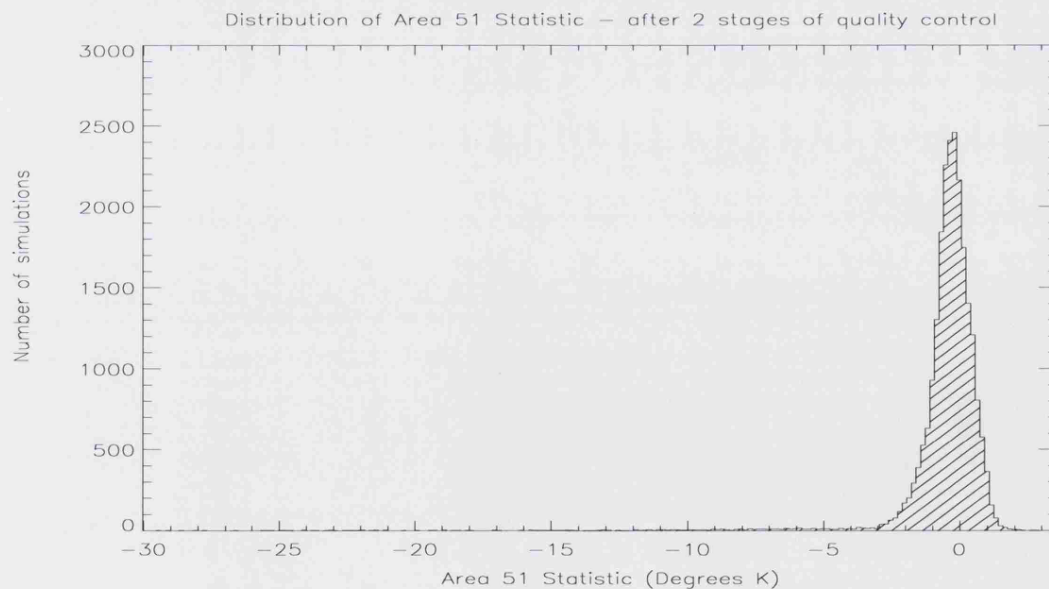


Figure 4.8: The distribution of values of the Area 51 anomaly for 23050 complete simulations with a non-significant GMST drift. The distribution is uni-modal about 0, with a long tail in the negative values.

