

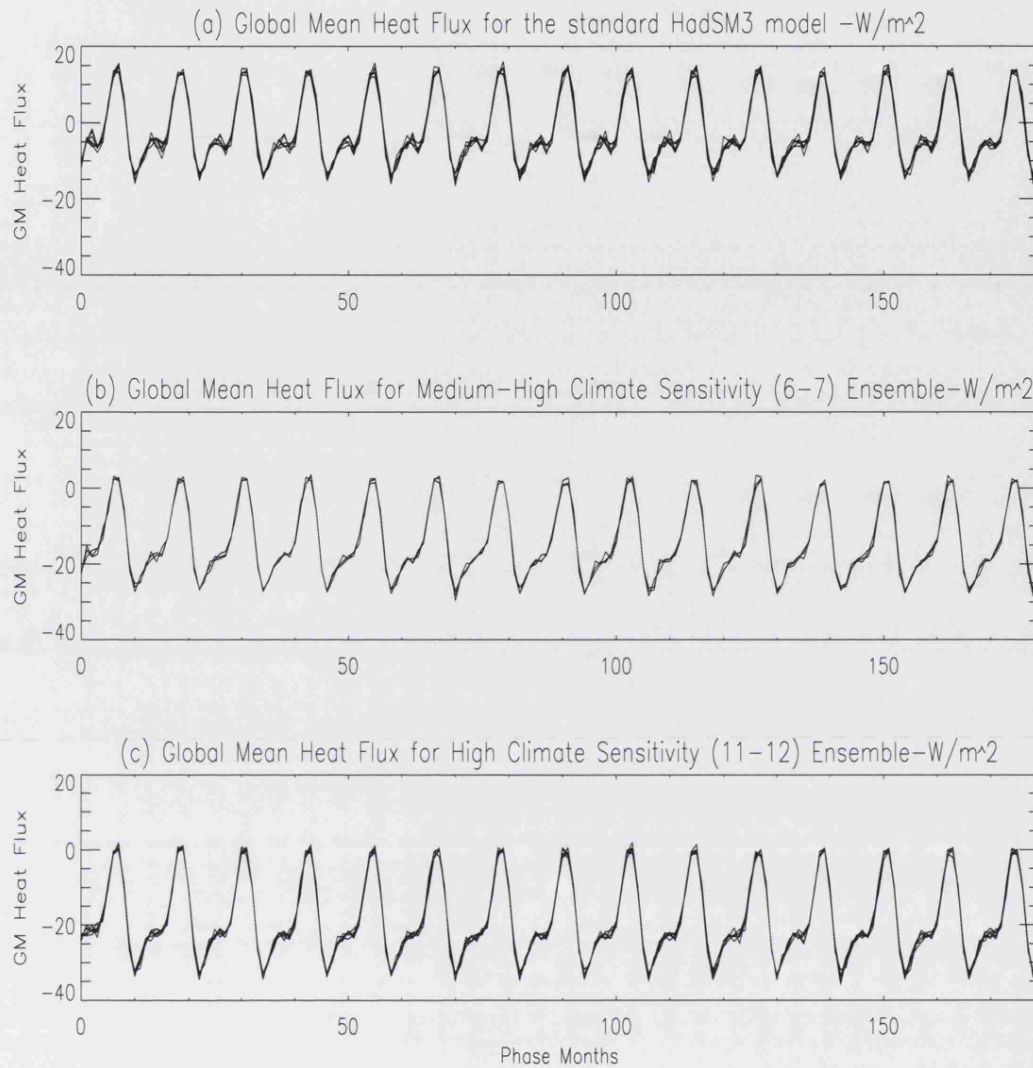


Figure 5.9: These graphs show the global mean HFA for the calibration phase. Time runs in months throughout the phase. Panel (a) shows the control ensemble (of 6 simulations, with an average CS of 3.4 degrees Celsius), panel (b) a randomly selected ensemble whose CS is 6.4 degrees Celsius (3 simulations)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

5.3 Seasonality in the HFA

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

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

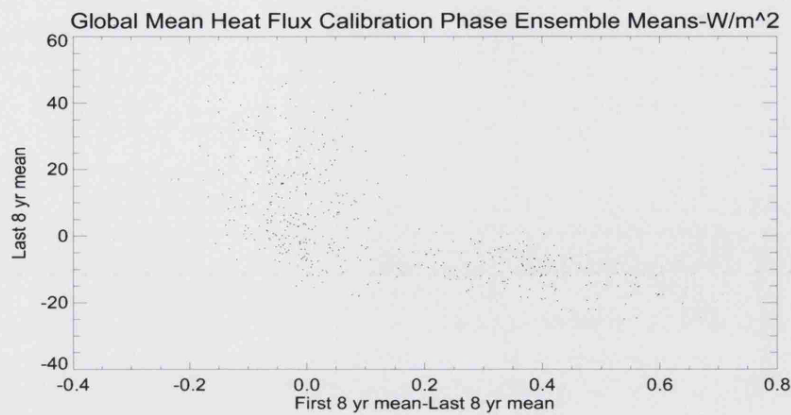


Figure 5.10: The y-axis shows the final 8 year mean global mean HFA for each of 484 model versions. The x-axis shows the *first* 8 year mean minus the last 8 year mean. These values are fairly close, but with a tendency for simulations with negative values of global mean HFA to remove more heat during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

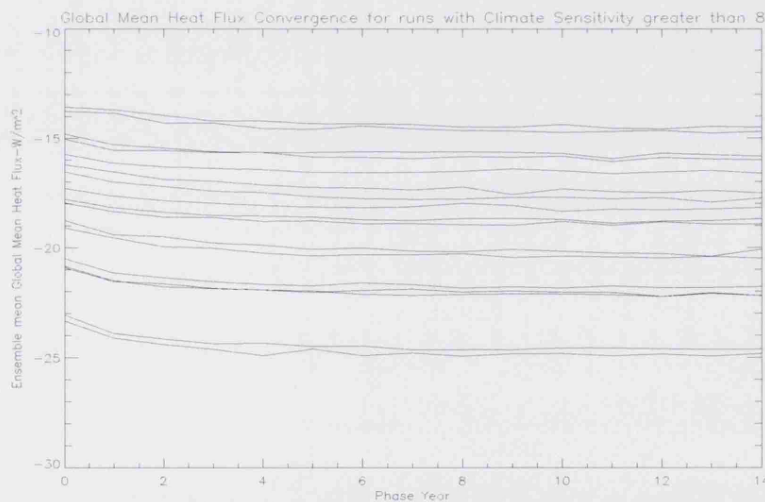


Figure 5.11: The global mean HFA is shown here for model versions with an average CS of 8 degrees or higher. There is an initial drop in the global mean HFA field, followed by a stabilisation.

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

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

systematic regional biases.

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

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

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

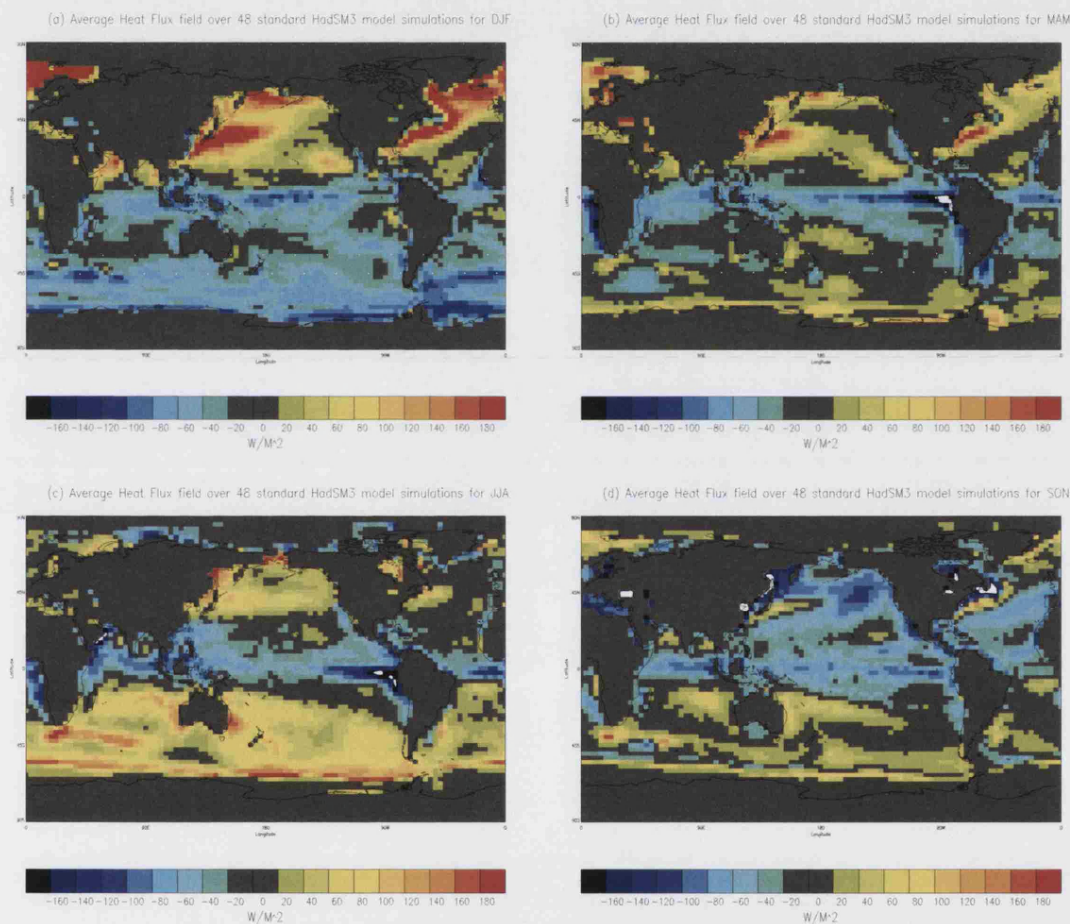


Figure 5.12: The HFA field for (a) DJF, (b) MAM, (c) JJA and (d) SON averaged over the Standard ICE.

5.4 HFA and Climate Sensitivity

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

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

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

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

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

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

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

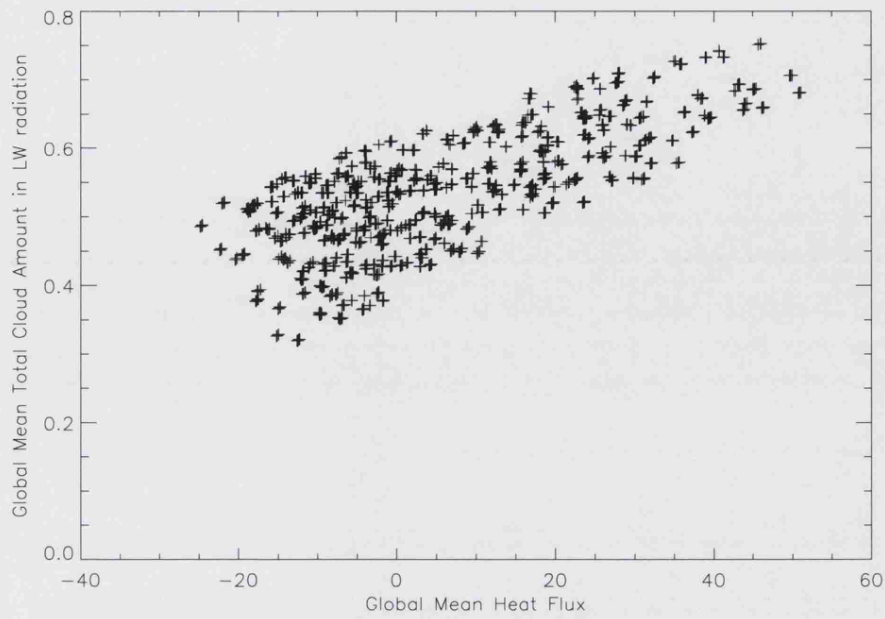


Figure 5.13: Total global cloud cover (as a fraction) is plotted against the global mean HFA for $PPE_{quality}$. There is a pattern for simulations with a low total cloud amount to have negative global mean HFA.

