Climate Model Ensemble Generation and Model Dependence

by

Ned Haughton

Supervisor: Gab Abramowitz

University of New South Wales

November 2012

Declarations

I hereby declare that this thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis; that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due acknowledgement is made in the text of the thesis; and that this thesis does not contain any material which infringes copyright.

Signed: _____

Ned Haughton

Abstract

Climate model projections have major environmental, social and economic ramifications, and, if taken seriously, could contribute to saving millions of species and lives, or causing losses of billions of dollars. With so much at stake, it is vital that these projections are as accurate as possible. Recent research has indicated that the properties of the climate model ensembles used to make these projections are not optimal. One of the major goals of the climate modelling community for the foreseeable future must be to resolve these problems. We must better understand what properties optimal ensembles need to have, and develop tools for analysing these properties. We must understand the effect that different ensemble generation approaches have on these properties, and the effect that different weighting methodologies have on ensemble projections. We must then use this knowledge to help direct the creation of model ensembles, either by optimally combining existing model simulations, or by actually designing ensembles with the properties we need. This is a process that will take years at best, and can not be encompassed in a single thesis. This thesis aims to be a stepping stone in this process.

We use a low-resolution general circulation model (GCM) to generate three ensembles using different perturbation techniques: initial conditions perturbations, physical parameter perturbations, and structural perturbations (representing a multi-model ensemble). We compare the three ensembles using simple averaging, performance based weighting, and dependence weighting to look at both ensemble mean and spread. We find that the different techniques produce very different ensembles. Initial conditions perturbations produce ensembles that are too narrow relative to variability in the observations, while perturbed parameters and structural ensembles apparently exhibit too much spread. We then use the same procedures to compare projections generated by the same ensembles. Projections using unweighted averaging are likewise too narrow for initial conditions ensembles, and too broad for structural ensembles. Performance weighting is shown to improve the mean estimate, but may actually degrade the estimate of variance. We show that Bishop and Abramowitz's (2012) independence transformation can improve ensemble mean and variance projections.

Acknowledgments

My thanks go first to Gab Abramowitz, who, despite having a fairly tumultuous year, was the best supervisor a person could hope for, and an amazing editor. To Andy Pitman, for acting as supervisor while Gab wasn't available and making sure I knew what happened to students who strayed from the path, as well as for being an excellent editor. To Steve Phipps, for being incredibly patient and helpful, and cutting months off the time I would have otherwise needed to get my models running[1]. Thanks to Anni and Dad and Ollie for being supportive and for all the editing and comments. And to Mum, for editing my thesis twice, and asking hard questions.

Software used for this project and thesis included Mk3L Climate System Model (Mk3L), the R statistical language, and ggplot2, LaTeX and Vim. All code used for model setup, data manipulation, and data analysis was stored in private Git repositories on BitBucket.org. The LaTeX code for this document was stored in a similar fashion. The model runs were conducted on the University of New South Wales (UNSW) Mathematics Department's "Tensor" computing cluster. All data was stored on the Climate Change Research Centre (CCRC)'s shared data facilities.

I would like to thank the ARC Centre of Excellence for Climate Systems Science for providing a generous scholarship, without which this year would have been much more difficult. And thanks to all the excellent crew at the UNSW Climate Change Research Centre, for excellent discussions, feedback, support, and for being so welcoming!

[1] Mum, these are the people who I'm referring to as "We" in the thesis

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Modelling the climate system

To predict the effects of global warming, we must understand the Earth's climate and how it changes. The chaotic nature of the Earth's climate makes prediction difficult (Giorgi, 2005). While many of the mechanisms of the climate system are reasonably simple, their effects can interact in complex ways, making the task of understanding the whole system a demanding one. Because of this complexity scientists create numerical models of the Earth's climate. Such models, beginning with simple energy balance models, have evolved over the last few decades to become highly complex and much more comprehensive. Climate models attempt to replicate the processes of the planet's climatic systems (for example thermodynamics, fluid dynamics, ecosystem processes) in a way that allows us to experiment with the parameters of the system, and draw conclusions about how different forcings create different behaviours. Modern general circulation models (GCMs) couple atmospheric models with cloud dynamics, eddy-resolving ocean models (Maltrud and McClean, 2005), land surface models, and biochemical models (Pitman, 2003).

However, even the most high powered modern supercomputers have limited power to process these models. GCMs must be run on quite coarse spatial and temporal scales – the most advanced models use grids on the order of hundreds of kilometers, and time steps of a few minutes (eg. the HadGEM3 model, Met Office, 2010). This is problematic, because the physical processes contributing to the climate occur on much smaller scales: land surface change occurs on the order of meters, while atmospheric turbulence, cloud physics, and biochemistry occur on microscopic scales (Sellers and Trenberth, 1992). Numerical models of the climate must therefore approximate or parameterise processes to work on larger scales. Because we are no longer dealing with the physics directly, but with a statistical representation of the physics, it is conceivable that multiple different models could provide appropriate approximations.

Climate modelling is a unique field in that there is only one realisation of a partially chaotic system, with no true replicates, and we are attempting to match

our models to that realisation. This makes it hard to draw inferences about the true distribution of potential climate states at any one time, just as if you only had one person's height as data, you would have difficulty drawing any sensible conclusions about the height of human beings in general. One possible solution to this problem is to create lots of models of the Earth, and use them together in an ensemble to draw information and projections about the climate.

1.1.1 How do we make use of multiple models?

A climate model ensemble is a set of climate models, run over the same period and region, used to draw inferences about the present or future climate. One of the most utilised and discussed examples of the climate model ensemble is World Climate Research Programme (WCRP) Coupled Model Intercomparison Project Phase 3 (CMIP3) 20th Century ensemble (Meehl et al., 2007), which formed the basis for the projections used in the Intergovernmental Panel on Climate Change (IPCC) 4th Assesment Report (AR4). This ensemble contains model runs submitted by various institutions that cover the entire globe, and run mostly over the time period 1850-2000 (see Figure 1.1).

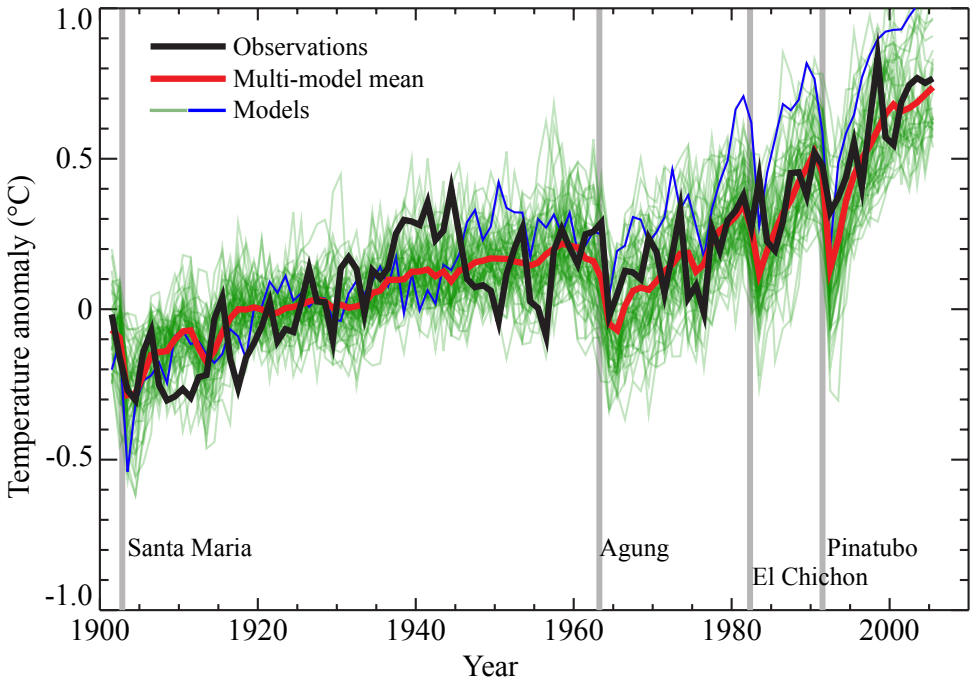

Figure 1.1: CMIP3 model ensemble, after Hegerl et al. (2007). Models are shown in green, the multi-model mean in red, and the observations in black. An arbitrary single model is shown in blue to highlight the differences between variability seen in the model runs and observations and in the multi model mean.

When attempting to predict the future based on models, we have two main possibilities: choose the best model (based on our assumptions), or somehow take

information from an ensemble of models and combine it into a single estimate. In some fields – especially those relying on empirical statistical models instead of numerical models for time series projection – the emphasis has been on model selection rather than model combination. Model selection generally works best when selection is stable. That is, when there are very small changes in the system being modelled, model selection does not change, or only changes slightly (Zou and Yang, 2004). The possibility of choosing the *best* model is most useful for producing a single prediction, but selection does not allow for uncertainty estimates as readily as combination does.