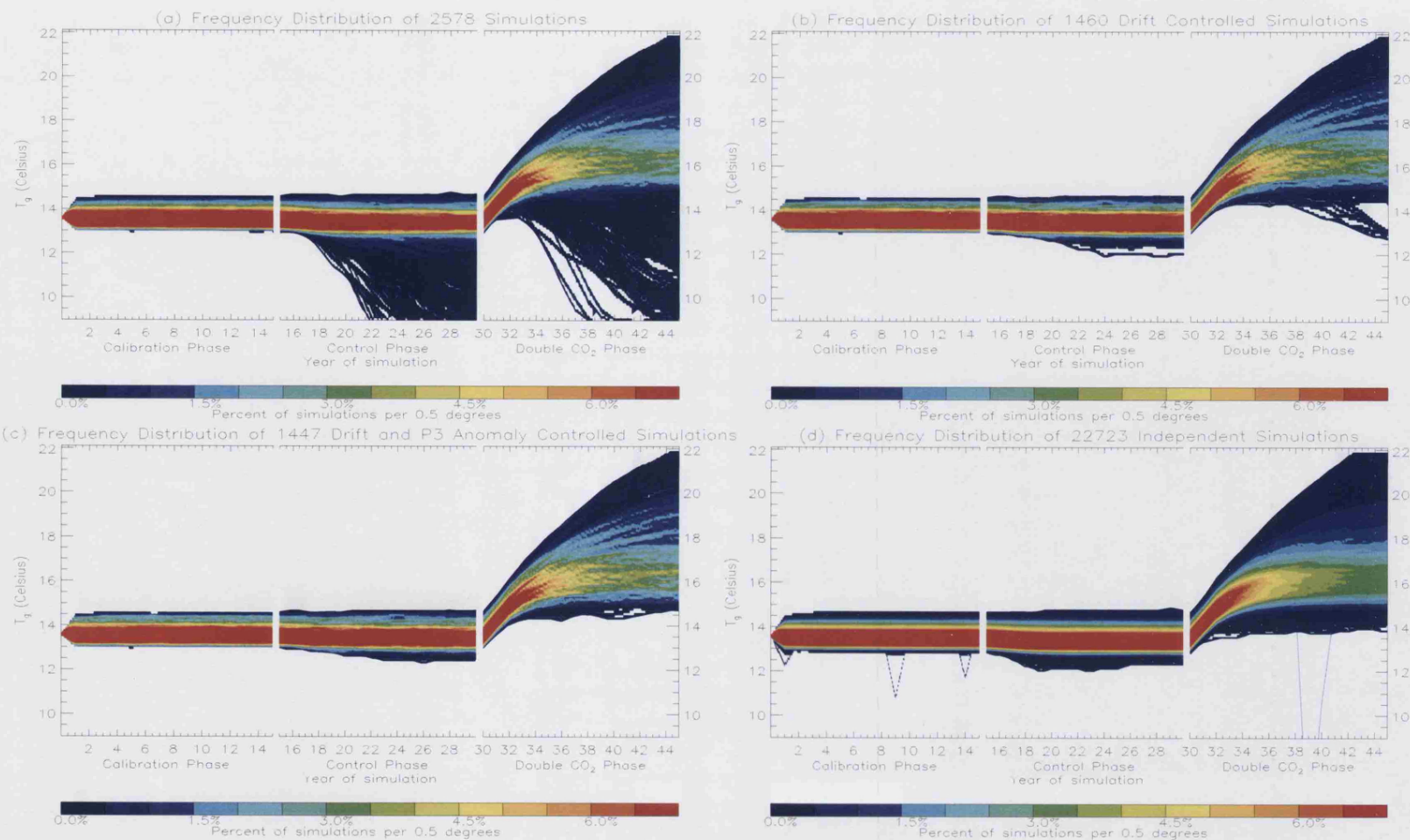


Figure 4.9: Panel (a) shows the time series of 2578 simulations with no quality control applied. Panel (b) shows the same time series, applying only the first two stages of quality control, leaving 1460 simulations. Panel (c) shows the time series with full quality control applied, leaving 1447 simulations. 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 than 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 than 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 from 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 distinct¹ 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 larger 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.

¹Some 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 is 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.