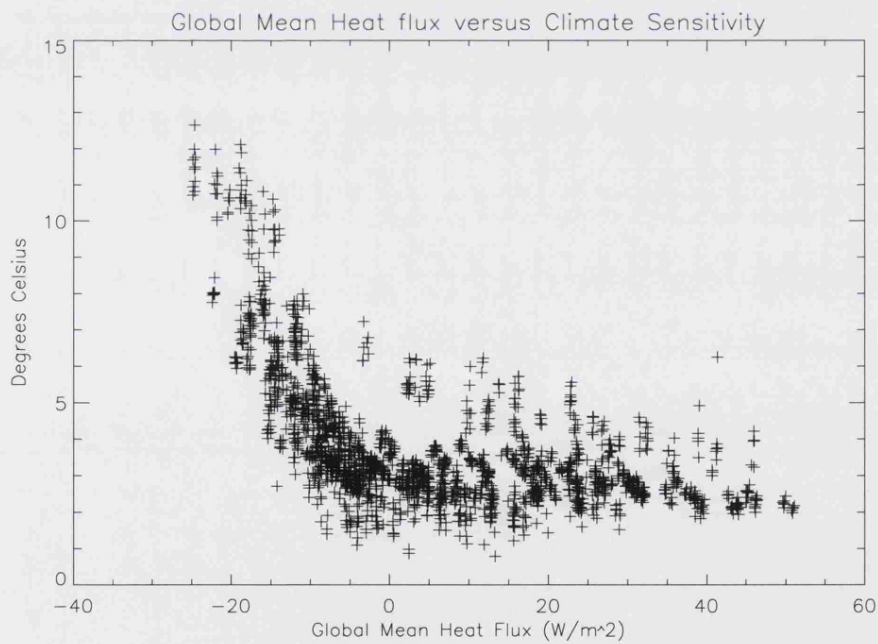


Figure 5.14: The global mean HFA is plotted against CS for $PPE_{quality}$. There is a distinct tendency for simulations with a large negative global mean HFA to produce simulations with very high CS.

5.5 HFA and drift

Despite the use of HFA in the CPDN experiment, a proportion of simulations ($\sim 30\%$) of PPE_{2578} display a strong negative temperature drift¹ during the control phase. These simulations are disregarded from the analysis of the effect of external forcings on model climate since the simulation exhibits significant unphysical climatic changes in the absence of external forcings. A smaller proportion of simulations show a positive temperature drift, but generally of a much lower magnitude than simulations with a negative drift (in PPE_{2578} , of 319 simulations with a significant positive drift, 3 have a positive drift greater than 1 degree Celsius per decade, whereas of 766 simulations with a significant negative drift, 265 have a negative drift of more than 1 degree Celsius per decade). An important negative feedback mechanism, first identified in Stainforth *et al.* (2005) and discussed in this Section, is a key component of negative drift. No such mechanism has been detected that results in unphysical positive feedbacks. Negative drift could arise from a mis-calibration of the HFA field or from some other source. This Section presents evidence to suggest that the HFA field (or its mis-calibration) is unlikely to be the sole cause of such model drift.

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

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

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

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

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

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

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

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

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

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

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

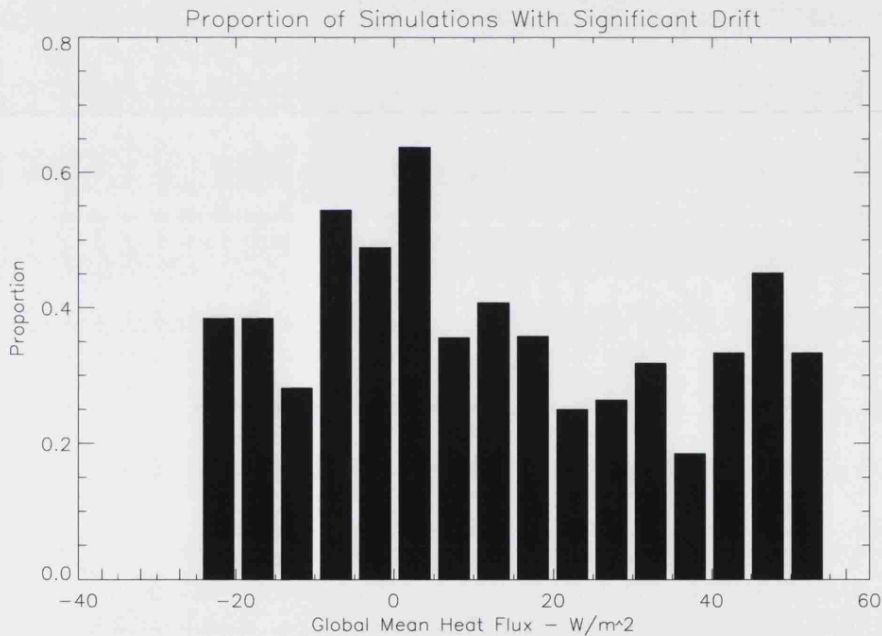


Figure 5.15: The proportion of simulations from PPE_{2578} with significant GMST drift is plotted for categories of global mean HFA of width $2\text{W}/\text{m}^2$. There is no clear tendency for simulations with a significant negative global mean HFA to have a significant GMST drift.

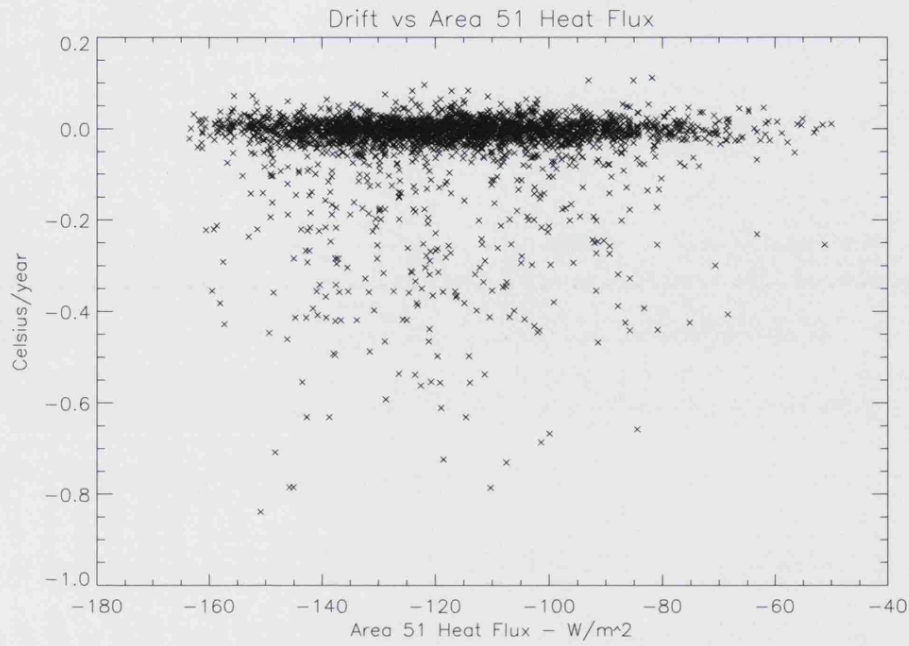


Figure 5.16: The Area 51, JJA, HFA is plotted against the control phase drift for PPE_{2578} . There is no clear pattern for simulations with a strong reduction of heat over Area 51 to have a strong negative drift.

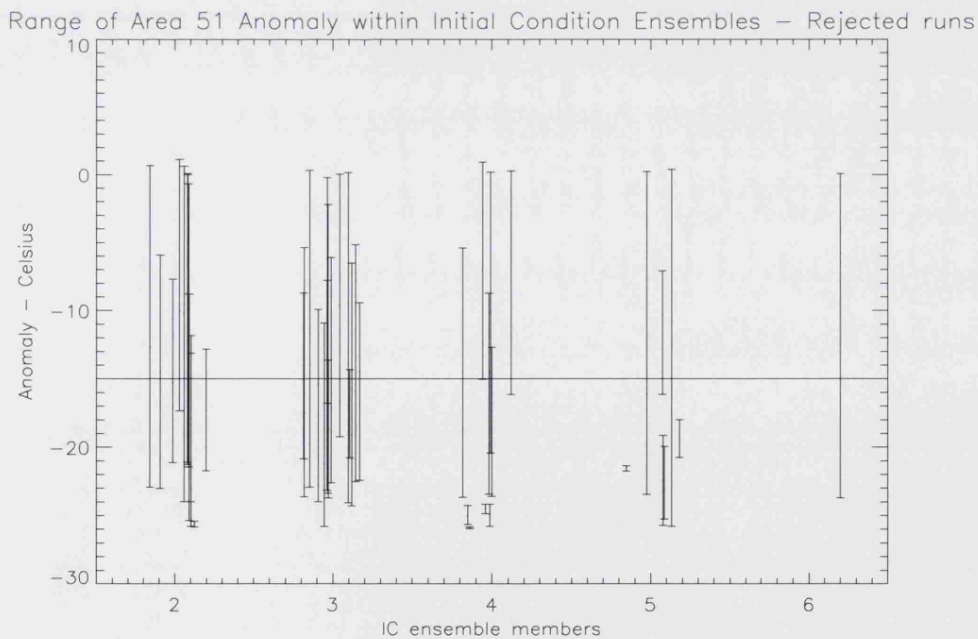


Figure 5.17: The range of anomalies within an ICE is plotted over the problematic grid-box. Of 484 ensembles, only those with at least one unacceptable simulation (an Area 51 statistic less than -15 degrees) are plotted (47 ensembles).

5.6 Conclusion

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

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

Original results presented in this Chapter are:

- Perturbation of Initial Conditions has little effect on the global mean HFA (the greatest difference between IC members across 418 model versions is $0.419W/m^2$). On the other hand, 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.
- It has been shown by analysing the leading Singular Vector in the HFA fields of ICEs that parameter perturbed model versions of HadSM3 can require significantly different global HFAs. The leading Singular Vector has been shown to explain significantly more variability in the HFA fields where model simulations share parameter values than where simulations are drawn at random. It was also also shown that whilst perturbing Initial Conditions has an impact

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

- Significant seasonal variations exist in the HFA both globally and regionally. This effect is likely mimicking the seasonally-dampening effect of a deep ocean and means HadSM3's seasonal cycle might not respond to rising CO_2 in a physical way.
- There is a relationship between global mean HFA and CS. Simulations with high values of CS (over 8 degrees) tend to have strong negative global mean HFA (less than $-10W/m^2$). This relationship is potentially important for interpreting simulations with very high values of CS (greater than 8 degrees Celsius).
- Relationships between HFA and model drift have been investigated with the use of global mean and refined local statistics. No discernible pattern between global mean HFA and model temperature drift was found. The same model version can produce simulations that either drift or do not drift, suggesting drift is not dependent solely on parameter perturbation.

Chapter 6

ICEs and the Internal Variability of Climate Models

6.1 Overview

Climate is a distribution¹, consisting of a range of possible outcomes. A single climate model can exhibit a range of different states; this range of model behaviour is the internal variability of the model. Climate modelling is thus concerned with a long-term distribution and determining whether or not a range of forcings will result in significant changes in this distribution. This Chapter explains how the distribution of model climate can be evaluated using ICEs.

In a dynamical systems context, this internal variability can be understood in terms of the climate model's "attractor", the set of points to which a model will converge on long time scales. In a dynamical systems' context, internal variability arises from not knowing on which point of the model's attractor the system will lie at any given time point; there are a variety of possible states, depending on the precise specification of the IC². The notion of simultaneously possible states is a multi-world

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

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

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

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

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

response.

1. ICEs are used here to assess variability in the HadSM3 model's response to a doubling of CO_2 concentrations for different regions and variables. An ICE is necessary to understand how the distribution of model climates change in response to some external forcing. This method can be similarly applied to the comparison of different emissions scenarios or models that differ by parameter values or structure.
2. ICEs can be used to test for consistency of information in the model output, as laid out in Chapter 2. Such tests can provide a check for robustness of information in climate predictions through an evaluation of the magnitude of the model's internal variability on various length scales and in different variables.

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

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

6.2 Introduction to ICEs

Uncertainty in climate projections can be ascribed to a number of different sources Stainforth *et al.* (2007a) (also see “Uncertainties and Ensembles” in Chapter 2). These uncertainties vary in magnitude with length scale and the variable of interest. One of these types of uncertainty, ICU, is considered here. ICU is an irreducible uncertainty and is an expression of the model’s internal variability.

The distribution of model climate needs to be understood in order to make a judgement on whether the model’s response lies outside the range of internal variability. There are at least 2 different ways to estimate the internal variability of GCMs:

1. Internal variability can be assessed using a very long simulation under the same scenario Min *et al.* (2005); this approach is only valid in *equilibrium experiments* and not *transient experiments*.
2. ICEs can be used to assess internal variability in either equilibrium or transient settings. A single, long, simulation can be useful for detecting drift in the model but can not be used in transient experiments. In contrast ICEs can provide a distribution of climate over time.