While there may be some value in climate model selection on the basis of theoretical grounds, a single model will still produce significant variability due to slight variations in initial conditions, or parameter values. Selection of a single climate model *run* is not viable: initial conditions can never be known to a high enough degree of accuracy, significant uncertainty exists in the values of various physical parameters, and grid and time-scale limitations mean that a single run will never adequately represent the true climate. In these situations, model selection becomes highly unstable, and combination becomes more suitable (Zou and Yang, 2004). In contrast, combination potentially allows us to explore the range of uncertainty in the system, and present a probabilistic best estimate. Tebaldi and Knutti (2007) argue that a "variety of applications, not only limited to the weather and climate prediction problems, have demonstrated that combining models generally increases the skill, reliability and consistency of model forecasts".

It should be noted, however, that even when using model combination, "de-facto model selection" still occurs: older versions of models and outliers (models that perform oddly, or show extreme changes) are often discarded (Knutti et al., 2010b). There are problems with both procedures. Knutti et al. (2010b) note that "If we indeed do not clearly know how to evaluate and select models for improving the reliability of projections, then discarding older results out of hand is a questionable practice". The question of how to deal with outliers is far broader, and is applicable to most science. Chatfield (2004) states that "The 'outlier' may be a perfectly valid but extreme observation which may for example indicate that the data are not normally distributed." It is possible that the distribution of possible climate states is highly non-normal, and assuming normality could cause estimates based on selected models to be overly optimistic (Tebaldi and Knutti, 2007). If we take into consideration the political pressure surrounding climate science, it is clear that decisions relating to model selection should not be taken lightly.

Once we have a selection of appropriate models – an *ensemble* – we must find some way of combining the information it contains. How we do this depends on our assumptions about the relationship between models and observations.

1.2 Paradigms for interpreting model ensembles

The question of how to interpret a climate model ensemble remains an area of active research. There are a number of different existing paradigms.

1.2.1 Truth plus error paradigm

This paradigm assumes "that the climate system and all the processes that affect it are entirely (i.e. deterministically) predictable from climate forcing variables" (Bishop and Abramowitz, 2012). Under this paradigm, we implicitly see model runs from non-perfect climate models as centred around the observations, with pseudo-random noise that represents flaws in the model, computational inadequacy, or initial condition uncertainty. This paradigm has been called the "truth-centred paradigm" (Annan and Hargreaves, 2010), or the "truth-plus-error" conception of modelling (Knutti et al., 2010b). This is the prevailing approach to interpreting model ensembles (Annan and Hargreaves, 2010).

The assumption of random distribution of error noise in the truth-plus-error paradigm leads to the conclusion that, as an ensemble increases in size, the mean of the sample should converge toward the observations, and uncertainty will converge to zero as model errors are averaged out. Unfortunately, if we accept this paradigm as reliable, we should immediately be worried about the state of current climate models. If we examine Figure 1.1, we see that while the models do a reasonably good job in many respects, there is significant difference between the multi-model mean (red), and the observations. In particular, there are long periods where the multi-model mean and the observations are quite distant, and, except where there are strong volcanic forcings, the multi-model mean exhibits significantly less variability than the observations.

This begs the question of whether the climate is purely deterministic or exhibits chaotic behaviour. If we could re-run the Earth given the exact conditions from some date in the past – with some arbitrarily small change – should we expect that the historical patterns of climate and weather to be exactly the same?

1.2.2 Indistinguishable paradigm

Annan and Hargreaves (2010) provide the first paper to explicitly refute and provide and alternative to the truth plus error paradigm. They assert that models and observations should be treated the same, that is, as if they were indistinguishable random draws from an underlying probability distribution. They term this the "statistically indistinguishable paradigm".

The indistinguishable paradigm addresses the assumption of convergence to the mean: because the observations act like a model, and contain a certain amount of error, as the model ensemble size increases the ensemble mean should *not* converge to the observations, but to a statistical centre of the distribution, and uncertainty will converge to a value relative to the width of the distribution. The multi-model

mean does not have the same attributes as a true Earth-like climate, because the combination process flattens out extremes. In particular, Gleckler et al. (2008) show that the multi-model mean has smaller errors, and that the variance of the mean is lower, than individual models. The multi-model mean does not represent a potentially real climate. This paradigm anticipates that the mean of an ensemble will have much lower variance than the observations (as is observed in Figure 1.1), as the observations can be expected to contain some error relative to the underlying distribution. The meaning of variability in ensembles is not explicitly described in Annan and Hargreaves (2010), but is assigned in a later blog post to "collective uncertainties about how best to represent the climate system" (Annan, 2010).

1.2.3 Replicate earth paradigm

Bishop and Abramowitz (2012) expand on this idea with their "replicate earth paradigm", arguing that it is not safe to assume, as the indistinguishable paradigm does, that models represent independent draws from the underlying distribution. They argue that if we assume that the Earth's climate is, to some extent, chaotic, and that no two Earth-runs are likely to be the same, then we must view the true Earth observations as a sample picked from a climate probability density function (CPDF). Under the replicate earth paradigm, variability in the CPDF is assigned to the chaotic component of the climate system, while the CPDF mean represents the deterministic component. The CPDF mean is much smoother than the observations (in the same way that the multi-model mean is smoother than the models in an ensemble), and has an instantaneous variance that is close to the variance of the observations around the CPDF mean, over time. If we re-ran the Earth from some point in the past, assuming the same boundary conditions, the new "replicate earth" would also be drawn from this distribution.

However, there is no guarantee that climate models adequately represent true replicate earths:

> "Climate models can be viewed as imperfect attempts to create replicate earths. We suggest that a perfectly independent model's predictions should be a random draw from the time-evolving CPDF. In this case, the mean of an ensemble of perfect models is simply an approximation of the mean of the CPDF. Since the real Earth itself is also a random draw from the CPDF, we should not expect observations of it to match this mean, but rather be equivalent to a perfect model. (Bishop and Abramowitz, 2012)

Bishop and Abramowitz (2012) set out two key properties that a model ensemble must meet before it can considered an adequately representative sample of the CPDF:

1. "The equally weighted mean of an ensemble of replicate earths is the linear combination of replicate earths that minimizes the dis-

tance from our Earth's observations." This implies that the models representing replicate earths must be independent.

2. "The time average of the instantaneous CPDF variance should be approximately equal to the variance of the real Earth about the CPDF mean over time." This essentially states that the variance of replicate earth-like models must have a variance about the CPDF mean similar to that of the observations.

Bishop and Abramowitz (2012) show that the CMIP3 ensemble does not fit these criteria, and cannot be considered as an ensemble of true replicate earths. They present a transformation methodology (described in Section 2.4.1) that brings an ensemble of poorly performing models closer to approximating these criteria.

1.3 Methods of model combination

How then do we best take information from multiple models and combine it into a single projection? Significant challenges face those wishing to decide on methods of interpreting ensembles. "Among these challenges are that the number of models in these ensembles is usually small, their distribution in the model or parameter space is unclear, and that extreme behaviour is often not sampled" (Knutti et al., 2010a). Even when we can overcome these problems, we also need to take into account the fact that different models perform differently, and that ensembles do not always behave like replicate earth ensembles.

1.3.1 Unweighted averaging

The simplest, most intuitive way is to take the arithmetic mean of the model outputs. This methodology has been a commonly used in the past, for example in the IPCC's AR4 (Solomon et al., 2007). The method has a number of advantages: It is a simple computation, and it provides a mean (although not necessarily the best performing mean, see discussion of model dependence below), as well as estimates of uncertainty (e.g. a confidence interval). However, for this process to be valid, two assumptions must be met:

- That the underlying distribution is symmetrically centred about the true climate (e.g. a normal distribution), and
- That the ensemble distribution is representative of that underlying distribution (i.e. randomly drawn with no bias).

1.3.1.1 Assumption of underlying symmetrical distribution

It is not clear that this assumption is valid. Firstly, it is not known, and possibly can not be known, whether a true climatological distribution exists. We have only a single realisation of the Earth's climate, which can give no indication of the breadth

of an underlying distribution (or the total lack of breadth, if the climate is purely deterministic). Bishop and Abramowitz (2012) argue that there is no evidence indicating that it is better to assume an entirely deterministic system.