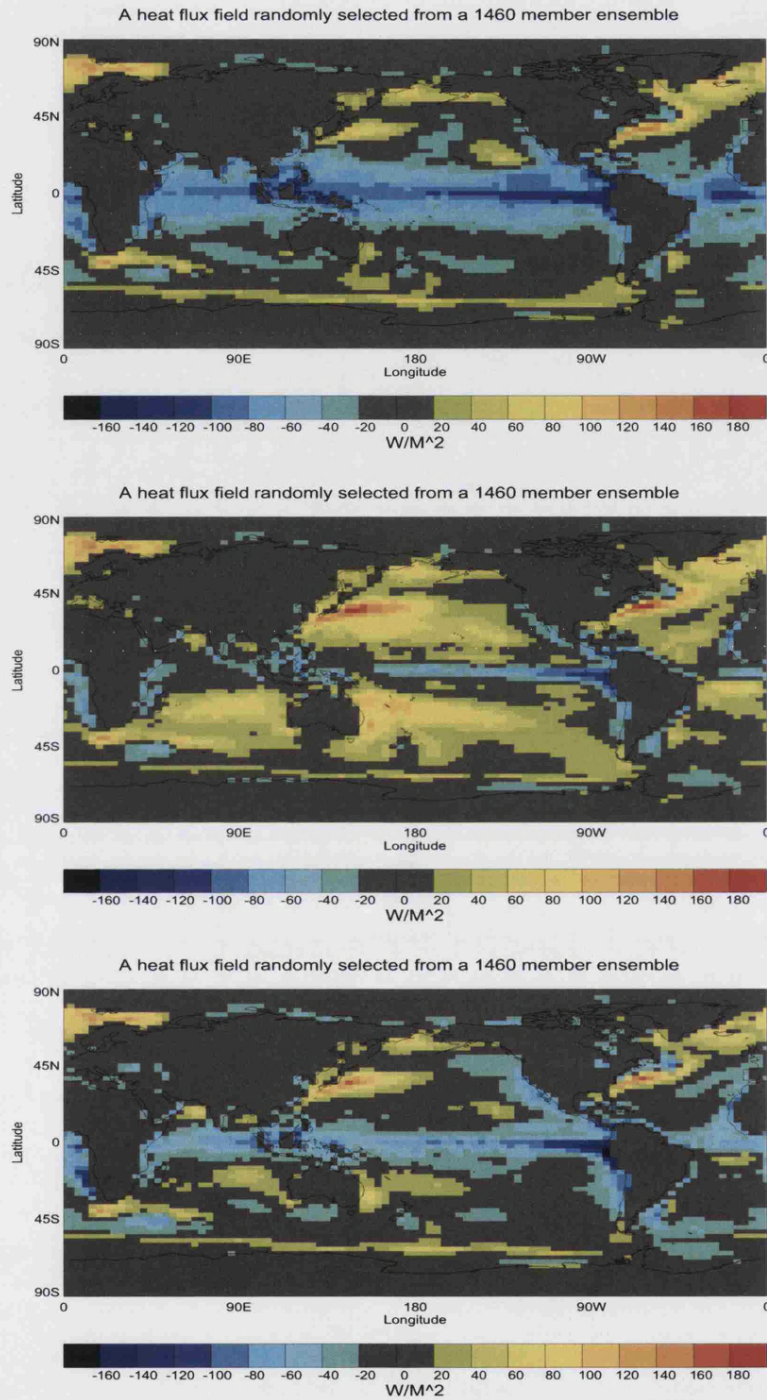


Figure 5.1: Three randomly selected HFA fields from $PPE_{quality}$. There are significant differences in HFA by region. Some areas require a reduction of heat in the ocean by more than $150W/m^2$ whereas others require more than $200W/m^2$ to be added. A HFA of $200W/m^2$ is approximately the same effect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