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

A 64 member ICE is studied in this Chapter in order to understand the standard HaDSM3 model better and to provide additional understanding of the model’s internal variability.

Since members of an ICE are samples from the same dynamical system (the model attractor, in the case of equilibrium experiments) ICEs might be used to make probabilistic statements about the model attractor. These probabilistic statements are reliant on our ability to assume that members of an ICE are, in effect, independent draws from the same distribution. These assumptions are not tested here, nor might they be easy to test rigorously since this would require further assumptions regarding the nature of the model and its attractor. It is important to note that whilst the terminology of dynamical systems is often used (e.g. climate models are often referred to as “chaotic” Liu *et al.* (2008)), dynamical systems theory is of limited use in the case of climate models; no computer model is truly aperiodic nor would a 10^8 dimensional model attractor be easily amenable to analysis. A pragmatic approach is therefore adopted and terminology used for the sake of convenience. Other types of ensemble, such as multi-model ensembles, draw from diverse attractors, and thus the statistics produced are critically dependent on the sampling strategy and choice of models used Frame *et al.* (2005). ICEs explore variability *within a particular model*, whereas perturbing parameters or changing model structures explores uncertainty *across models*. Within a single model, it is possible to make probabilistic statements assuming independent and identically distributed samples. In contrast, the metric of model space (or even parameter space) is not well-defined in multi-model ensembles Allen & Stainforth (2002) and thus is not readily amenable to objective *probabilistic* analysis.

6.3 The internal variability of HadSM3

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

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

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

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

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

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

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

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

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

There is a significant amount of variability due to IC perturbation alone, especially for regional¹ simulations of precipitation. This suggests that the internal variability of climate models can be large on regional length scales.

6.4 ICEs and Robust Model Response

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

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

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

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

6.4.1 Regional response in the 64 member HadSM3 ensemble

An important use of ICEs is to distinguish between models or forcing scenarios. In this Section, a 64 member ICE is used to examine whether a significant response can be detected in model behaviour to a doubling of CO_2 concentrations. A single model run under each of these conditions does not allow the two different scenarios to be statistically distinguished since we can not reliably estimate the model's internal variability. With a few simulations under each scenario, it is only possible to make a limited assessment regarding the internal variability of the model under each level of CO_2 and judge whether the two distributions are significantly different. On the other hand, large ensembles enable robust analysis of model response; it is highly likely that two sets of 64 simulations would overlap in distribution if there were no difference in model behaviour between the pre-industrial and doubled CO_2 scenarios. Distributions of model response are compared in this Section in terms of the magnitude of overlap observed; this overlap represents a measure of the distinguishability of these distributions¹. In the case of the CPDN experiment, the degree of overlap between the pre-industrial and doubled CO_2 phase distributions of the standard HadSM3 model ICE is a quantification of the consistency of model response to doubled CO_2 . Thus, when no overlap between model distributions is seen

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

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

The probability of observing no overlap between two distributions of size m and n can be calculated non-parametrically as follows: under the assumption of independent and identical draws from the same distribution, an ensemble of size n is expected to cover $100 \cdot (\frac{n-1}{n+1})\%$ of the probability mass¹ Weisheimer *et al.* (2004). Take the ensemble of size n to be the base ICE (e.g. pre-industrial CO_2) to which the distribution of size m is to be compared. The probability of a simulation, drawn from the same distribution as the base ICE, not falling within the range already covered is $\frac{2}{n+1}$ – a probability of $\frac{1}{n+1}$ of falling below all n simulations and a probability of $\frac{1}{n+1}$ of falling above. The probability that m simulations are drawn from the same distribution as the base ICE and all fall outside the range covered would be $2 \cdot (\frac{1}{n+1})^m$, which becomes extremely small for large values of n and m . Where n and m are greater than 3 this probability is less than 5% and in the case of the 64 member standard HadSM3 ICE ($n=m=64$) the probability of observing no overlap, under the assumption that both sets of simulations are drawn from the same distribution, is less than 10^{-117} . Where distributions do overlap, which is almost certain to occur if the 64 member ICEs are drawing from a common distribution the model produces a result that is within the range of internal variability. Clearly, these probabilities only represent whether there is a statistical difference between distributions and not whether the difference is meteorologically significant. In the case of overlap, the level of distinguishability between the distributions can be estimated as follows.

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

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

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

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

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

in the distribution of simulated precipitation changes across the ICE.

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

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

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

6.4.2 Comparison of two Perturbed Physics ICEs

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

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

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

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

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

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

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

6.5 ICEs in Transient Experiments

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

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

Thus ICEs play a different role in transient experiments; providing insight into how the model’s dynamics change under a particular forcing. The model’s internal variability may not be directly comparable to the variability in the observed climate¹. An ICE can help understand the internal variability of the model under transient forcings.

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

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

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

6.6 Conclusion

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

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

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

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

New results presented in this Chapter are:

- The availability of a large ICE allows for a quantification of the HadSM3's internal variability, which has been shown to be significant on length scales relevant for impact studies and adaptation decisions. The various roles of ICEs are discussed and their increased use encouraged. These results are significant since it has typically been assumed that the effect of perturbing Initial Conditions on climate simulations was negligible Tebaldi & Knutti (2007).
- It has been shown for the first time that Initial Condition perturbation can have a significant effect on model behaviour on relevant length and time scales in temperature and precipitation. The sign of change for 8 year mean seasonal precipitation is unanimous in only $\sim 3\%$ of grid boxes. In temperature, 8 year mean seasonal differences is shown to be as large as 10 degrees Celsius in some grid boxes. Such large differences are not usually considered possible to have arisen from Initial Condition perturbation alone and these results could affect the experimental designs and the interpretation of model variability.

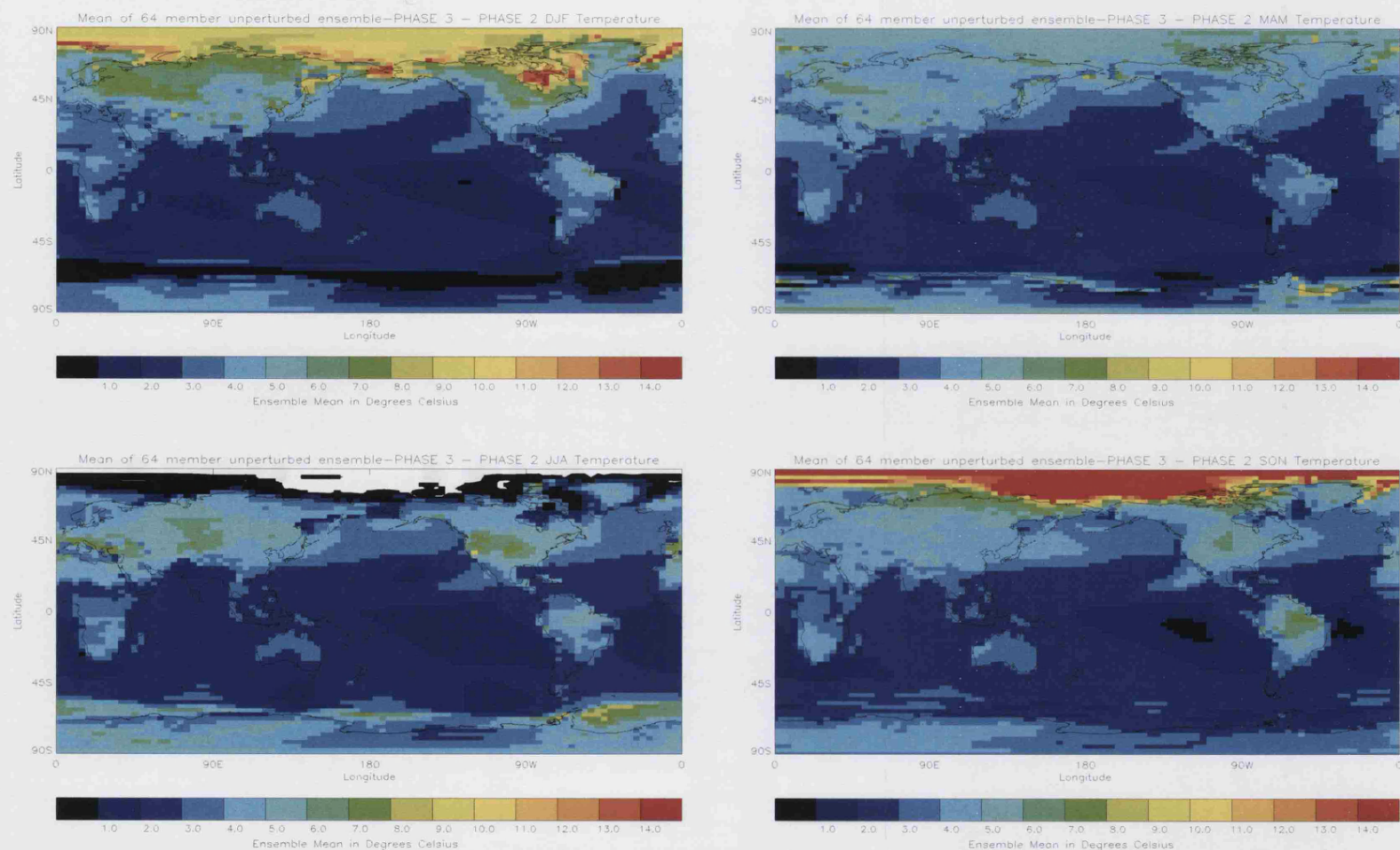


Figure 6.1: The standard HadSM3 model 64 member ICE mean for 8 year temperature change is shown for each season. Black areas show little or no cooling, white areas a cooling. Red areas show very high warming of over 9.5 degrees Celsius. Warming is strongly non-uniform and varies significantly with season.

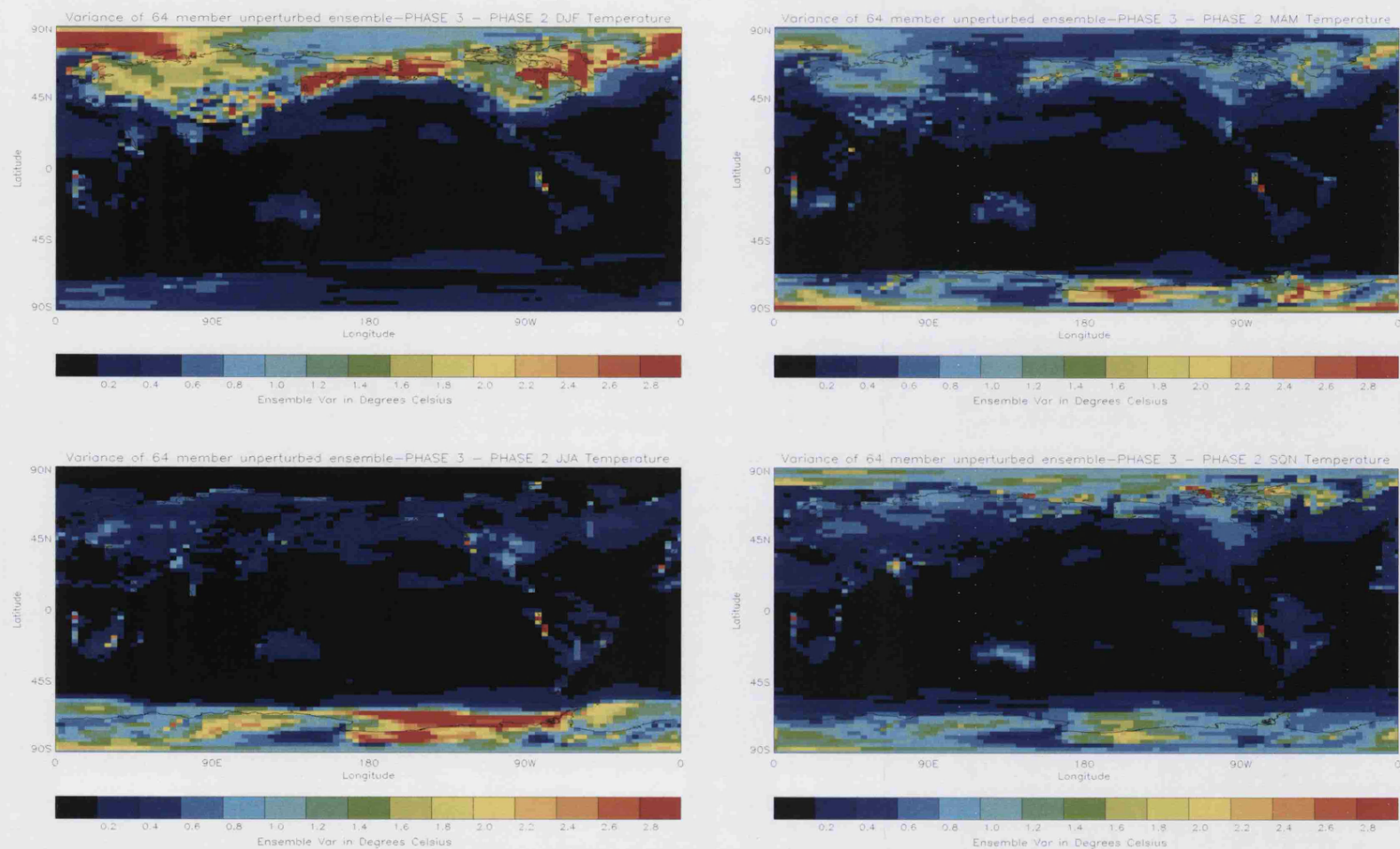


Figure 6.2: The variance in seasonal 8 year mean temperature change fields over the 64 member standard HadSM3 model ICE. Black areas show a variance of less than 0.2 degrees Celsius. Variance is typically higher over land – over 2 degrees Celsius in some cases.