If we assume that there *is* a chaotic component, the next question is whether the Earth's CPDF is reliably symmetrically distributed at any point in time. If the distribution is significantly skewed, the mean and variance will provide a biased estimate of the distribution. Since we have only one sample, there is no way to determine whether this distribution is skewed or not. If higher moments of the distribution are also non-normal, we should also expect variance estimates to be wrong. Whether current models appropriately represent this distribution is hard to verify or falsify (Masson and Knutti, 2011).

1.3.1.2 Assumption of a representative sample

The assumption that an ensemble is a representative sample of the underlying distribution is only true if the models in the ensemble are independent. Under the truth-plus-error paradigm, independence is required for the model ensemble average to converge to the true climate (Tebaldi and Knutti, 2007). Under the indistinguishable paradigm and the replicate earth paradigm, independence is required for the ensemble to accurately approximate the CPDF. If the models in the ensemble are not independent, i.e. they are not truly randomly sampled from the CPDF, then the ensemble mean will not converge to the true climate (under the truth-plus-error paradigm), or to the CPDF mean (under the replicate earth paradigm).

Annan and Hargreaves's indistinguishable paradigm assumes that the CMIP3 models are relatively independent. Bishop and Abramowitz (2012) show that this is not the case, and strengthen the requirements of this convergence, by showing that for the ensembles to converge to the true CPDF, the models must be "replicate earths" – maximally independent models with statistically similar properties to the observations.

Unfortunately, in large multi-group modelling ensembles, such as the CMIP3 ensembles, the assumption of model independence is unlikely to be met:

> For the most recent coordinated modelling effort archived at Project for Climate Model Diagnosis and Intercomparison (PCMDI) [CMIP3, 2005], several groups submitted more than one model or model version, e.g. one model was run at two different resolutions but the same physics; one ocean was coupled to two different atmospheres. In those cases, the models are clearly not independent, and their biases against observations are probably highly correlated. Sharing components and knowledge is not bad a priori, but it will result in persistent biases in a multi-model mean, whether weighted or not. (Tebaldi and Knutti, 2007)

This is likely to become even more of a problem in Coupled Model Intercomparison Project Phase 5 (CMIP5), as some modelling groups submit hundreds of runs, while

others submit only a handful (Taylor et al., 2012). A method allowing us to deal with model dependence is urgently required, and is discussed further in Section 1.4.2.

1.3.2 Performance weighting

Some models perform better than others. This may be due to, for example, more accurate algorithms and parametrisations; higher resolution, capturing more complexity; or the inclusion of more physical components. It makes intuitive sense to treat the output of such models with higher regard. We can adjust our predictions by calculating the performance of each model, and weighting better performing models more heavily.

Performance (or *skill*) is generally calculated by some measure of distance between a model run and observations. The difference from observations can be calculated in any number of ways, depending on the purpose of the experiment. Common measures include root mean square error (RMSE), and covariance. These measures may be calculated for any number of variables, regions, or time-spans. Specific cost-functions used in this thesis are discussed in Section 2.4.

One way of dealing with performance differences is to remove from an ensemble models that are below a specified performance threshold. This is a form of automated ensemble selection. Such a removal process can also be used to select a subset of an ensemble for other practical reasons, such as computational limitations for further processing. The use of performance measures can help to remove some of the subjectivity from the decision to remove outliers, although subjectivity is still required when deciding on a threshold.

Another way of dealing with the problem of model performance differences is to weight the multi-model average based on performance. Tebaldi and Knutti (2007) advocate this, arguing that "models with small bias and projections that agree with the ensemble 'consensus' should be rewarded while models that perform poorly in replicating observed climate and that appear as outliers should be discounted." A weighted average can be quite flexible, in that performance differences can be mapped to large differences in weights (giving precedence to well-performing models), or to smaller differences (allowing the less well performing models to still impact the results). This has the benefit of taking some information from all models, while reducing the susceptibility of the mean to outlier-introduced bias.

Weigel et al. (2010) show that performance-weighting-based projections can result in significantly improved accuracy. However, they also show that if the performance weighting is not related to the underlying uncertainty, performance weighting can have a detrimental impact on projection accuracy.

Macadam et al. (2010) raise a critical question: "is the skill of an [coupled atmosphere-ocean general circulation model (AOGCM)] in the past a useful guide to the skill of the AOGCM in the future?". Reifen and Toumi (2009) show that performance ranking based on temperature anomalies is inconsistent over different in-sample periods (different decades of the 20^{th} century), suggesting that performance over out-of-sample periods (i.e. projections) would similarly be inconsistent.

Even if these problems are resolved, a key problem remains: dependence between models is not addressed by performance weighting. By reducing the impact of *unusual* model runs, performance weighting may actually increase inter-model dependence in an ensemble. We can imagine a small ensemble, where some models perform better than others, and some pairs of models are more dependent other pairs: in some situations it may make sense to weight based on performance, in others based on independence (see Figure 1.2). There is a trade-off to be made here in most cases, but before we can understand that trade-off, we must find ways of dealing with dependence.

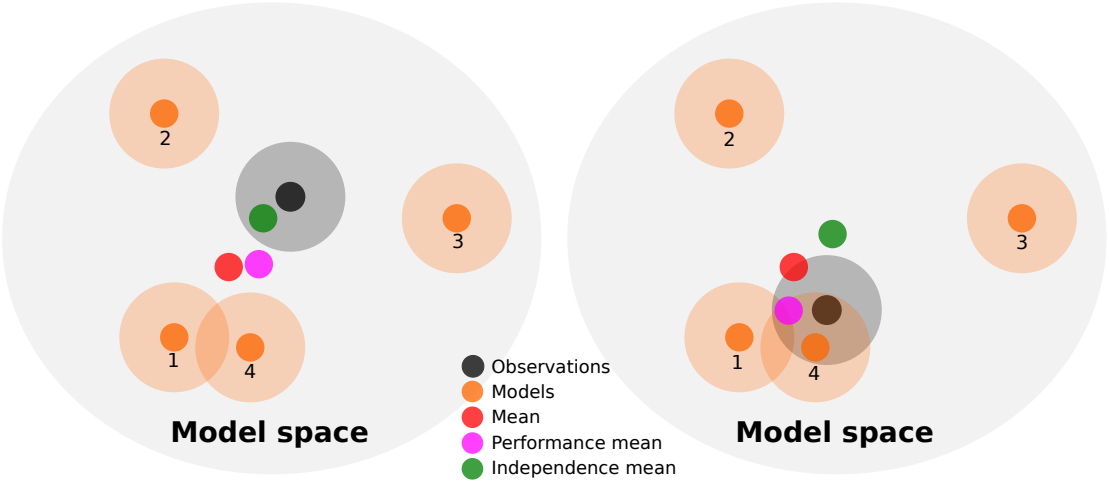

Figure 1.2: Conceptual diagram of two potential cases where performance weighting must be traded-off against independence weighting, after Abramowitz (2010). In some instances, dependence can skew an ensemble – in the left example, down weighting models 1 and 4 might be beneficial. In other cases, performance weighting is more important – models 1 and 4 are simply more accurate in the right diagram.

1.4 Dealing with model dependence

> *"Despite the ever-recurrent comment about the need of accounting for model dependence, no formal approach at quantifying this dependence has been worked out yet. A distance in model space is definitely a difficult concept to formalize."* Tebaldi and Knutti (2007)

Independence is a difficult concept, but may be loosely defined as *the ability for new data to add new information to a dataset.* Adding more dependent data to a data set makes it *more* difficult to generalise about the data, since assumptions of data distribution (e.g. normality) are broken.

A simple one dimensional example is instructive. Imagine you are measuring the height of children of a given age bracket, using a single school class as your

sample. This class happens to contain a set of identical twins, who are more or less exactly the same height. These two individuals are dependent data points, in so far as measuring one twin gives you the information about the second twin, and measuring the second twin adds no new data to your data set. If these twins are of average height, then you may not have any problems. But if they are exceptionally short or tall, they may bias your average. Even if their heights are not extreme, the dependence between data points may affect the data set variability, or higher moments[1].

A mathematical analogy is also useful. We can think of models as a vector in *model information vector space*, where different orthogonal dimensions represent different types of information. If we represent a new model vector as a combination of information taken from other model vectors already in the ensemble, then this model introduces no new information (i.e. no new dimensions that are not already represented in the information space). Indeed, if we want the distribution of information in the information space to be distributed in a similar way to the true climate system, adding a highly dependent model may simply skew our dataset, by shifting the centre of our information set (the same way the twins might shift the mean of the height data set).

Masson and Knutti (2011) define dependence in the context of climate model ensembles thus: an additional model is dependent "if it provides little insight into why and how models differ from each other in the existing ensemble, and from observations." That is, if a new, dependent model run is added to an ensemble, then it adds little or no new information to the ensemble, and probably skews the data. The obvious question here is, can't some information be more important than other information? If the sample is producing a skewed probability density function (PDF), how do we know that this is *not* representative of the underlying CPDF? It is possible that this question is ultimately an unanswerable one.