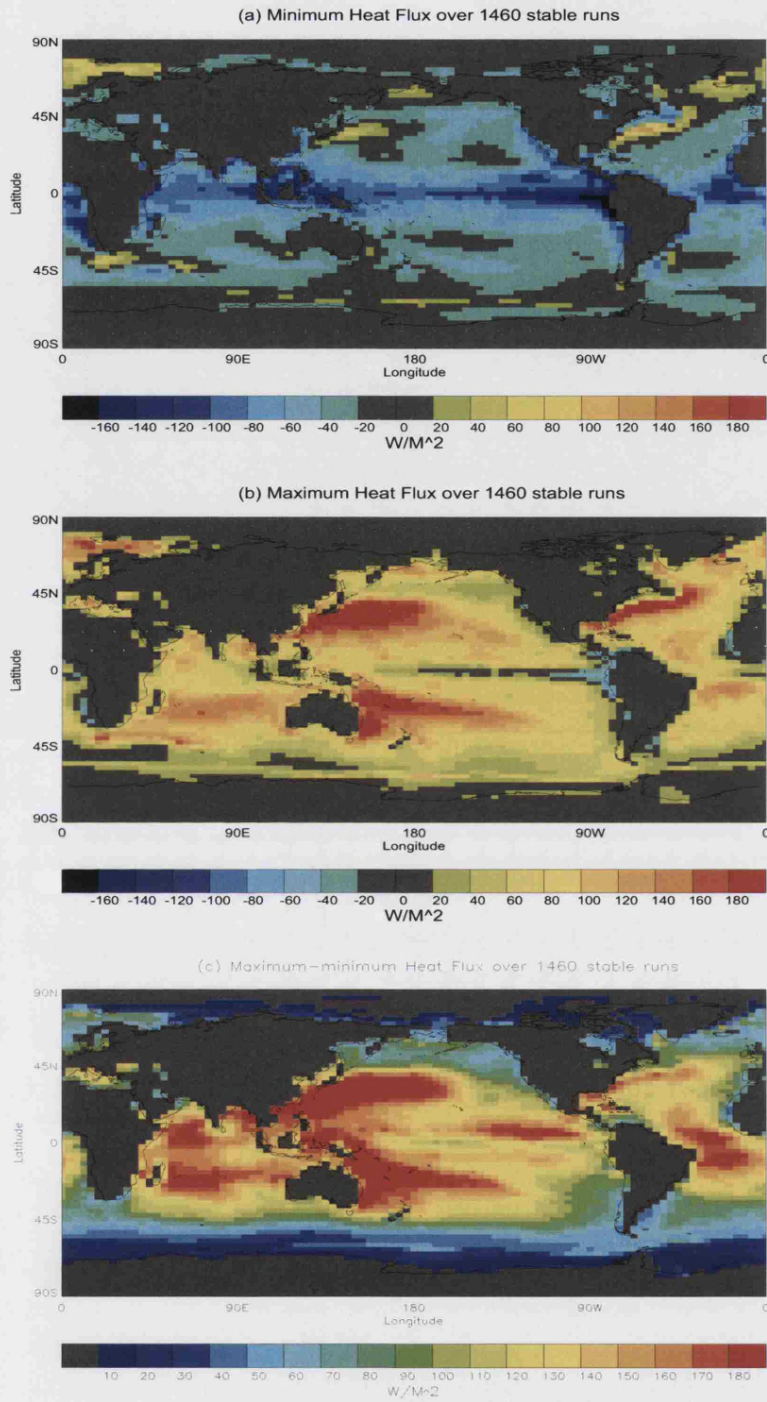


Figure 5.2: The three panels show the (a) minimum, (b) maximum and (c) range of the HFA field for $PPE_{quality}$. The bounding box relates to the spread of the ensemble at each grid-box. Positive values denotes heat into the ocean. The range of values shown in panel (c) can be as large as $200W/m^2$.

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.

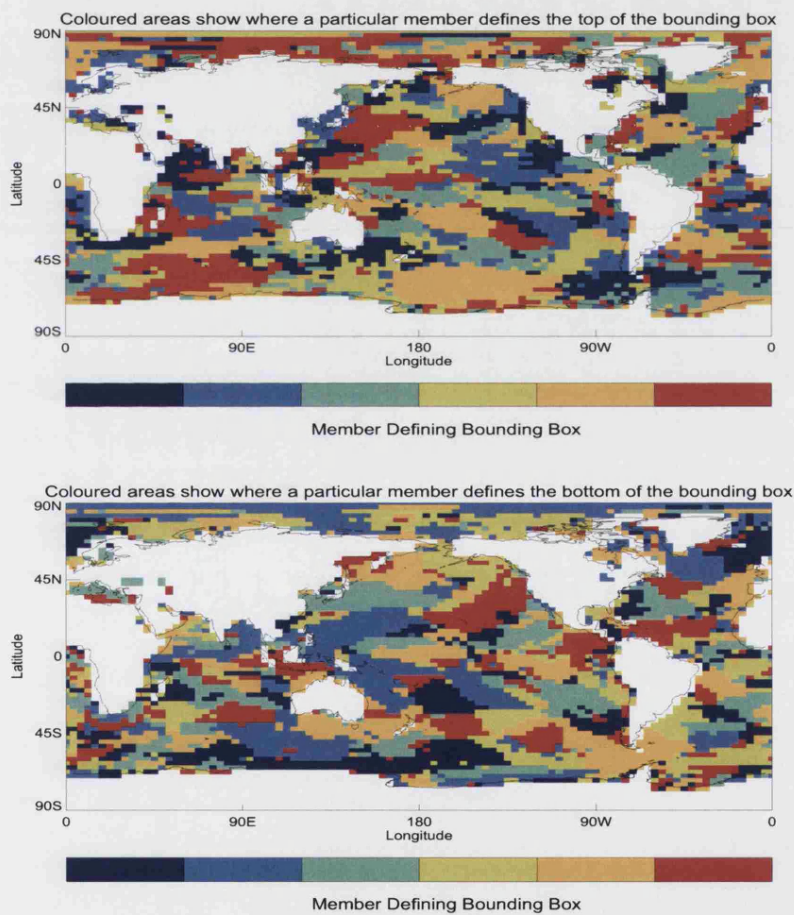


Figure 5.3: A colour is plotted for each of 6 members of the Initial ICE where it defines the bounding box. The top picture shows the top of the bounding box, and the bottom the bottom of the bounding box. The roughly even distribution of colours indicates that all members contribute to defining the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