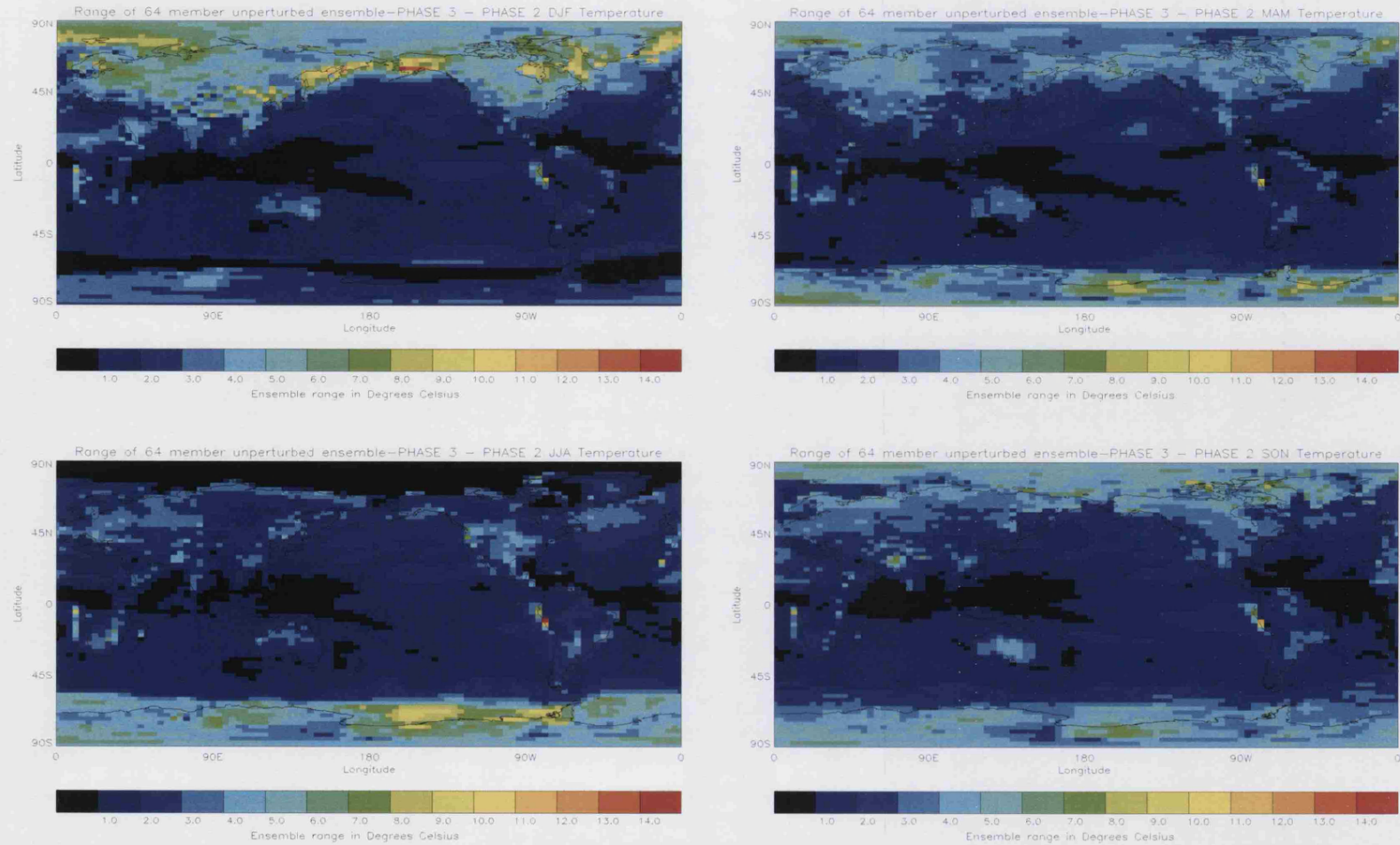


Figure 6.3: The range (maximum - minimum values) for temperature change over the 64 member ensemble. Areas in black indicate a spread of less than 1 degree Celsius over the whole ensemble. The range of seasonal temperature change within this ICE is over 10 degrees Celsius in some cases.

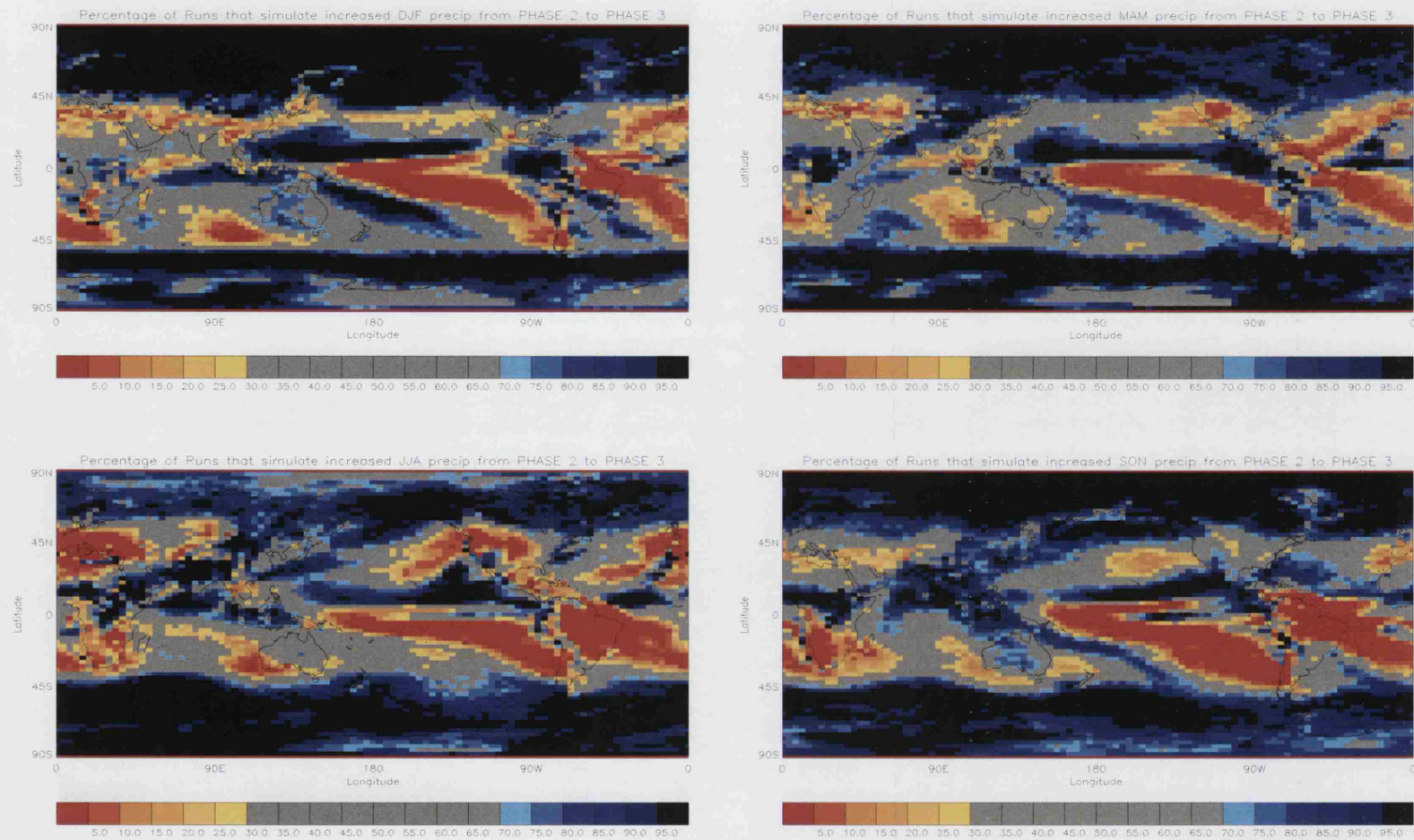


Figure 6.4: A democracy plot of precipitation change. The percentage of simulations (over the standard HadSM3 model ICE for which the 8 year seasonal precipitation increases from control to doubled CO_2 . Areas in black (red) indicate that more than 95% of simulations show an increase (decrease) in precipitation.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

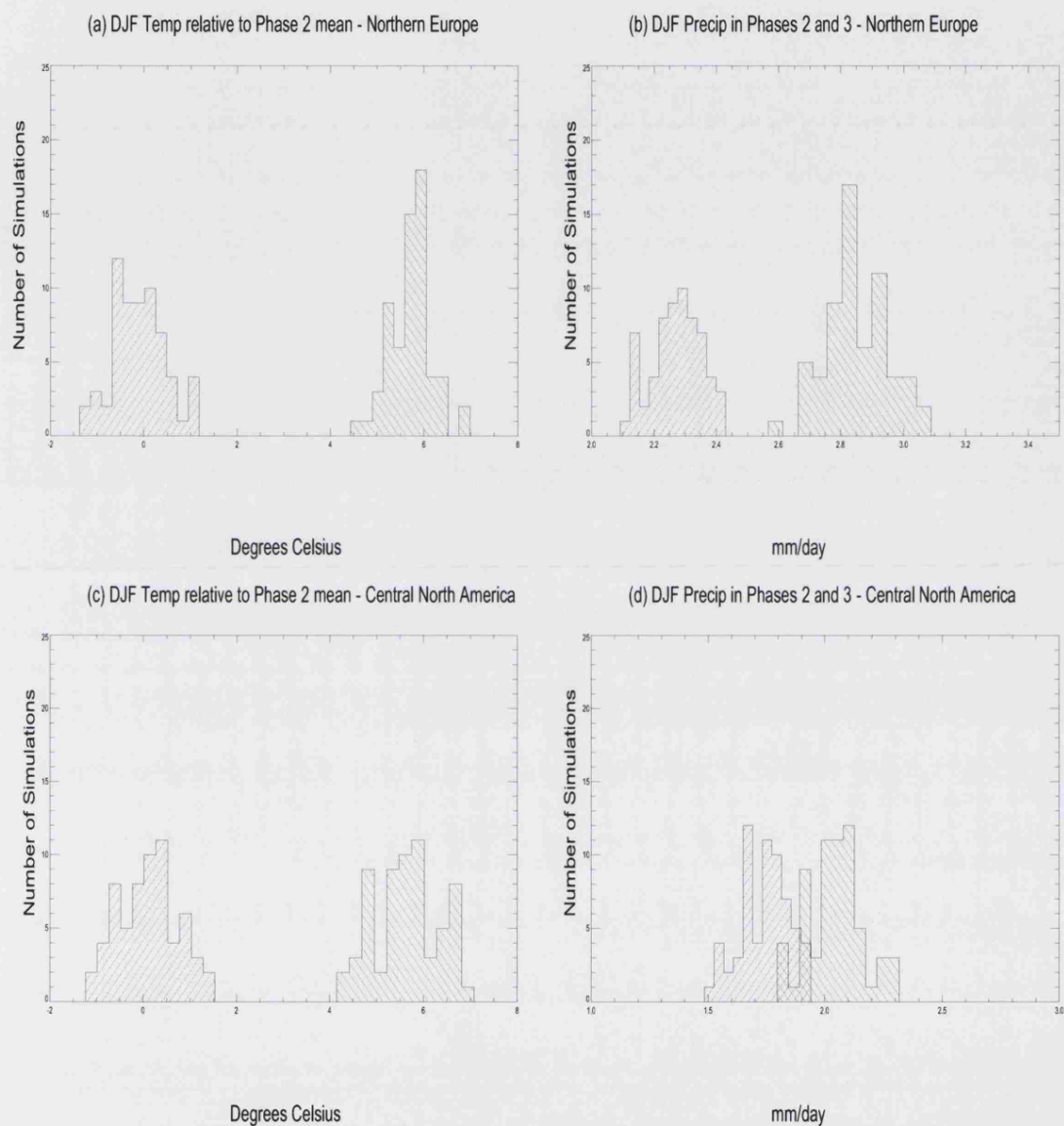


Figure 6.5: The distribution of 8 year means for the 64 members standard HadSM3 model ICE for (a) Northern Europe Temperature, (b) Northern Europe Precipitation, (c) Central North American Temperature and (d) Central North American Precipitation. The control phase is shown in green and the doubled CO_2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