1.4.1 Types of dependence

Dependence between climate models is not particularly well defined. Any number of components of the modelling process can be considered for dependence. Since a model is, in effect, a complex input-process-output (IPO) model, there are three major groups of dependence: Input dependence, process dependence, and output dependence (see Figure 1.3).

- *Input dependence* refers to dependence in the input data (e.g. initial conditions, boundary conditions, forcings) for the models.

[1]These data-points are not *entirely* dependent, as the twins may have had different environmental conditions during their growth (e.g. different nutritional input). Likewise, the heights of all other children in the class are not entirely *in*dependent, as they have probably grown up in a similar environment, and culture to each other (so the height result might only be applicable within a local context).

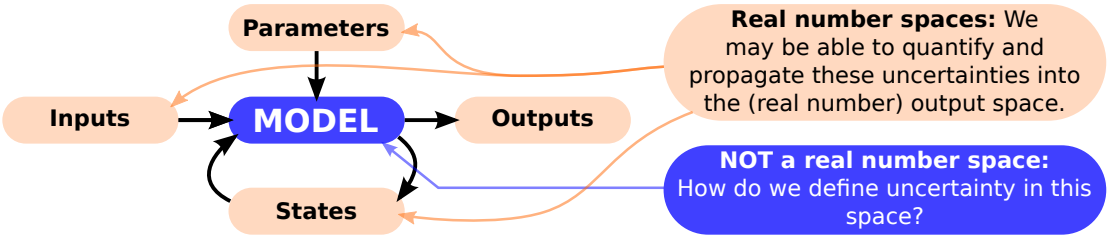

Figure 1.3: Model conceptualisation after Abramowitz (2010)

- *Process dependence* refers to the dependence between numerical representations of physical processes in models, including equations, parameterisations, and time step length and grid scale.
- *Output dependence* refers to statistical similarities between the patterns observed in model outputs.

Both input and output dependence are real-valued spaces, and can, at least to some extent, be compared objectively. Process dependence is decidedly not a real-valued space: how can one accurately measure the distribution of uncertainty between different modelling approaches? Should one attempt to quantify support for various approaches among the scientific community? Even if that were possible, should we expect the distribution of beliefs in the scientific community to usefully represent reality?

Although process dependence may be impossible to quantify, in the real world the structures of the models *are* usually somewhat dependent. Indeed, Masson and Knutti (2011) showed that there exists a "genealogy" of climate models: "Strong similarities are seen between models developed at the same institution, between models sharing versions of the same atmospheric component, and between successive versions of the same model."

Even more difficult to elucidate is the relationship between input, process, and output dependence. Mappings between the spaces may be possible, but due to the complexity and non-linearity of the models, such mappings are unlikely to be simple. As we are primarily interested in obtaining reliable projections from ensembles, we focus solely on output dependence in this thesis, and leave the investigation of the links between dependence types for later.

1.4.2 Independence weighting

Bishop and Abramowitz (2012) introduce the first methodology for explicitly weighting model runs in an ensemble based on model dependence, by considering *error dependence* with reference to an observations data set. Under the truth-plus-error paradigm, the expected value of the correlation between model errors time series is zero, as the "truth" component has been removed: the correlation is between two series made entirely of noise. Under the indistinguishable and replicate earth

paradigms, model errors should actually be considered a linear combination of two time-series (model minus observations), and so expected correlation will be positive (Bishop and Abramowitz, 2012).

The algorithm Bishop and Abramowitz (2012) have developed uses error covariance as a measure of difference. This has the benefit that the error covariance of one model run with itself acts as a measure of performance (i.e. the variance in the error), allowing performance and independence weighting to be traded off against each other. We discuss this approach in more detail in Section 2.4.1)

1.5 The problem: how best to generate ensembles?

Climate modelling is a computationally intensive task, requiring scarce and costly resources to run even relatively small ensembles. How can we maximise the information we get out of our models, while minimising computational resource usage? Which ensemble generation methods provide the best performing, and most independent ensembles? There are three main ensemble generation techniques we will consider.

1.5.1 Initial conditions ensembles

Initial conditions ensembles (ICEs) consist of model runs from a single model with fixed structure and parameterisations, but use variations in the initial state of some or all model variables. These variations are usually constrained by the climatology of the starting period. Initial conditions ensembles are generally used to explore the internal variability of a model, which may be assumed to represent the variability of the Earth's climate, or uncertainty in true values of the system at the start of the model run. ICEs are often small, 3-5 member ensembles, and are commonly used for assessing the variability under certain conditions (Allen and Ingram, 2002). Some of the submissions to the CMIP3 ensembles are ICEs themselves.

Because of the partially chaotic nature of most numerical models, "For time horizons of several decades and longer, any detailed memory of the initial conditions has probably been virtually lost, so that the simulated climate statistics during a given period may be anywhere within the probability distribution determined by the external forcing and the model's internal variability" (Räisänen and Palmer, 2001).

1.5.2 Perturbed physical parameter ensembles

If we are to model the Earth's climate, we must provide appropriate parameters for model equations. Some parameters are directly measurable, and may be coded directly into a model. Others, especially those that form part of heavily approximated equations, must be estimated. For example, the diffusion of heat through the oceans may be approximated as a single global parameter value, or different val-

ues for different regions. This parameter may be used in different ways in different models.

Perturbed physical parameters ensembles (PPEs) are sets of simulations from a single model with fixed structure and initial state, and with one or more physical parameters varied for each run. PPEs allow us to explore uncertainty in physical parameters, and combinations of parameters, usually under the assumption that the model structure and initial state are accurate representations of the Earth's climate system.

The Climate*Prediction*.net ensembles, some of which have tens of thousands of members, consist of a combined PPE and ICE: for each set of parameter values, an initial conditions ensemble is run, and the results are combined into a "grand ensemble" (Stainforth et al., 2005).

1.5.3 Perturbed physical structure ensembles

Because of the shared background of many climate models, structural uncertainty "would be hard to capture by changing parameters within a single model, no matter how wide the range of parameters is chosen" (Tebaldi and Knutti, 2007). So there is value in comparing divergent theoretical understandings that describe different mathematical models of real physical processes.

Perturbed structure ensembles (PSEs) consist of sets of model runs that come from entirely different models, or from a single model with alternative structural components. Initial conditions and parameters between different models are not necessarily comparable.

1.5.4 Ensembles of opportunity

There is no limit to the possible values of initial conditions and parameters, or of potentially viable component structures that could be used to represent aspects of the true climate. Only a small subset of these values and structures will ever be used, because of the constraints imposed by computing power and time. For this reason we must make do with whatever model data are available. Because for such ensembles, models are generally collected rather haphazardly, and as opportunity presents them, they have been termed "ensembles of opportunity" (e.g. Tebaldi and Knutti, 2007; Annan and Hargreaves, 2010).

Most large, international multi-group ensembles, such as the CMIP3 and CMIP5 ensembles, are part PSEs, in the sense that multiple research institutes submit runs from their own models, but many of these submissions contain smaller PPEs and ICEs. There is very little overarching design behind the choice of model components represented in these ensembles: they fall firmly in the category of ensembles of opportunity. This highlights how little thought has been directed to the question of how best to generate model ensembles.

1.6 Aims

Model ensembles such as CMIP3 and CMIP5 underpin our best estimates of future climate. Such projections have major environmental, social and economic ramifications, and, if taken seriously, could alter the course of history, saving millions of species and lives, or costing billions of dollars. It is vital that these projections are as accurate as possible.

In order to achieve the best possible projections, ensembles used to make those projections must be in some sense optimal. However, we know that this is not the case for current projections, such as those based on CMIP3: Annan and Hargreaves (2010) show that the ensemble is over-dispersive for both surface air temperature and sea level pressure, and under-dispersive for precipitation. Bishop and Abramowitz (2012) also show that the ensemble members are not totally independent. The effect of these problems on projections is uncertain, but it is unlikely to be beneficial.

A major goals of the climate modelling community for the foreseeable future must be to resolve these problems. That means finding ways to reliably generate ensembles with the properties needed to make ensemble-based projections consistent and accurate. There are a number of steps required to get to such a point:

- First, we must understand what properties optimal ensembles will have. This process is underway (e.g. Tebaldi and Knutti, 2007; Annan and Hargreaves, 2010; Bishop and Abramowitz, 2012), but much more can still be learned.
- We need to find and develop tools for analysing these properties in ensembles (some are used in Annan and Hargreaves, 2010; Bishop and Abramowitz, 2012).
- We must understand the effect that different ensemble generation approaches have on these properties in the resultant ensembles. This area is particularly unexplored to date.
- We need to develop a better understanding of the effect that different weighting methodologies have on ensemble projections. Work on solving this part of the puzzle is underway, although the focus has been on performance-based weighting, and mostly only on mean estimates (e.g. Weigel et al., 2010; Gleckler et al., 2008).
- We must then use this knowledge to help direct the creation of model ensembles, either by optimally combining existing model simulations, or by actually soliciting specific simulations to ensure that the ensembles have the properties we need.