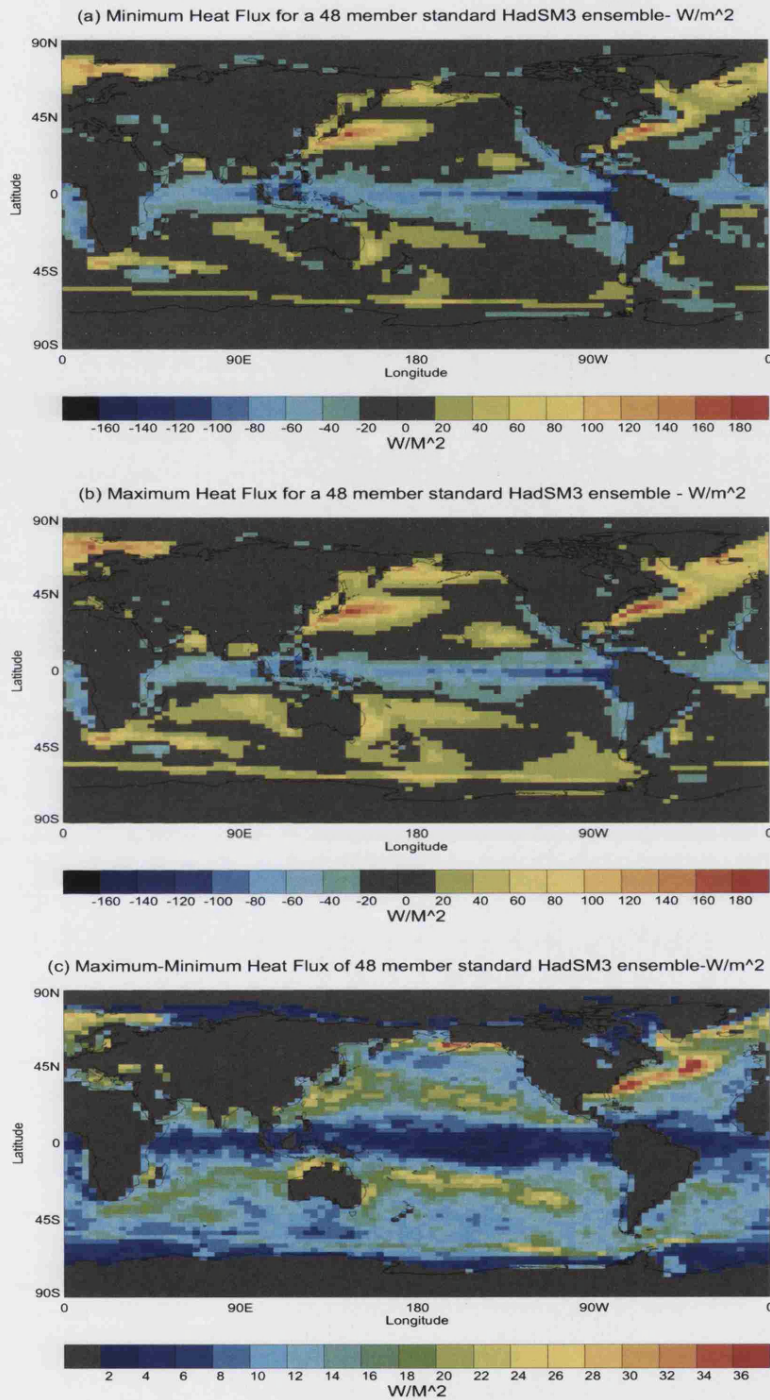


Figure 5.4: Panels show (a) minimum, (b) maximum and (c) range of HFA for the Standard ICE. Panel (c) shows that there are regions for which members of the Standard ICE require HFAs differing by less than $4W/m^2$ but other areas where the differences can be as large as $40W/m^2$

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:

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

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 $n \times 7008$ 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.