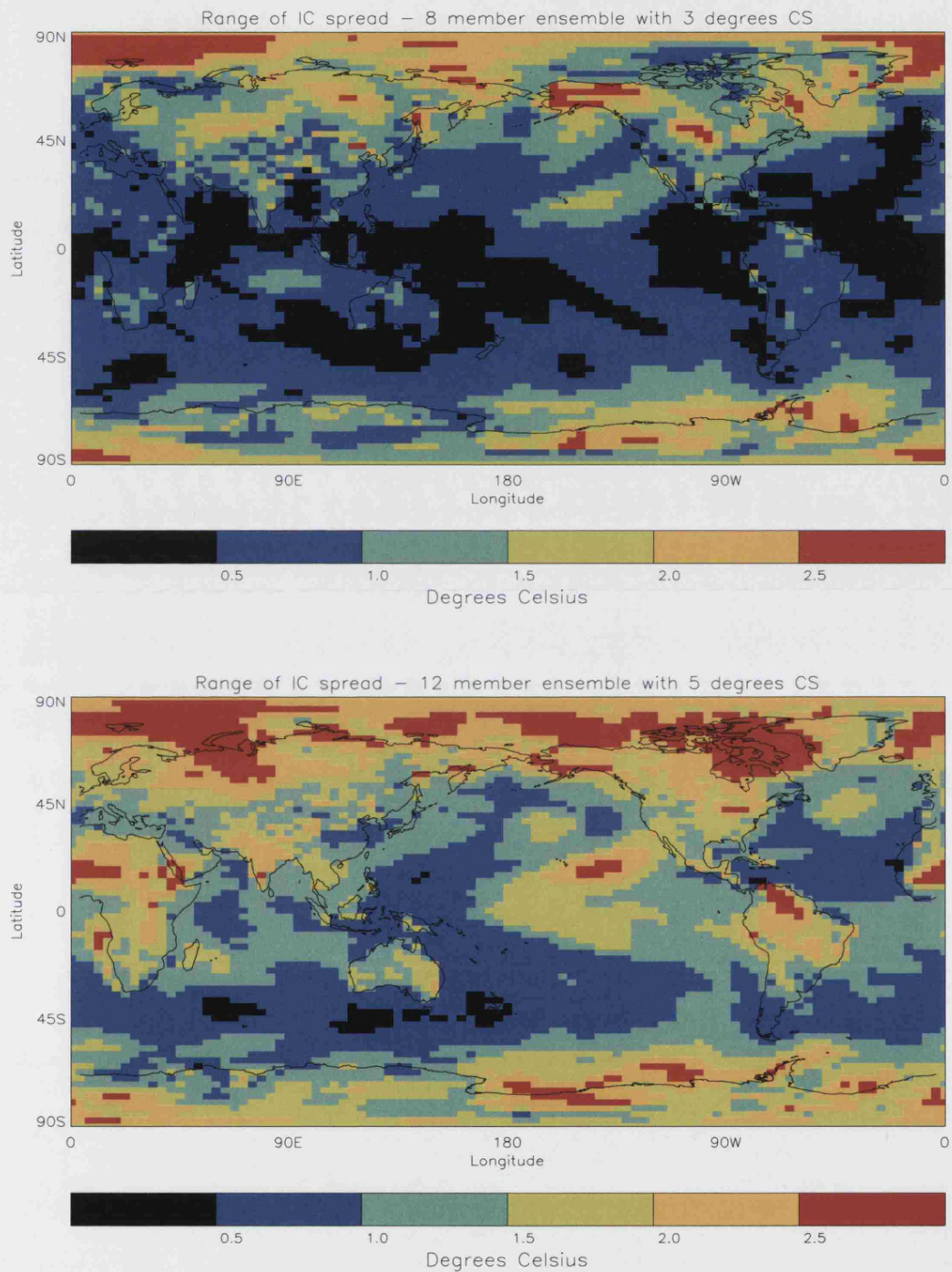


Figure 6.6: Range of 8 year mean temperature change under a doubling of CO_2 for 2 ICEs of 8 and 12 members and 3 and 5 degrees CS respectively. The magnitude of this internal variability is typically one degree Celsius, but can be over 2.5 degrees Celsius, particularly for the larger, 5 degree CS, ICE.

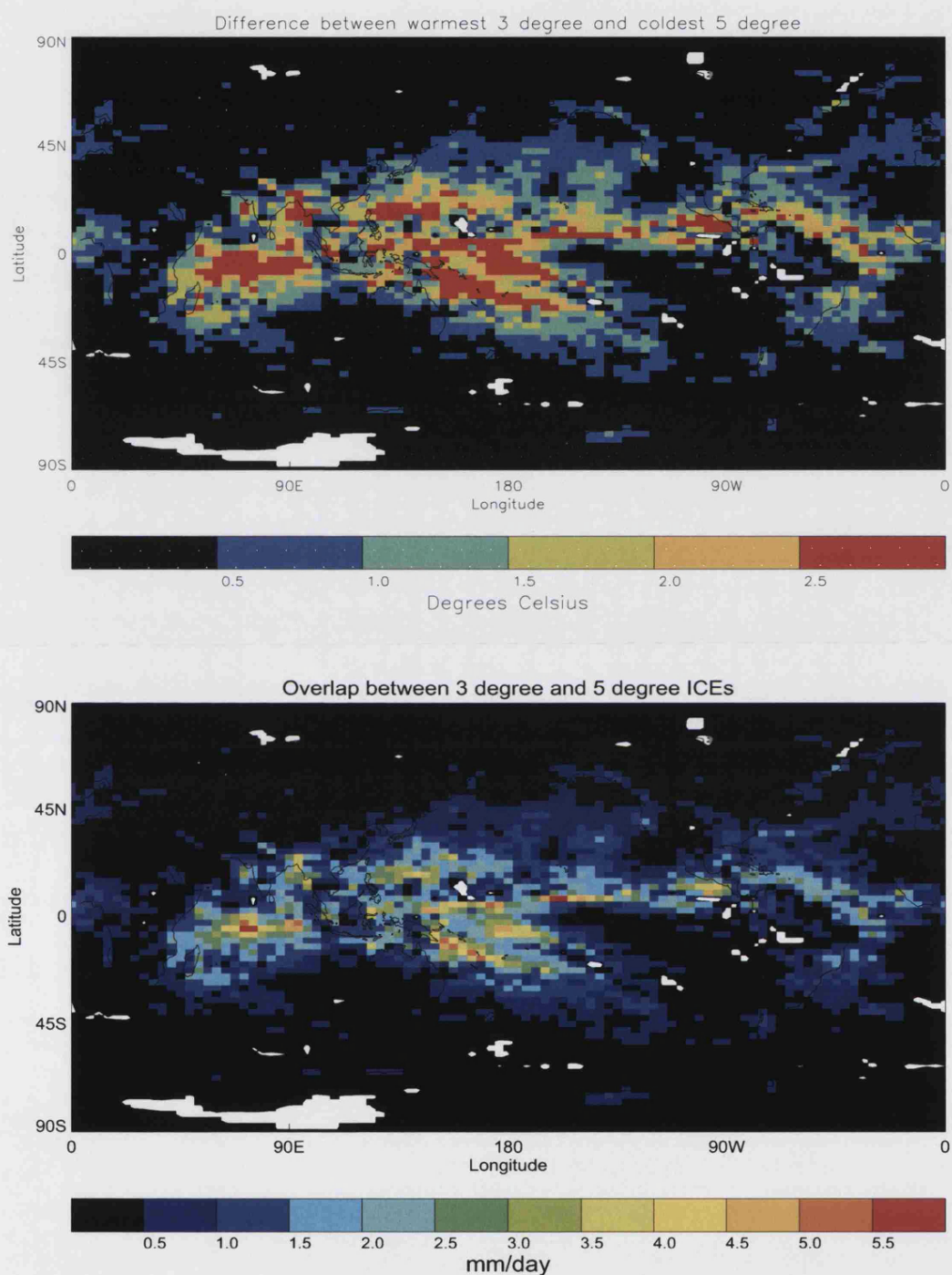


Figure 6.7: The difference in 8 year mean temperature/precipitation change under a doubling of CO_2 between the maximum of an 8 member ICE with 3 degrees CS and the minimum of a 12 members ICE with 5 degree CS. The extent of this overlap is shown in temperature and precipitation. Positive values in temperature 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 form 3 to 5 degrees. Negative values denote areas with no overlap.

Chapter 7

Constraining New Results from the CPDN grand ensemble

7.1 Introduction

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

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

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

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

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

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

the magnitude of uncertainties present in a grand ensemble of CPDN simulations. Constraining is understood here as attempting to reduce uncertainties by eliminating simulations¹. Mathematically, simulations receive an effective weight of either zero or one. Whilst fractional weights have been attempted Murphy *et al.* (2004) the simpler, yet still challenging, approach of assigning weights of zero or one is used in this Chapter. The difficulties in constraining the range of results from climate ensembles are highlighted in this Chapter in the context of their physical rationale and statistical good practice.

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

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

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

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

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

- *Heat flux adjustment.* The HadSM3 model used in this experiment uses a Heat Flux Adjustment (HFA) (details have been presented in Chapter 5). It might be suggested that a global mean value of this HFA close to zero would be a desirable property of the model – the global mean HFA is sometimes forced to equal 0 in flux-adjusted experiments Min *et al.* (2005). In particular, if the global mean HFA is close to zero this would suggest the atmosphere and ocean components of the model are in close equilibrium during the calibration phase. Furthermore, Chapter 5 established a relationship exists between HFA and CS; this might be used as the basis of a post-hoc constraint on CS. Such post-hoc filters are not recommended in this Thesis, rather this method is applied to make two points: **a)** It is statistical bad practice to choose a method for constraining CS *a posteriori* and **b)** The effect of ruling out simulations with a significantly non-zero global mean HFA is shown to have limited effect on the range of CS. Even when a stringent HFA filter is applied, simulations with high values of CS (over 8 degrees Celsius) are still admitted.
- *In-sample fit.* An important requirement for models to provide useful predictions is their ability to match the past, as discussed in Chapters 2 and 3. One approach to constraining model simulations could then be to disregard simulations that do not capture past observations within a certain level of accuracy. There are a number of different ways to do assess such in-sample fit. A simple criteria is set in this Chapter, using the in-sample RMSE error, as in Reichler & Kim (2008); Stainforth *et al.* (2005), compared to the observational fields for 1961–1990. This RMSE is used to constrain model

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

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

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

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

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

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

7.2 The Data Set

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

7.3 Climate Sensitivity

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

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

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

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

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

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

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

1. The peak is around 3.5 degrees for all three distributions of CS, close to the value of CS for the standard HadSM3 model (about 3.4 degrees Celsius, as stated in Chapter 6) and consistent with estimates from other state-of-the-art models Solomon *et al.* (2007a). This suggests that the effect of simultaneous parameter perturbations often cancels out (or the effect of perturbing some parameters is small) for GMST change (this need not be the case for other variables and length scales), often resulting in values of CS close to the standard model version.
2. The distribution is fairly smooth, with no obvious discontinuities or multimodality. There is a wide range of values, especially at the high end. Rather than a few isolated high values, there is a smooth tail, with a monotonically decreasing number of very high CS simulations.