This is a process that will take years at best, and may potentially stretch into decades. There is no way that even one of these points can be wholly encompassed in a single paper. This thesis offers a stepping stone in this process. Our aims are:

1. To examine the effects of different ensemble generation techniques, using existing ensemble analysis techniques,

2. To examine the effect of different weighting methodologies on the ensembles produced in 1, and
3. To develop new ensemble analysis tools and methodologies, and use them to compare different ensemble interpretation paradigms.

We make no attempt to delve into the creation of optimal ensembles, but hope that the work presented here will provide a decent foundation for future progress in that area.

Chapter 2

Experimental Methodology

2.1 Overview

This chapter describes the experimental methodology used to obtain and analyse the data for this thesis. The experiment involves generating and comparing multiple ensembles of model runs, covering the period 1971-2010. This period was selected to have a reliable high-coverage surface air temperature data set, the UK Met Office Hadley Centre/University of East Anglia Climate Research Unit observations dataset (HadCRUT3), for comparison (described in Brohan et al., 2006).

Mk3L (Phipps, 2011) was chosen because it is a low resolution model, and so runs quickly, which allowed the generation of many simulations within the available time. Even though Mk3L is primarily designed for long time-span modelling, it performs reasonably well even on short time scales, including the 20^{th} Century (Phipps et al., 2012a).

Section 2.2 describes the HadCRUT3 data set. Section 2.3 describes the model setup, and the perturbations used for each of the three ensembles we generated: an ICE, a PPE and a PSE (as described in Section 1.5). Section 2.4 describes the methods used to analyse the individual model runs, and the ensembles as a whole.

2.2 Observational Data sources

HadCRUT3 consists of a combined land surface and sea surface temperature record, from 1850-2012, on a 5 deg $\times$5 deg grid (72×36 grid points). It is a combination of in-situ and satellite-based measurements. As depicted in Figure 2.1, of the 2592 grid points, 820 (31.6%) have data for every month of the period 1971-2010. A further 677 grid points have >80% coverage (57.8% in total), and 608 (23.5%) have less than 20% coverage in the same period. The coverage is better in the low latitudes – most of the data from polar regions are missing, and there is also better coverage in the northern hemisphere than the southern. To ensure reliability of data-model comparisons, all analyses presented here are limited to those grid cells where >80% of months have data coverage.

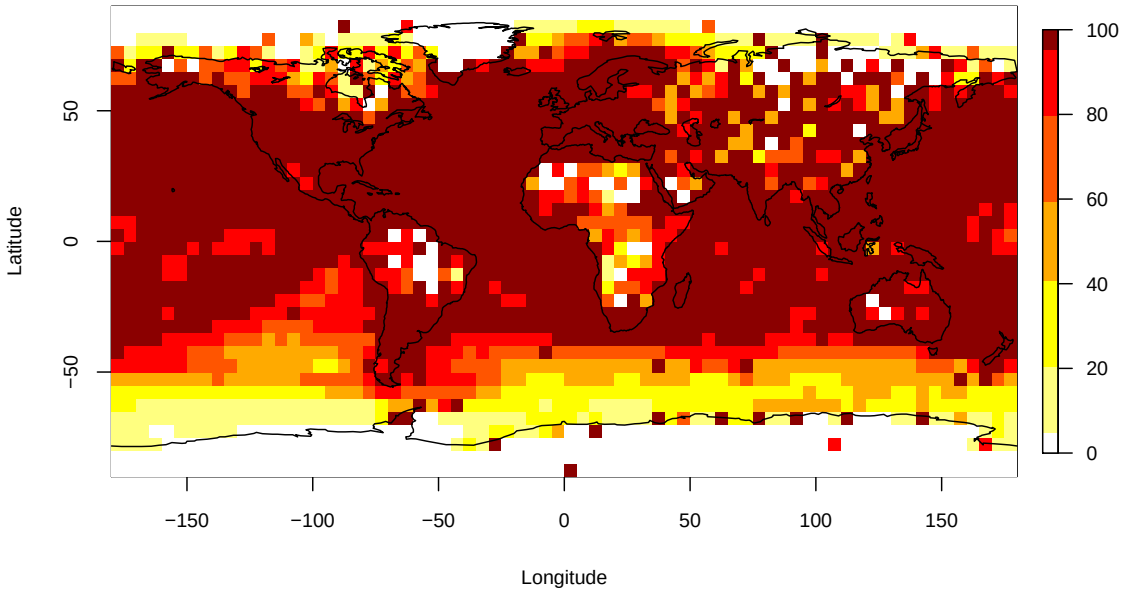

Figure 2.1: HadCRUT3 data availability 1971-2010 (percentage of months that have data per grid point). We used only the data in grid cells coloured red and dark red (>80% data availability).

2.3 Model data generation

The experiment examines model ensembles generated by the Commonwealth Science and Industrial Research Organisation (CSIRO)'s low resolution Mk3L Climate System Model (Mk3L). Three ensembles were generated, each using a different perturbation methodology:

- A perturbed initial conditions ensemble (ICE).
- A perturbed physical parameter ensemble (PPE)
- A perturbed physical structure ensemble (PSE).

2.3.1 Basic Model Set-up

All run ensembles were based on the default Mk3L setup (Phipps, 2011) with each run perturbed from that baseline. The full coupled ocean-atmosphere mode of Mk3L was used. While the atmosphere-only mode of Mk3L would require less computing time, the prescribed sea surface temperatures required for this model mode would likely have significantly mitigated the effect of the model perturbations.

The time span on which to run the model (1851–2010) was based largely on the requirement for a long period of reliable observational data for comparison (1971–2010), plus the requirement of a spin-up period. Two key auxiliary files were extended: atmospheric CO_2-equivalent – a combination of all the major greenhouse

gasses; and total solar irradiance (TSI) – solar irradiance combined with the changes in insolation due to aerosols (based on Schmidt et al., 2012). The CO_2-e file was extended using data from the Goddard Institute for Space Studies (GISS), calculated using the method described in Table 6.2 of the IPCC Third Assesment Report (TAR) Working Group 1 report (Ramaswamy et al., 2001). The TSI file was simply extended using the 2000 value for all following years, as there had been no major volcanic activity over this period. Both of these files are annual inputs, and the values are show in Figure 2.2.