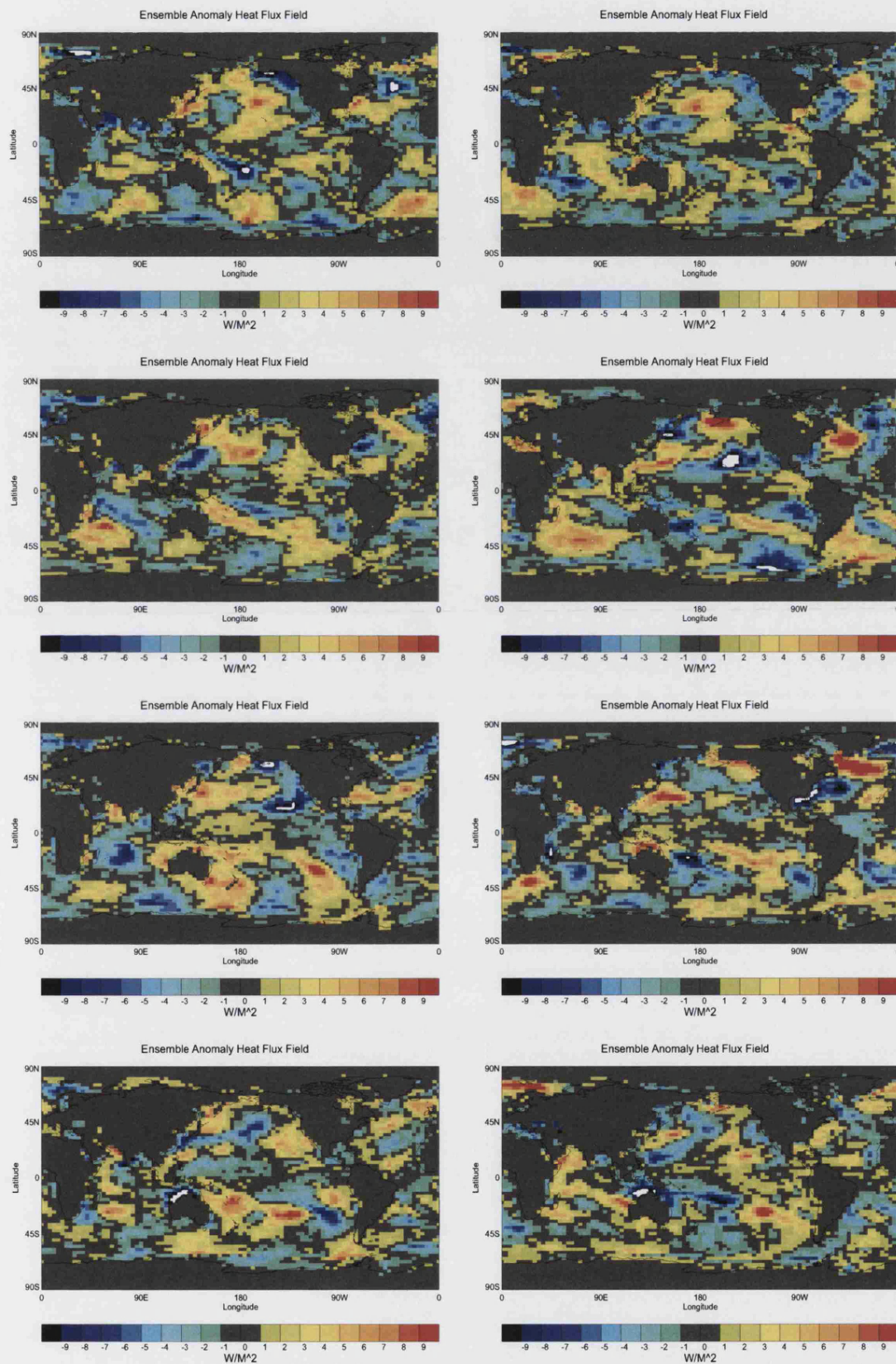


Figure 5.5: The HFA for 8 simulations randomly selected from the Standard ICE. The ensemble mean is subtracted from each simulation, giving an anomaly field.