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

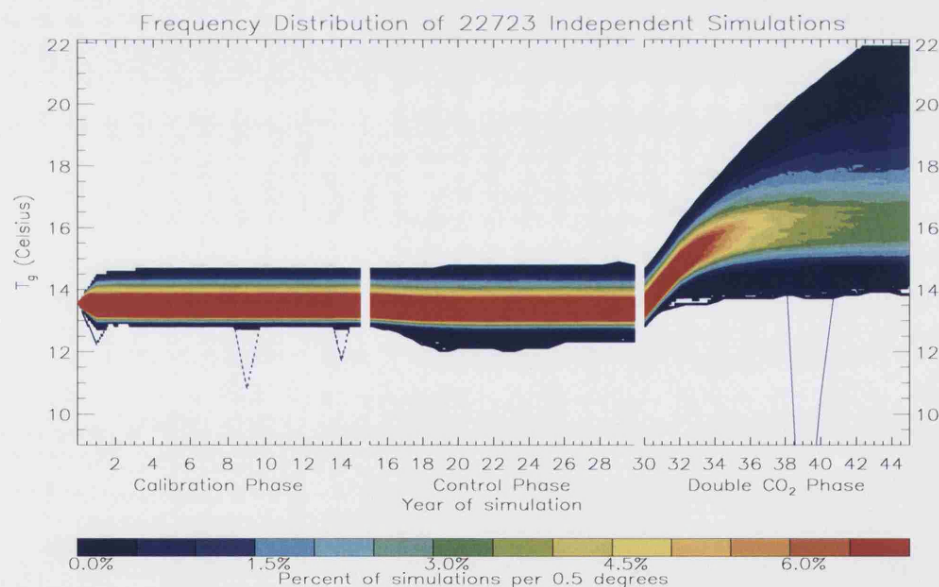


Figure 7.1: The time series of model version mean (averaged over available quality controlled ICE members) for the three phases of the experiment. Most simulations warm rapidly in the final phase, some by over 8 degrees by the end of the 15 year doubled CO_2 phase. There are some simulations with unsmooth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

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

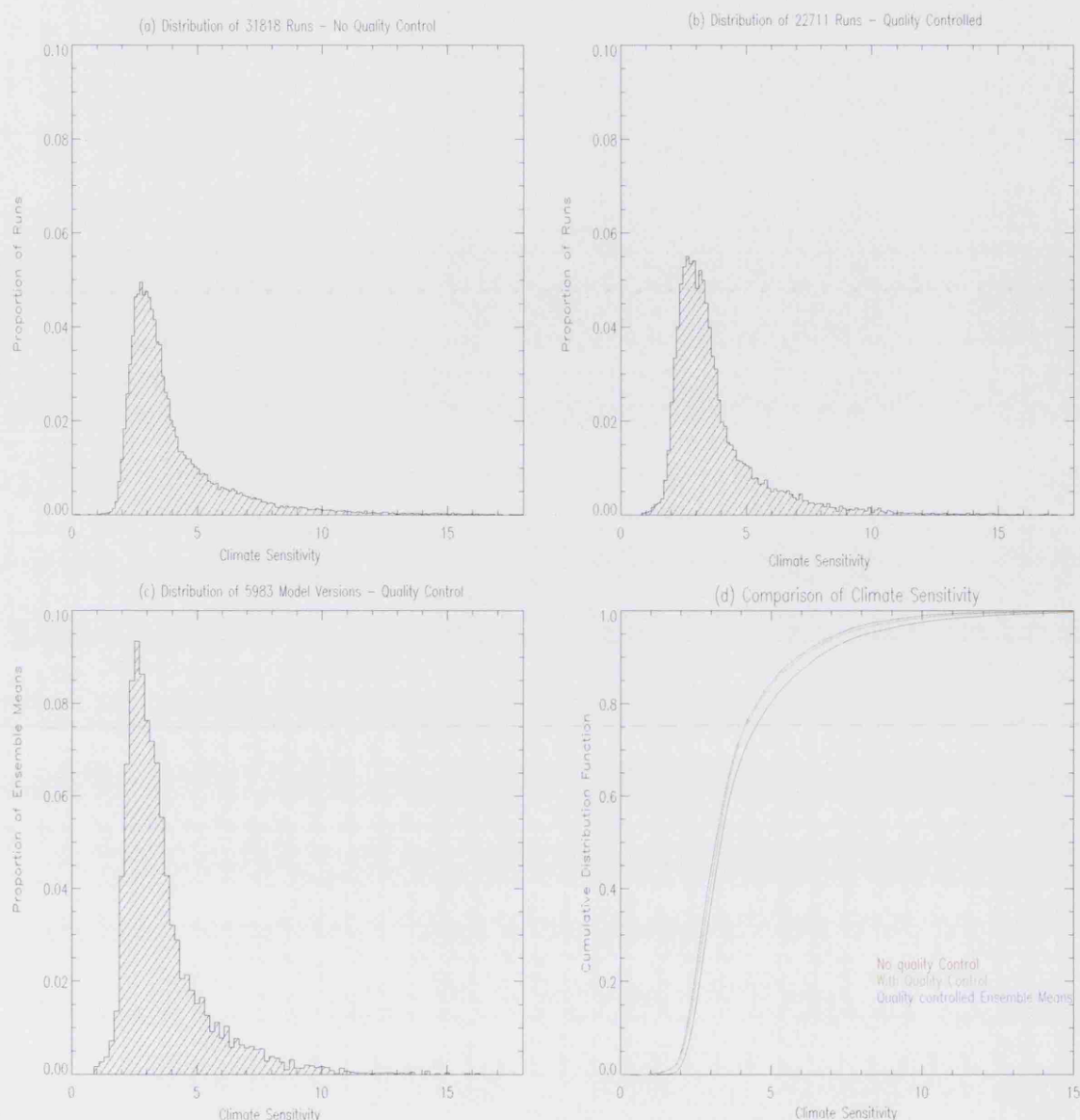


Figure 7.2: The distribution of CS in the CPDN PPE. Panel (a) shows the distribution of all simulations, panel (b) the distribution of quality controlled simulations. Panel (c) shows the ICE mean over all model versions, for quality controlled simulations. Panel (d) shows a comparison of the three different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

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

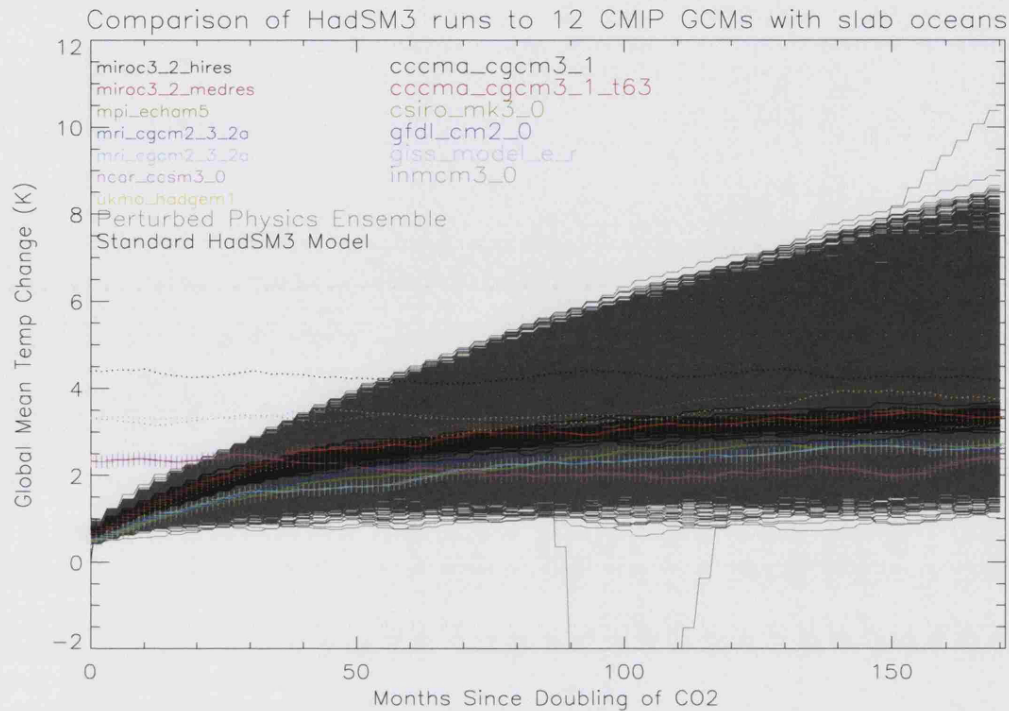


Figure 7.3: The change in temperature following a doubling of CO_2 is shown for 12 CMIP II models and the CPDN ensemble. For some CMIP simulations, data pertaining to the transient period of warming immediately following a doubling of CO_2 was not available.

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

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

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

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

7.4 Sub-global Behaviour

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

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

Figure 7.4 reflects the spatial variations in temperature change for each grid box and the associated variance. Figure 7.4 shows the mean (upper panel) and variance (lower panel) of the change in 8 year annual mean 1.5 metre surface temperature between the control and doubled CO_2 phases for 22698 quality controlled simulations¹.

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

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

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

¹22698 simulations are used in this Section in contrast to the 22711 quality-controlled simulations used in Section 7.3 since 13 simulations have requisite data for the calculation of CS but not for producing the required regional fields. Similar variations exist elsewhere in this Thesis due to a corresponding effect.

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

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

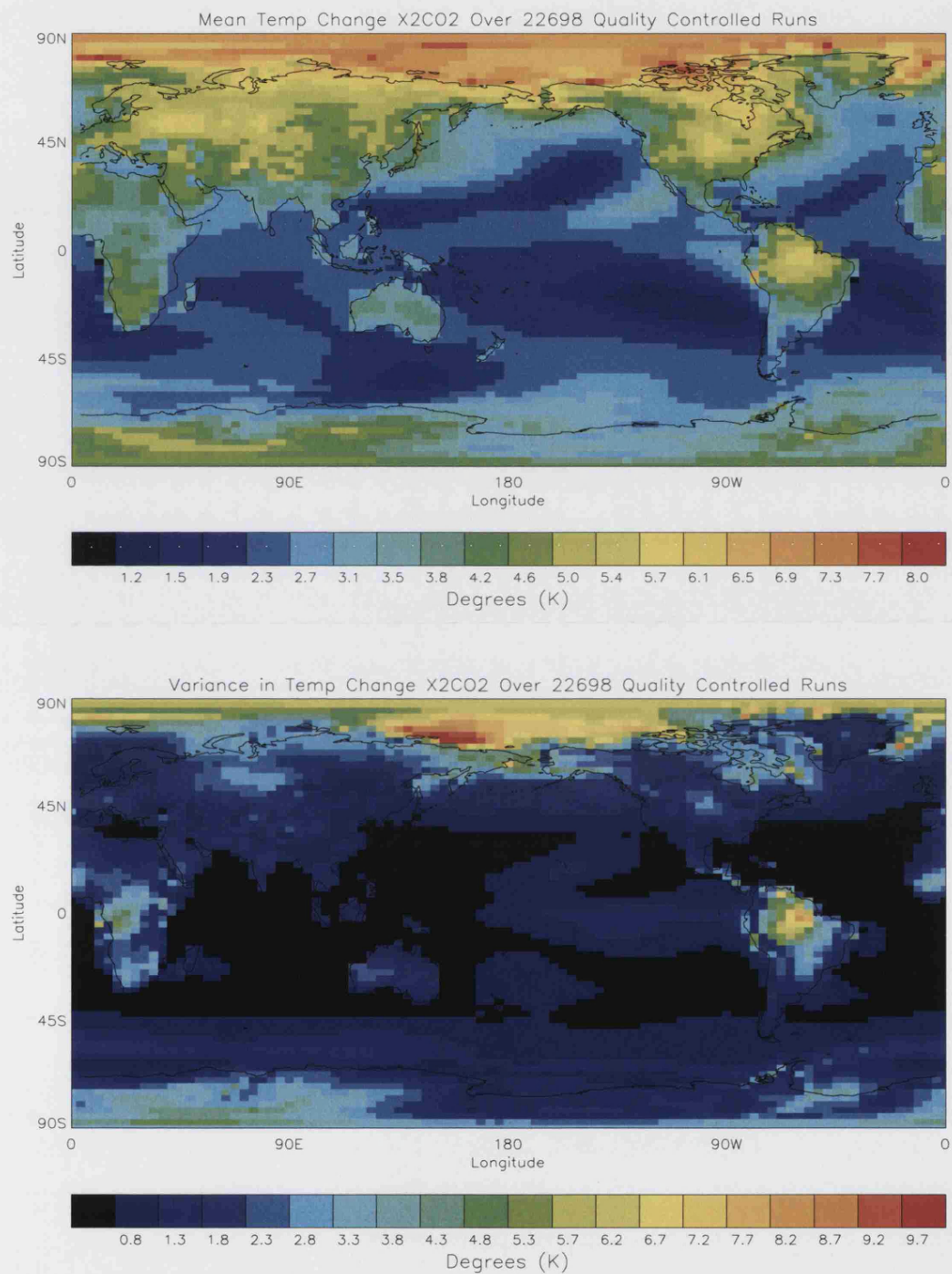


Figure 7.4: The mean (upper panel) and variance (lower panel) of 8-year mean annual mean temperature change between the pre-industrial CO_2 calibration phase and the doubled CO_2 phase over 22698 simulations. Warming is greater in the centre of large masses and in the Northern high latitudes. Warming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