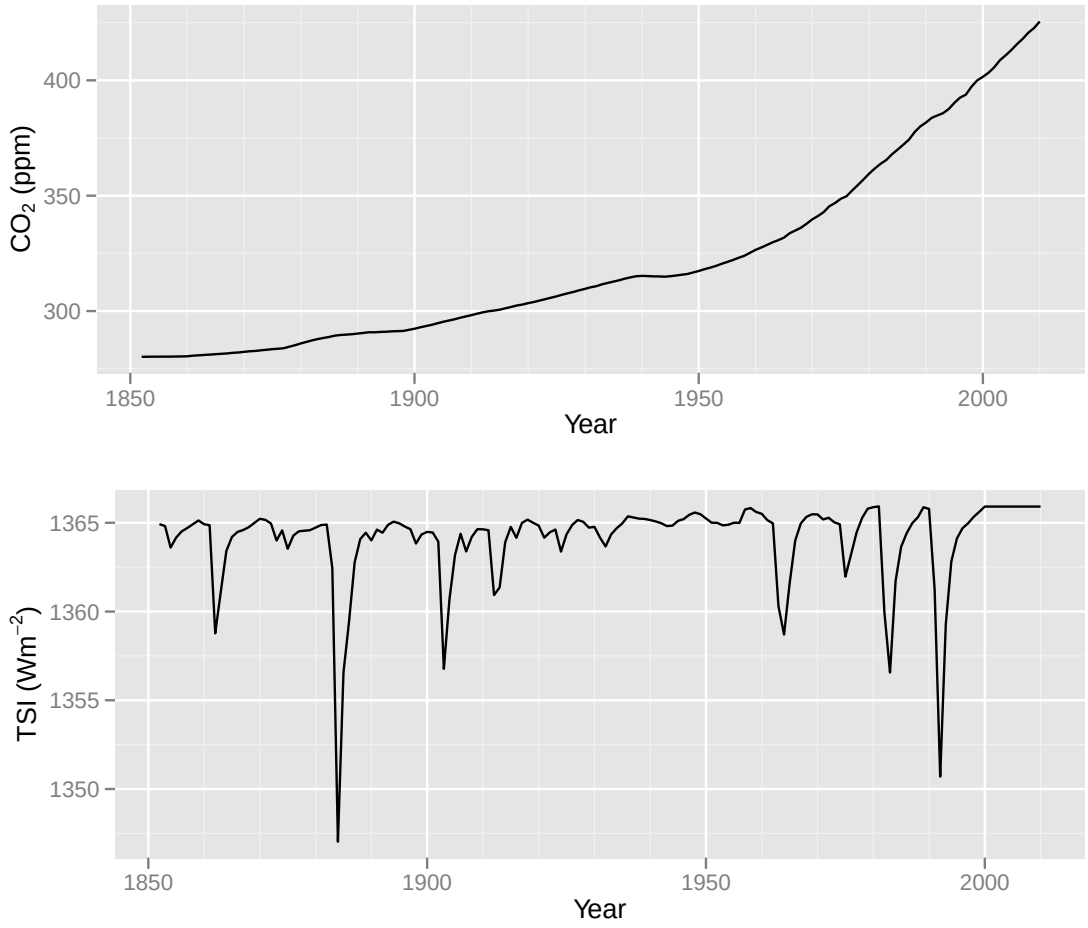

Figure 2.2: CO_2 and TSI values for the simulation period. Note the large drops in TSI in the early 1980s and 1990s, due to volcanic events (see Figure 1.1).

A number of variables were added to the Mk3L control file, as some of the parameters we wanted to perturb were not easily modifiable in the standard model code. These included soil moisture variables, surface roughness, and albedo. For each, we created a scaling factor that was accessible from the model control file. After the modification, the code was tested using various values. Outputs were compared to ensure that they differed (if the variable was not acting, the output

should not change). Some parameter values within the ranges we wished to use caused model test runs to crash. In particular, the model would not run successfully with significantly reduced soil moisture variables, forcing us to abandon running the model at the lower end of our intended range (see Section 2.3.3.1).

The atmospheric component of Mk3L uses a 64×56 grid ($\sim 3.2° \times 5.625°$), while the HadCRUT3 observational data set uses a 72×36 ($5° \times 5°$). The individual model run variable files were re-gridded to match the HadCRUT3 grid, using area weighted averaging.

2.3.2 Initial Conditions Group

For the initial conditions group, the default control file values were used for all runs (Phipps, 2011). Restart files were generated by running the model under 1850 conditions for 10000 model years, and recording restart files every 50 years (Phipps et al., 2012b). This process allows the model state to be recorded while in a climatic equilibrium condition, while also capturing natural variability seen in the model. Restart files with 100 year spacings, starting at year 100, were used as the initial conditions for the group runs. This spacing is long enough to avoid short-term autocorrelation of years close together.

2.3.3 Perturbed Parameters Group

The model parameters modified in this ensemble were chosen because they are known to often significantly affect model output, and as such, should maximise behavioural diversity. Six parameters, or groups of related parameters, were perturbed; including parameters from each of the land, atmosphere, and ocean components of the model. They are:

- Ocean diffusivity.
- Soil field capacity/water holding capacity (varied together with wilting point and saturation point).
- Land surface roughness length.
- Land surface albedo.
- Relative humidity threshold for cloud formation.
- Cloud albedo reduction factor.

2.3.3.1 Justification of parameters and values

Land surface albedo has a direct impact on how much incoming radiation is absorbed by the land surface, resulting in a global land surface net radiation of -2.0Wm^{-2} for a 20% increased albedo, and 2.4Wm^{-2} for a 20% reduction. This results in a ± 0.16K global land surface temperature change, with significantly regional

variation (Fischer et al., 2010). Fischer et al. (2010) used a albedo parameterisation low-high range of $\alpha \pm 20\%$. We used a range of 0.5–1.5 for the scaling factor.

Roughness length is related to the height and variation in height of the canopy, and corresponds to the height at which wind speed is theoretically reduced to zero, due to friction. Roughness length affects surface wind speed, has a negligible effect on surface temperatures, and a small impact on precipitation (Fischer et al., 2010). Murphy et al. (2004) perturbed roughness length over forests only, and use ranges of 0.5–2.0 and 1.05-2.9, depending on the forest type. We used a range of 0.5–1.5 for the scaling factor.

Soil field capacity is the ability of soils to hold water against gravity. It is related to the soil composition, and varies significantly globally (Dunne and Willmott, 1996). Field capacity is bounded above by soil saturation point, the point at which water can no-longer enter the soil, and instead becomes run-off; and below by wilting point, the point at which moisture availability is too low to sustain plants. Soil moisture affects evapotranspiration, which in turn affects cloud formation, latent heat transport, and precipitation (Milly and Dunne, 1994), and has a significant impact on global temperatures (Ducharne and Laval, 2000). We used a range of 0.9–1.5 for the scaling factor.

Ocean diffusivity is a measure of how fast heat diffuses though the oceans. Washington and Meehl (1989) perturbed horizontal ocean diffusivity, using values of 200m^2s^{-1} and 400m^2s^{-1}. The Mk3L default value is 600m^2s^{-1}, and we perturbed this parameter within a range of 400m^2s^{-1}–800m^2s^{-1}.

Critical relative humidity threshold for cloud formation (RH_{crit}) defines the humidity level at which clouds appear in the model. It has been used in large perturbed parameter ensembles, such as Murphy et al. (2004), where a range of 0.6-0.9 was used. We used a range of 0.65–0.75 for RH_{crit} over land, and 0.75–0.95 for RH_{crit} over sea, varied together. This is ± 0.1 around the Mk3L defaults[1].

Cloud albedo reduction factor accounts for the fluffiness of clouds at the subgrid level, and reduces the albedo of the model clouds, which are calculated as a simple plane (Gordon et al., 2002). There are two variables: one for convective cloud, and one for other clouds. We used values in the ranges [0.495, 0.695] for convective cloud, and [0.765, 0.965] for other cloud – ± 0.1 from the recommended values. This range should be expected to produce changes of the order of $\pm 10\text{Wm}^{-2}$ in incoming surface radiation (Phipps, 2011), and have a large impact on global temperature.

2.3.4 Perturbed Structure Group

Ideally, this experiment would involve the comparison of single-model initial conditions and perturbed parameters ensembles with a true multi-model structural-differences ensemble. However, such a process would have large set-up time, and difficulty in comparison between the output of the various models. As a compromise,

[1] RH_{crit} over land was intended to be 0.65–0.85, but a coding mistake was made in the set-up scripts

we ran a structural perturbations ensemble using only Mk3L, but using a number of control file switches to turn components of the model on and off. Although we should clearly not expect as much structural independence between model simulations produced this way as we might with a true multi-model ensemble, we should still expect to see more behavioural diversity than in the perturbed parameters ensemble.

2.3.4.1 Identification of key structural components

We identified a number of switchable Mk3L model components that would be expected to have significant impact on model output. The components are (Phipps, 2011):

- The sea ice model, which has four states: off (prescribed sea ice), basic Semtner sea ice model, Semtner model with leads, and dynamical sea ice (this is actually three separate but interdependent binary switches, each requiring the one before to be on).
- The NCAR boundary layer scheme,
- The New SIB land surface scheme,
- The prognostic cloud scheme,
- The UK Met Office convection scheme,
- The McDougall equation of state, and
- The new gravity wave drag scheme.

The experiment used a number of combinations of each of these options, including the default (all on), and default with each option changed individually (9 runs), and a further 15 runs with pseudo-random sampling of model structure, described in Section 2.3.5.

2.3.5 Sampling strategy

For parameter value selection a low-discrepancy sequence, the Sobol' sequence (Reichert et al., 2002), was used to sample both values from a uniform distribution over the intervals described in Section 2.3.3.1, and combinations of discrete values for the PSE.

Sobol' sequence sampling involves calculating an m-dimensional (here $m = 6$ - the number of parameters) Sobol' sequence, of length n (the number of samples – runs – we want to generate). The elements in the Sobol' sequence are m-dimensional vectors, $\mathbf{s}_n$, with each component a quasi-random value between 0 and 1. We can then take each of these samples, and map each component to the interval required for each model parameter – i.e. we take the first element to correspond to the ocean diffusivity values, which we want to map to [400,800], and we use *ocean diffusivity* $= (800 - 400) \times \mathbf{s}_{n,k} + 400)$.

The Sobol' sequence is constructed in such a way that each dimension is relatively evenly sampled, and that no two dimensions are highly correlated (see Figure 2.3). Low-discrepancy sequence sampling has the advantage over processes such as Latin-hypercube or orthogonal sampling that an arbitrary number of additional samples can be taken without significantly changing the evenness of the sampling.