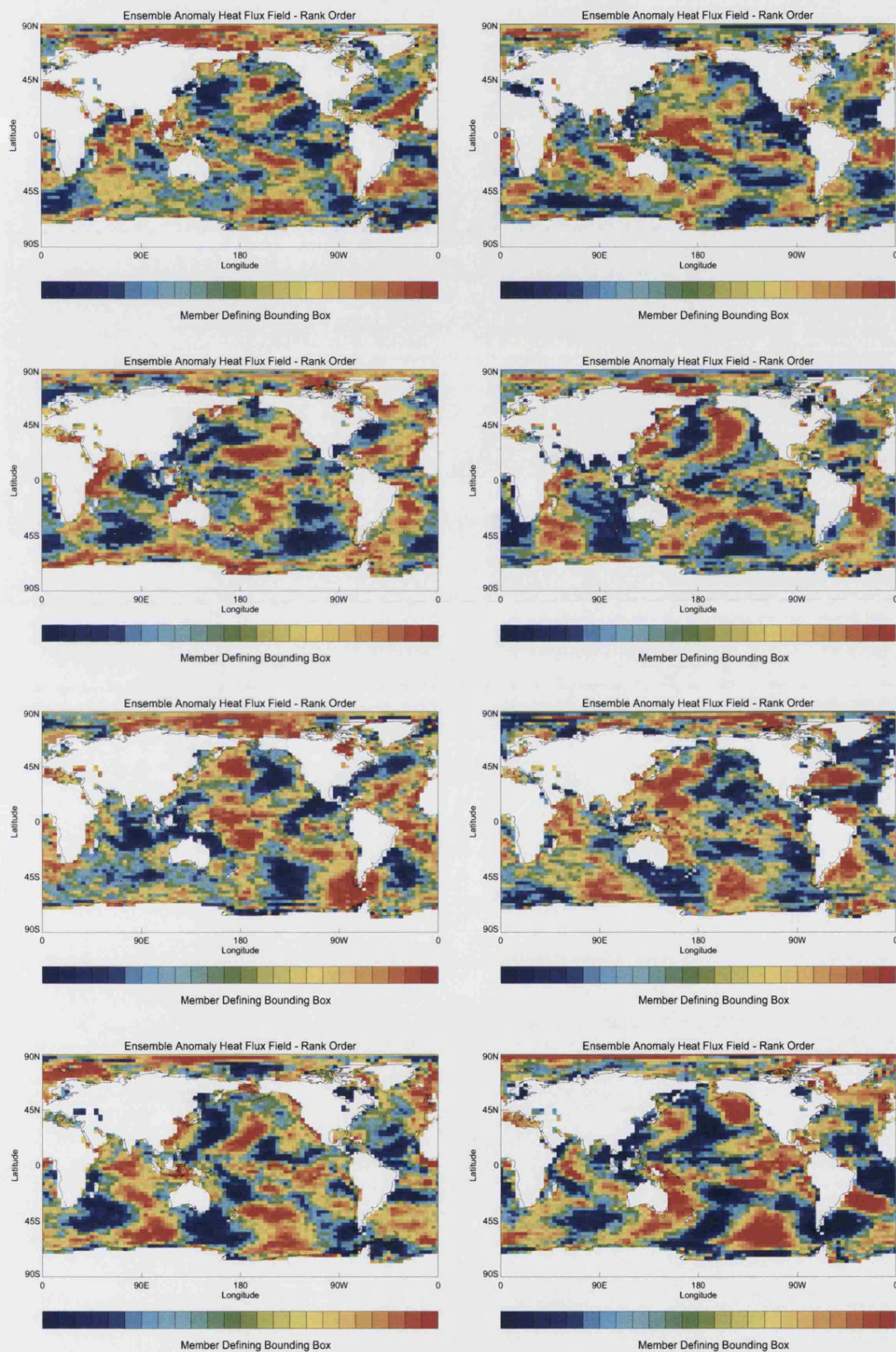


Figure 5.6: The HFA for 8 simulations randomly selected from the Standard ICE. The ensemble mean is subtracted from each simulation, giving an anomaly field, expressed in rank order.