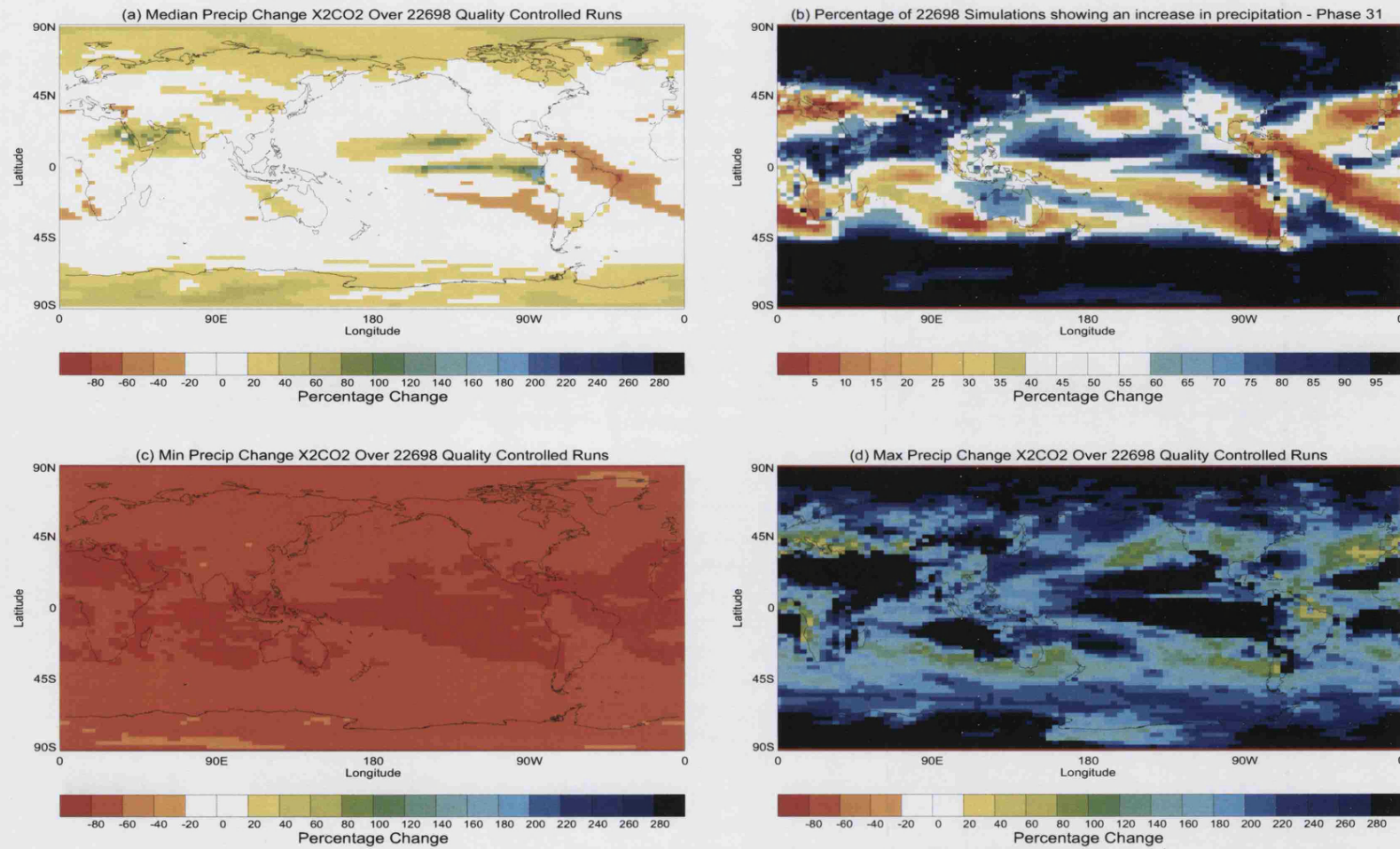


Figure 7.5: The mean (a), democracy plot (b) and bounding box of 8-year (the minimum is shown in panel (c) and the maximum in panel (d)) mean precipitation change between the pre-industrial CO_2 calibration phase and the doubled CO_2 phase over 22698 simulations. The democracy plot shows the percentage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

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

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

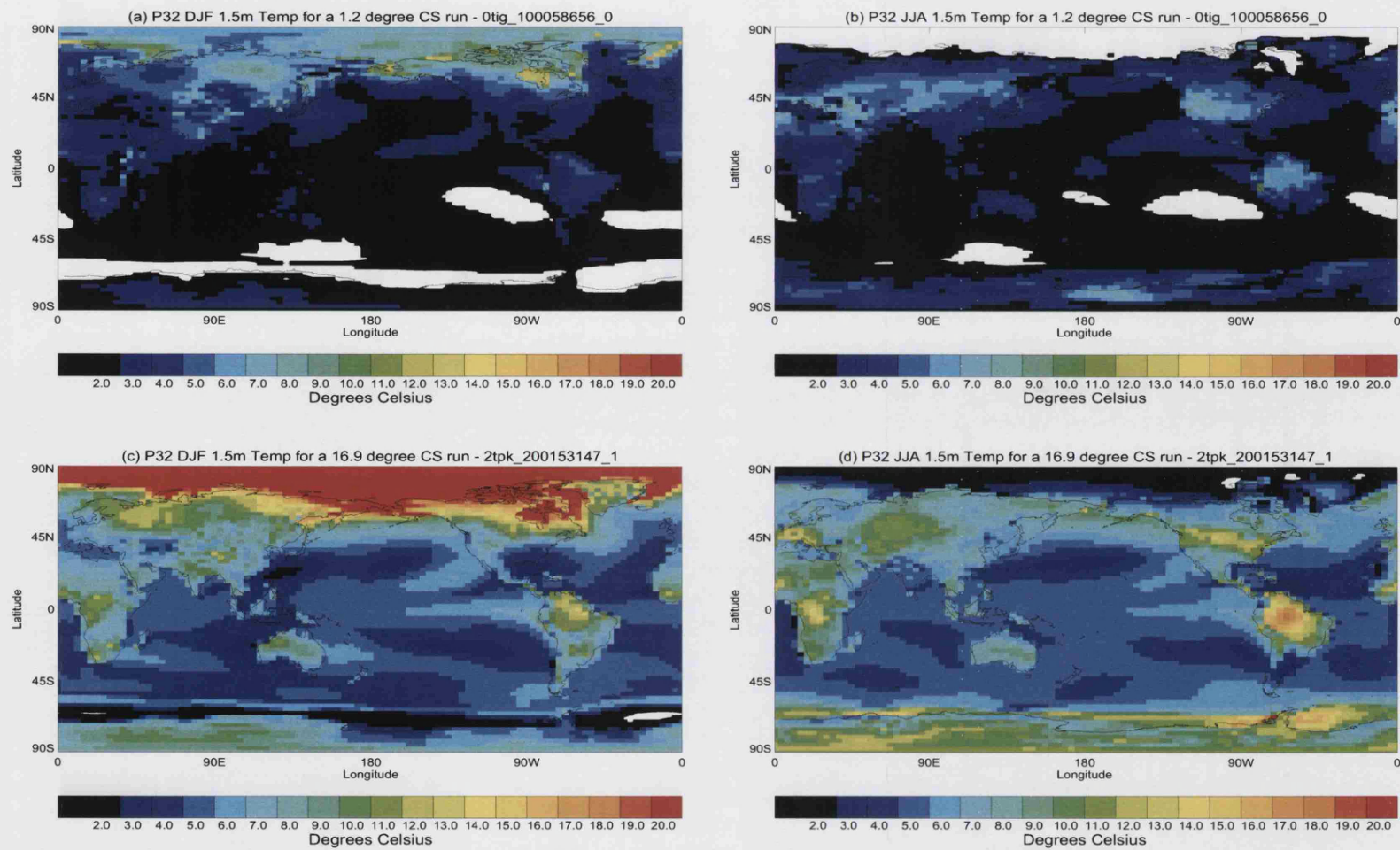


Figure 7.6: The change in temperature from phase 2 to phase 3 is shown for two simulations. These simulations were selected on the basis of having very low and very high climate sensitivities of 1.2 and 16.9 degrees, respectively. Panel (a) shows the DJF change for the 1.2 degree CS simulation and panel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

7.5 Constraining Model Simulations

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

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

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

7.5.1 Constraining using the Entrainment Coefficient

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

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

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

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

of CS would have been obtained.

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

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

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

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

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

7.5.2 Constraining using HFA

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

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

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

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

7.5.3 Constraining using in-sample fit to observations

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

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

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

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

The Metric

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

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

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

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

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

$$\theta_{s,j}^2 = \frac{\epsilon_{s,j}^2}{\epsilon_{h,j}^2} \quad (7.2)$$

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

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

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

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

Perturbed Physics Ensemble

The performance of simulations from the CPDN PPE is examined in this Section. Figure 7.10 shows the mean RMSE over 5 variables (to allow comparison of RMSE to CMIP simulations, 5 variables are used here) of 22712 simulations versus CS. The 5 variables used here are: annual mean temperature, sea level pressure, precipitation, and atmosphere–ocean sensible and latent heat flux; these were the only five variables were available from the CMIP II simulations used for comparison¹. The CMIP GCMs are taken from the Coupled Model InterComparison Project II experiment Covey *et al.* (2003), developed at modelling centres across the world. The mean RMSE is found by taking the arithmetic mean of RMSE over these 5 variables, as in Stainforth *et al.* (2005). The use of such a composite score is not recommended here, since differences in the variability of RMSE between different variables could, without standardisation, lead to some variables gaining more influence than others. Nevertheless, this mean RMSE score is shown in order to highlight the importance of considering the performance of simulations in the individual variables that comprise the mean score.

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

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

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

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

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

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

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

Applying the constraint in all variables (a simulation must pass in all 7 variables to be included) give the top-most CDF in Figure 7.11, shown in the upper panel; the highest admitted CS is about 5 degrees Celsius. This shows that it is possible to constrain model diversity by using more tests and drastically reducing the number of simulations (408 pass the test in all variables – $\sim 1.4\%$ of the original number). Constraining based on a single variable gives different results depending on which variable is used. Using more variables applies a stricter constraint but it is not clear how many should be used. Using all the data available to constrain model simulations would result in the CDF tending towards the standard model, since this is standard for in-sample fit used as the reference point². If enough variables were included, at a sufficiently strict level of compatibility with observations, all simulations could be ruled out. It is not clear what level of model–observation agreement should be demanded when constraining model simulations.

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

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

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

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

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

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

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

2. The global mean statistic of RMSE used is here can only be a rough guide to model performance. More detailed spatial and temporal analysis is required for decision-relevant model evaluation.
3. The constraint is based on a static climate; no dynamical changes are included. Arguably, the ability of models to match dynamical changes in the Earth's climate system are the most important test of likely out-of-sample skill. Such a test can not be carried out using the data analysed here since CO_2 is instantaneously doubled rather than increased transiently. Attempts by Annan & Hargreaves (2006); Knutti *et al.* (2006) to constrain CS using dynamical changes have yielded a reduction in high CS values, although the shape of this distribution itself is, to some extent, an artifact of the experimental design Frame *et al.* (2005).