For the PSE, a discretised version of this process used: Real-valued samples were generated using the Sobol' sequence, in the space $(0,1)^m$, and then each interval was split into p_i intervals, where p_i = number of states in dimension i. The 7-dimensional Sobol' sequences would then be mapped from $(0,1)^7 \rightarrow \{0,1,2,3\} \times \{0,1\}^6$ (4 sea-ice model switch states, 2 for each other structural switch). For example:

$$(0.826,\ 0.231,\ 0.345,\ 0.581,\ 0.918,\ 0.287,\ 0.614) \rightarrow (3,\ 0,\ 0,\ 1,\ 1,\ 0,\ 1)$$

Which is equivalent to (dynamical sea ice; NCAR boundary layer scheme off; New SIB scheme off; prognostic clouds on; UK Met Office convection on; McDougal equation of state off; new gravity wave drag scheme on).

As far as we are aware, no previous papers have been published using this sampling strategy. There are a number of potential problems: in particular, the discretisation may introduce some inter-dimension correlation in samples; the discretisation means that the sequence increases in length, some elements (vectors) will be repeated; and the elements will eventually cover the entire sample space (in this case $4 \times 2^6 = 256$ elements), such that all subsequent vectors will be repetitions.

The first potential problem does not appear to exist; correlation is not appreciably worse in the discretised version: all absolute pair-wise correlations remain lower than 0.3, and correlations change by at most 0.17. The repetition problem inevitably occurs, however this experiment (with sample size initially 25, unlikely to increase above 50) is not affected: the first 127 samples are unique.

A nice property of this sampling method is that discrete and continuous sample spaces can be sampled in the same process. For example, it would be possible to perturb 3 real model parameters and 4 logical switches in the same experiment by using a 7 dimensional Sobol' sequence, and discretising only the last 4 dimensions.

2.3.6 Bias correction

Models run over a specified time frame (e.g. the 20$^{\text{th}}$ century) must go through a "spin-up" phase, which means they are run for a period before the sample period, in order to reach a climatological equilibrium. This spin-up can introduce a systematic bias in results, such that some models, while exhibiting high correlation with observations, are 1-2 K lower than observations (see for example Macadam et al., 2010).

To address this, models are commonly "bias-corrected", which usually means aligning them to a common baseline, such as the average temperature of the first decade of output, or the average of the whole sample period. This kind of bias correction is standard practice (used, for example, in the projections shown in the AR4),

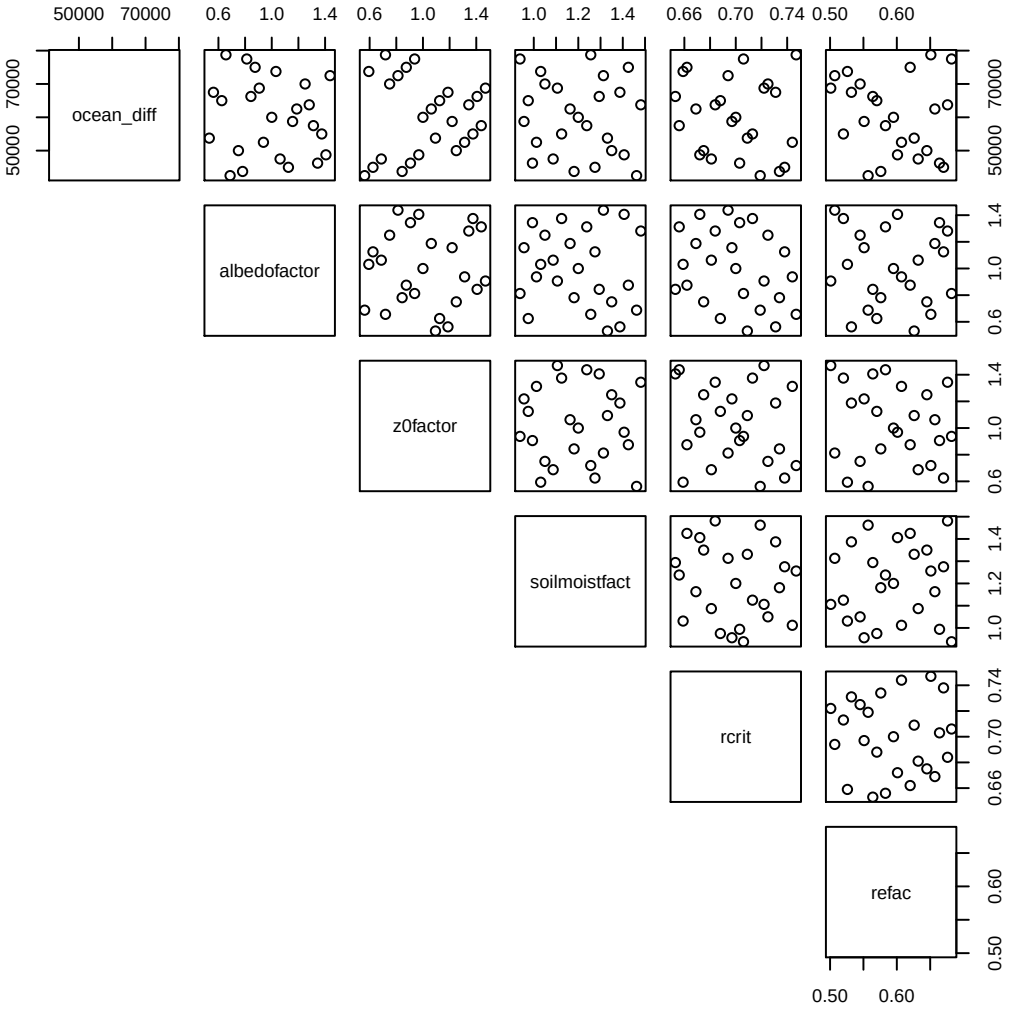

Figure 2.3: Parameter values used for the 25 perturbed physics parameter model runs. The distribution of the samples over the 6-dimensional sample space is relatively even; and pair-wise correlations between variables are low. The units for ocean diffusivity and cloud parameters, which are not important here, correspond to those given in the Mk3L manual (Phipps, 2011).

and consists of subtracting the global annual model mean for the relevant period from the model data, then re-adding the observations mean for the same period. The dependence weighting methodology introduced by Bishop and Abramowitz (2012) relies on bias correction: if the model runs are not bias corrected, then model errors are largely made up of bias, and independence weighting becomes more or less a bias-weighting.

2.4 Analysis

The three ensembles were analysed, using raw output, then using bias corrected output. We investigated ensemble spread by looking at the ensemble as a whole, and by comparing runs within the ensemble. Accuracy of means and ensemble variances were compared under three different weightings: unweighted averaging, performance weighting, and independence weighting (see Section 2.4.1). Dependence within ensembles was investigated using pair-wise comparison of runs.

The three weighting procedures were also applied to the three ensembles using the period 1971-2000, as an in-sample, or training period. The results were used to make projections over the out-of-sample, or testing, period, 2001-2010. Results of these 9 sets of projection approaches were compared.

Note that the experiment was originally designed so multiple variables could be analysed (e.g. temperature, sea level pressure, precipitation), but that due to time constraints, only surface air temperature was analysed.

For the purposes of this experiment, we used cost functions of surface temperature (screen temperature), over the entire globe (excluding areas with missing data – see Section 2.2). The main cost functions we used were covariance, correlation, and root mean square error (RMSE). We ran cost functions over two domains: per-cell data (using all data at each time-step and grid point), and global data (using globally averaged data at each time-step).

2.4.1 Bishop and Abramowitz's methodology

Bishop and Abramowitz (2012) use error-covariance-based weights to weight ensembles. Weights for each model are calculated as the inverse of the sum of the pair-wise covariances between the model and each other model. They show that this creates the optimal linear combination of models, that minimises the distance (mean square error (MSE)) between the ensemble mean and the observations, for a given domain. While this is just one way of defining independence, it is the only way we have currently available of tying model independence to ensemble performance.