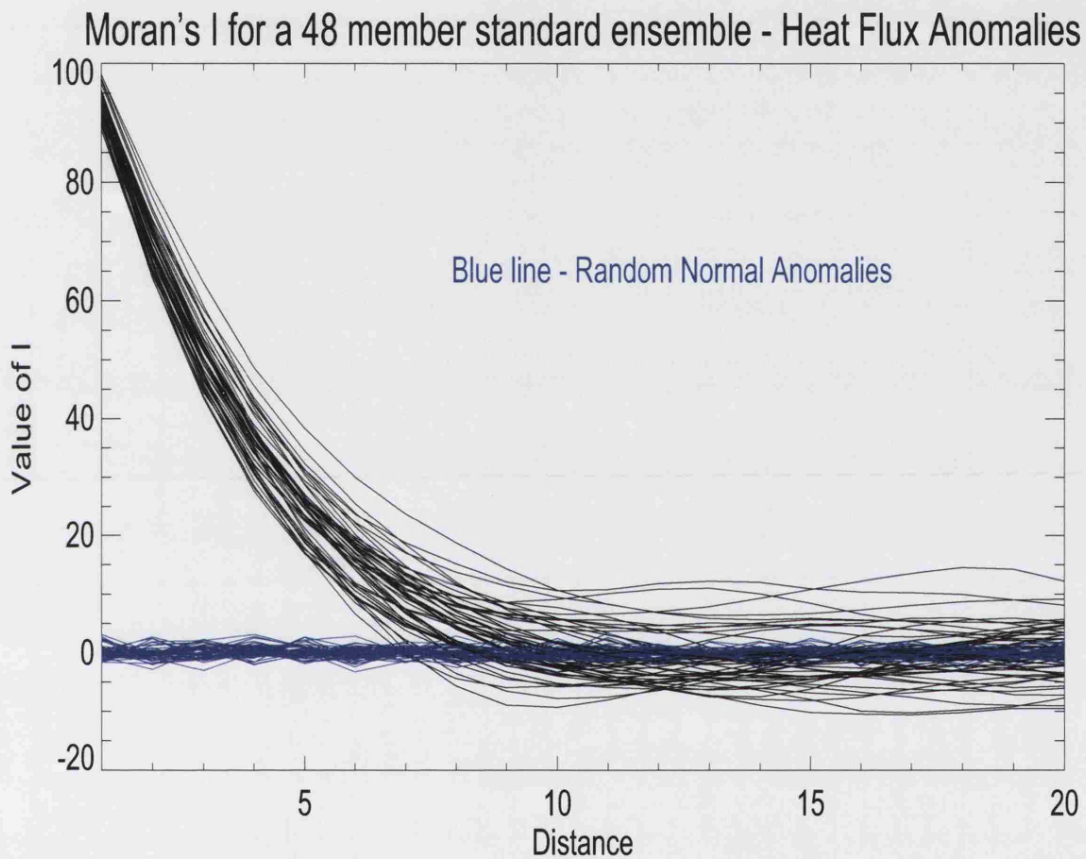


Figure 5.7: Moran's I statistic is plotted for the Standard ICE against distance in white. In blue, the statistic is calculated on a set of randomly generated data for comparison. Positive values of I indicate a positive correlation. For distances less than 7, grid-boxes show a positive correlation.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

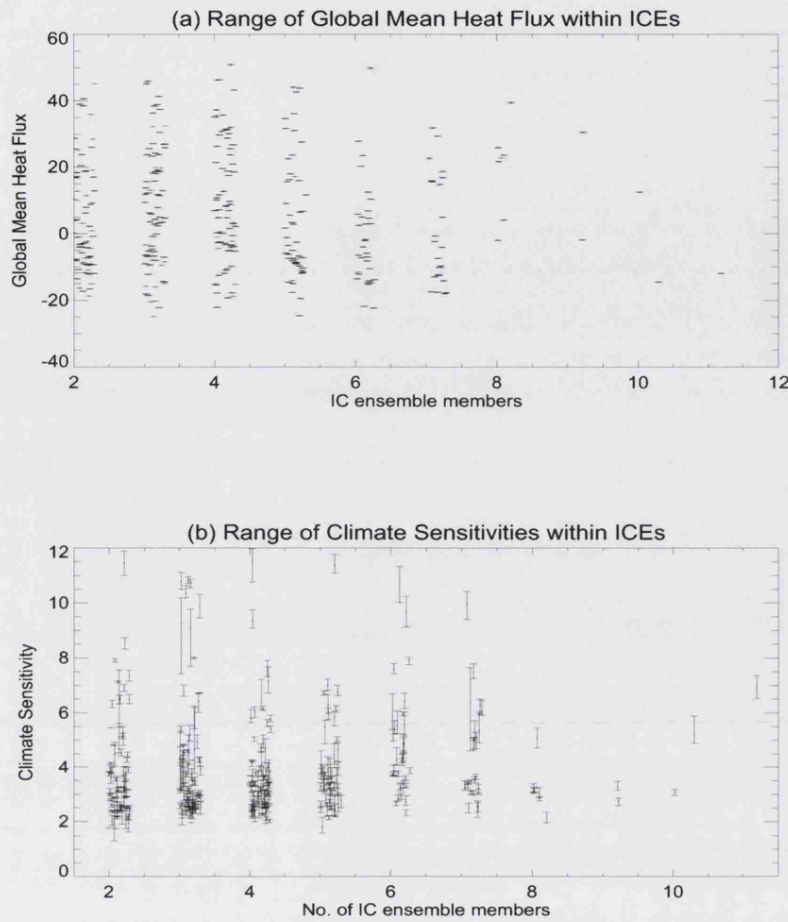


Figure 5.8: These graphs show the range of values for global mean HFA (top) and CS (bottom) as a function of ICE size. The range of values for global mean HFA is often so small that the minimum and maximum bars are almost indistinguishable. The number of ensemble members is “jittered” by adding a small 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 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 by 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.