7.6 Conclusion

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

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

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

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

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

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

- The range of behaviour shown in an ensemble of 45644 of GCM simulations is unprecedented, with estimated CS ranging from 0.9 to over 16 degrees Celsius.
- By comparing sub-sets of the CPDN PPE it has been shown that the shape of the distribution of CS is not an inevitable feature resulting from an approximately Gaussian distribution of feedbacks, as was proposed in Roe & Baker (2007). The distribution of CS can be changed substantially by a different choice of experimental design.
- The global mean HFA and the Entrainment Coefficient can be used to change the distribution of CS, but simulations with CS over 8 degrees are still admitted. The use of such post-hoc filters is criticised on the basis of bad statistical practice.
- When comparing model performance in-sample in 7 different variables, results depend on the choice of variable. For example, constraining in temperature tends to admit more high CS simulations and constraining in precipitation more low CS simulations.

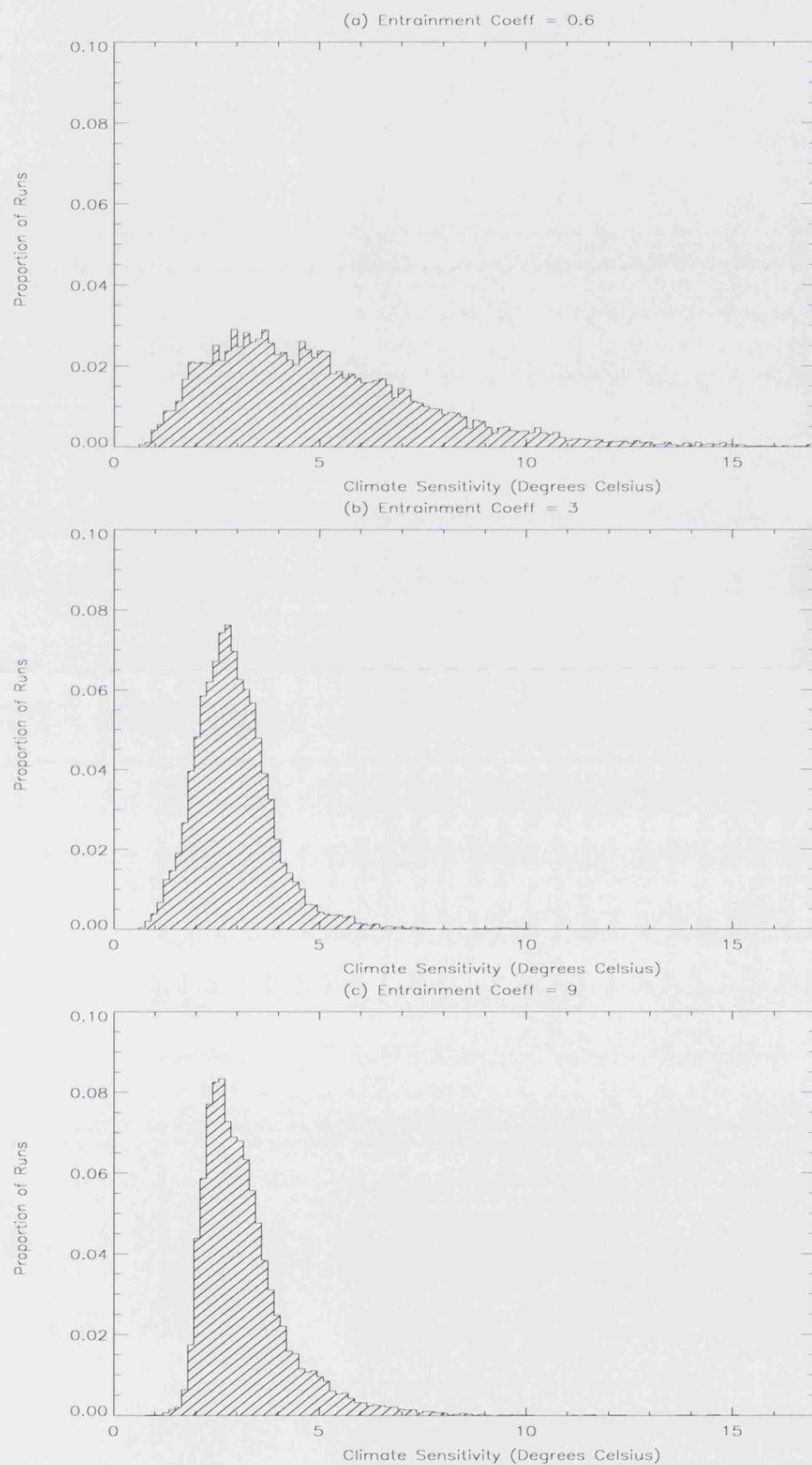


Figure 7.7: The distribution of CS is shown for all quality controlled simulations for three different values of the Entrainment Coefficient - 0.6 (low) in panel (a), 3 (standard) in panel (b) and 9 (high) in panel (c).

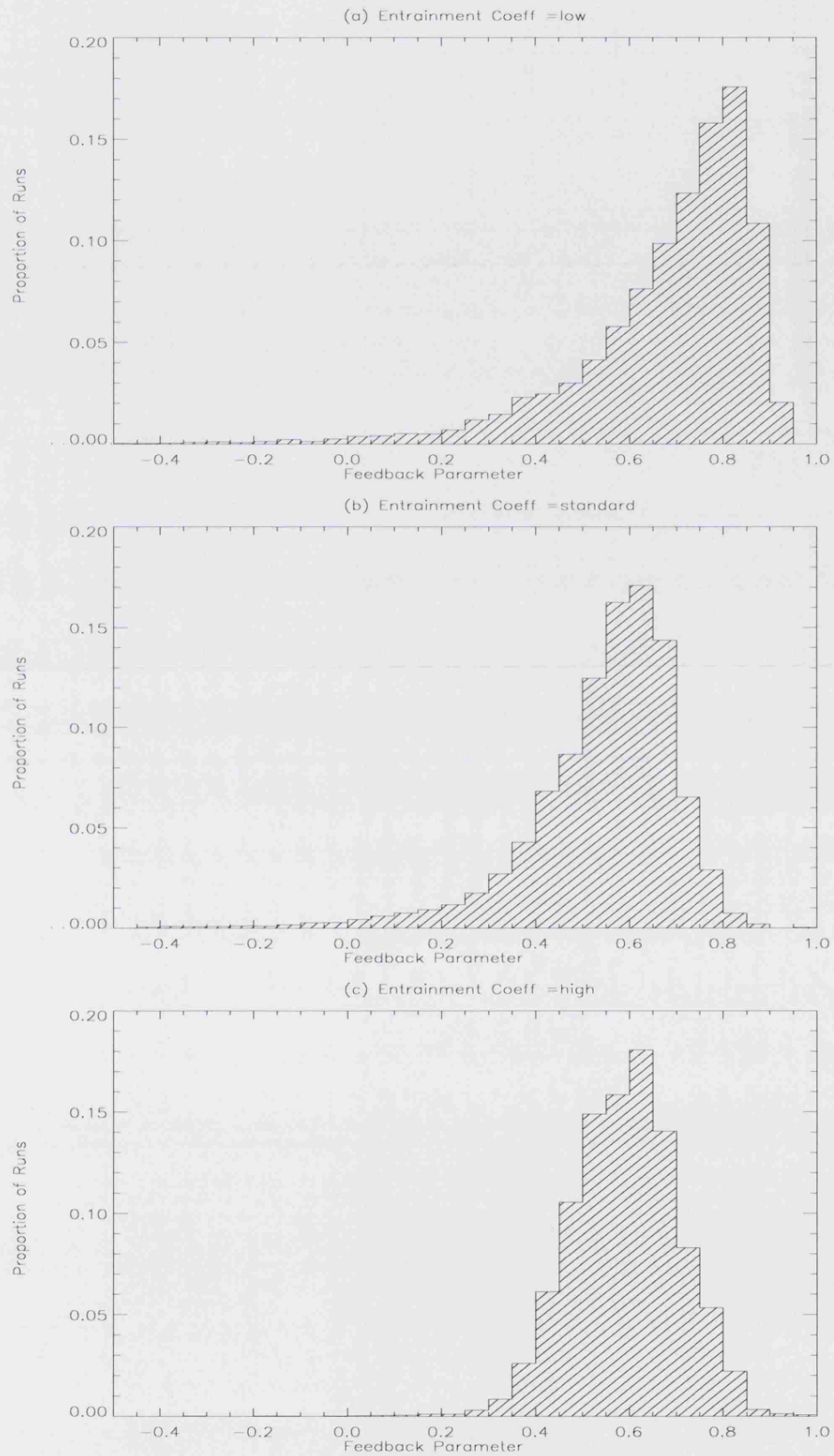


Figure 7.8: The implied distribution of feedbacks for three different values of the Entrainment Coefficient

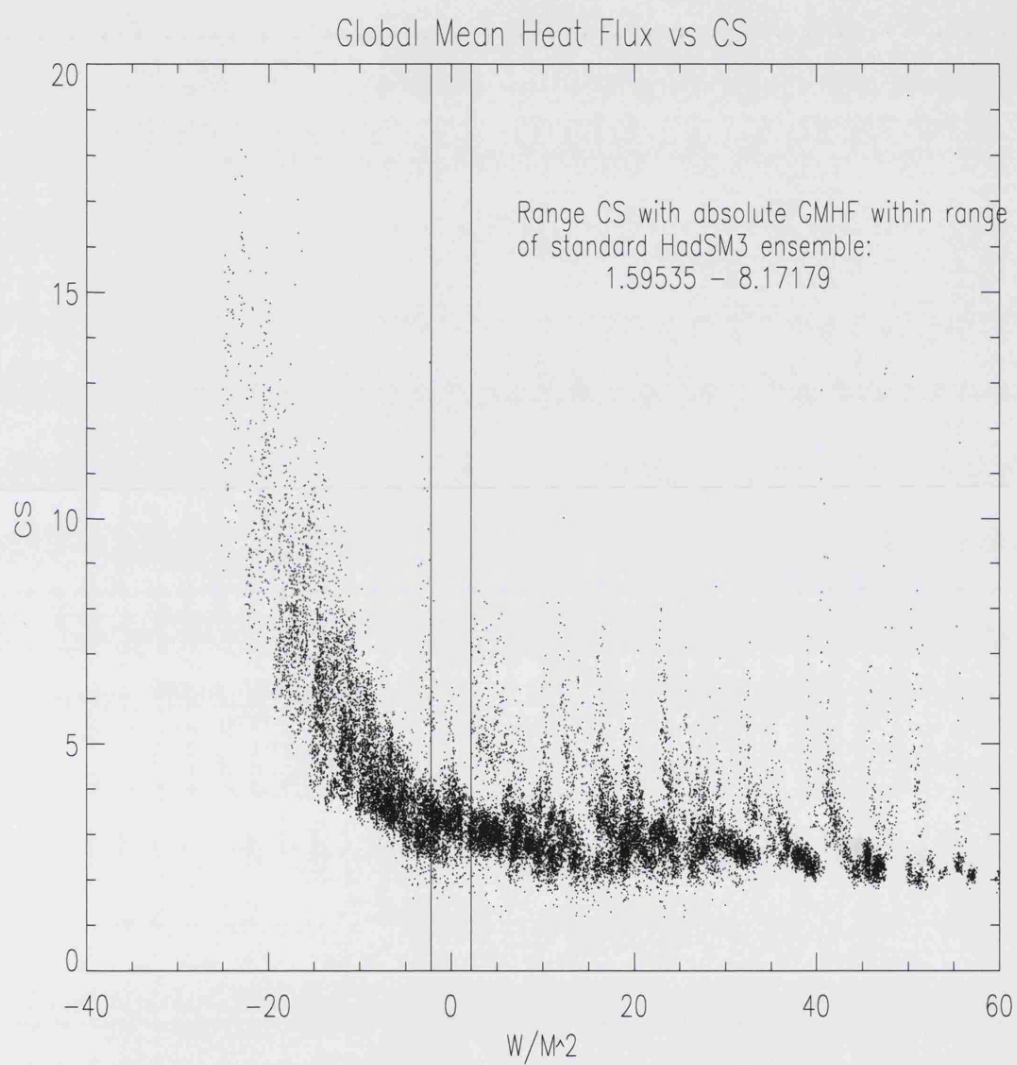


Figure 7.9: The global mean HFA is plotted against CS. The vertical lines denote the largest absolute values of global mean HFA in the 64 members standard ensemble. The range of CS captured by this range is (1.59535, 8.17179).