Covariance is calculated pair-wise between K model run errors ($\mathbf{x}_k$ = model k -

observations) to give the error covariance matrix (A):

$$A = \begin{pmatrix} \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_K) \\ \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_K) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathbf{x}_K, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_K, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_K, \mathbf{x}_K) \end{pmatrix} \tag{2.1}$$

The matrix A is then inverted, and the column corresponding to model run $\mathbf{x}_k$ is summed, and normalised by dividing through by the sum of the components of the inverted matrix, to give a value w_k for each model:

$$\mathbf{w} = \frac{\mathbf{A}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{A}^{-1}\mathbf{1}} \tag{2.2}$$

The elements of the vector $\mathbf{w}$ act as the coefficients of the linear combination of model runs that minimises the MSE between the models and the observations ($\mathbf{y}$), i.e. for time step $j \in \{1 \dots J\}$:

$$\mu^j = \mathbf{w}^T\mathbf{x}^j = \sum_{k=1}^{K} w_k x_k^j \quad \text{such that} \quad \sum_{j=1}^{J}(\mu^j - y^j)^2 \quad \text{is minimised.}$$

The values w_k are such that $\sum_{k=1}^{K} w_k = 1$, but can be larger than 1, or negative. Unfortunately that means that these $\mathbf{w}_k$ cannot be considered weights for taking an ensemble average, and can not be used to calculate variance. Because the independence coefficients are used to weight model *errors*, there is also no way to use them to weight for projections using just the bias corrected model data. As projections are made over a period for which the observations are unknown (or known, but only used for validation), we cannot calculate model errors for this period.

To overcome this problem Bishop and Abramowitz use a transformation that modifies the ensemble members themselves, such that their variance around the CPDF mean (as estimated by the independence-coefficient based weighted mean) is the same as the observations around the CPDF mean. The transformation is a two-step process first that normalises the independence coefficients $\mathbf{w}_k$ to weights $\tilde{\mathbf{w}}_k$ that are positive and sum to 1, and then inflates or deflates the variance of each ensemble member about the CPDF mean estimate. The resultant elements are linear combinations of the original model runs, rather than model runs. This transformation process does not modify the CPDF mean estimate: the independence coefficient-based linear combination of the original models is the same as the independence-weighted mean of the transformed ensemble members. The new ensemble elements retain high correlation with the original corresponding model runs ($\sim$0.95 with their toy model), but have different variability structures, and cannot be considered as models. A projection variance can then be calculated from these transformed ensemble members.

Bishop and Abramowitz's 2012 weighting can be applied in any number of ways, using various cost functions. In particular, it can be applied on a per-cell basis –

that is, each model run has one weight per grid cell, and the time series for each grid cell from different models are combined using independent weights – or it can be combined globally, using all data to calculate a single weight per simulation. Bishop and Abramowitz used both global and per-cell weighting in their paper. For this thesis we use only global weighting, based on per-cell data.

Weighting model runs for dependence is somewhat analogous to removing model runs from the sample – if two models are dependent, they contain very similar information. As such, one would expect the variance to increase correspondingly, if the weighting is very uneven. There is a sample weighted variance formula:

$$s^2 = \frac{V_1}{V_1^2 - V_2} \sum_{i=1}^{N} w_i \left(x_i - \mu^*\right)^2 \tag{2.3}$$

Where $V_1 = \sum_{i=1}^{n} w_i$ and $V_2 = \sum_{i=1}^{n} w_i^2$. If we use normalised weights, we get $V_1 = 1$, and so equation 2.3 becomes:

$$s^2 = \frac{1}{1 - \sum w_i^2} \sum_{i=1}^{N} w_i \left(x_i - \mu^*\right)^2$$

For very homogeneous weights, with a large sample size, the denominator approaches $1 - 1/N$, and the weighting has very little effect, but when weights are highly heterogeneous, and the sample size is small, the difference between unweighted and weighted variance becomes larger. We use the weighted variance formula for all variance calculations on weighted ensembles.

2.4.2 Other analysis tools

We have used a number of analysis tools uncommon in or absent from the climate literature. We describe two here, but the description of others (pair-wise error correlation and variance validation) are left to Chapter 3, as they are closely intertwined with the results.

Rank histograms compare the observations to the ensembles by calculating the rank of the observations relative to the models at each data point (Hamill, 2001). For example, if the observation value for a particular grid point and time-step is lower than all the models, it will be ranked $(n+1)^{\text{th}}$, where n is the number of models. If the observations and ensemble data are drawn from the same distribution, the distribution of ranks should be approximately uniform, as the observations have an equally likely chance of any given rank, for a given data point. If the ensemble is over-dispersive – meaning that the spread in the model ensemble is greater than that in the observational data set – then the distribution will be higher in the middle, as the observations are less likely to fall in the extreme high or low ranks. On the other hand, if the models are under-dispersive, the distribution will be U-shaped, as the observations fall outside the model range more often. Note that this would not be true with strongly biased data: if the ensemble contained a significant bias,

we would expect to see the observations rank consistently high or low. As we are using bias corrected data, we should expect that the ranks are spread relatively symmetrically over the interval – that is, the integral of the density function over the lower half of the ranks should be approximately equal to that of the upper half.

Quantile-Quantile plots (QQ-plot) provide an alternative representation to the rank histogram, but have not been used to examine climate model ensemble spread before. The QQ-plot compares the distributions of the model and observational data sets by point-plotting the quantiles of the two distributions, with model distribution quantiles as y values, and observation distribution quantiles as x values. If the model and observational datasets are drawn from the same distribution, the points should align on the $y = x$ line. If the model data set (on the y-axis) is over-dispersive relative to the observational data (on the x-axis), the graph will plot a steeper line; if the models are under-dispersive, the slope will be shallower. This analysis is useful in comparison to the rank histogram approach as it uses real values rather than ranks, which can help in the detection of outliers.

As there are more data in the model data set (because there are multiple model runs, with the same spatial and temporal resolution as the observations), the model data set must first be condensed into a single representative data set. This is done by linear approximation over the sorted data set, which acts as a simple down-scaling, giving us a derived data set the same size as the observational data set, but closely approximating the model data set.

Because the data is sorted before the QQ-plot data is calculated, any correlated trends in the data sets tend to over-ride the variability at each cell. There *are* strong correlations between the model and observations data sets, due to the seasonal temperature changes, the polar-tropics temperature gradient, and the over all temporal model trend. To overcome this, the data is first detrended using the per-cell monthly mean of the models (i.e. different detrending for each ensemble). This removes seasonal, spatial, and annual trends, leaving only variability at each data point, which is compared here. This is somewhat analogous, although not equivalent, to ranking, which happens on a per-cell and time-step basis, thus removing any spatial or temporal trends that might be present. Data points to the negative end of the scale are points where the temperature is significantly lower than the average for that grid point and time-step. Data are then sorted by value (regardless of time step, longitude, or latitude), in each data set, and plotted against each other.

Chapter 3

Results

This chapter describes the results of the experiments outlined in Chapter 2. We first analyse the results of the generation groups, and compare different methods of averaging over the three ensembles. We then examine the spread within the ensembles, and look at basic measures of dependence. Finally, we use three weighting methodologies – unweighted mean, performance weighted mean, and the transformation methodology from Bishop and Abramowitz (2012) – to compare projections from the three ensembles.

3.1 Scope of model output

The experiment initially aimed to have 25 model runs for each of the three ensembles (an ICE, a PPE, and aPSE). Of these runs, all of the initial conditions and perturbed parameters runs completed successfully. 20 of the 25 structural perturbations runs completed successfully – the other five all ran for 1-2 years and then failed, most likely due to numerical errors, and could not be recovered using restart scripts.

Of the model simulations that ran successfully, all appeared to display broadly acceptable behaviour: all appeared to track the last 4 decades of global surface air temperature reasonably well, with some drift; none were more than 6K different from the observations globally. All completed runs were therefore included in subsequent analysis, as there was no objective reason to exclude them.

The data from model simulations have been interpolated to the HadCRUT3 grid ($5^\circ \times 5^\circ$), using only grid cells for which at least 80% of the months in the sample period have data (see Section 2.2). All statistics, weights, and graphs are calculated using monthly, per-cell data (see Section 2.4.1).

3.2 Properties of the three ensembles

The models' global annual average temperature is provided in Figure 3.1. Annual means are plotted, because although all calculations are performed on monthly data, the strong seasonal signal makes it difficult see the difference between runs

and means. Each ensemble performs reasonably well, and the ensemble means follow a fairly similar path, likely because they share the same input files, including CO_2 trajectory and TSI: The CO_2 causes the upward trend, while the TSI (which includes aerosol effects) causes the two large dips in the early 1980s and early 1990s, associated with major volcanic events.

There are, however, major differences between the three model ensembles (see Figure 3.1). The mean and standard deviation of the simulation global-time means (taken over all grid cells and time steps) for each ensemble are shown in Table 3.1. The standard deviation between the means in the initial conditions ensemble are much lower than both the perturbed parameters and structural ensembles. The structural ensemble variance is also much larger than that of the perturbed parameters ensemble. This shows that there is an increase in diversity of run behaviour as the model is perturbed in increasingly complex ways – initial conditions provide little diversity, while structural changes provide the most.

Figure 3.1 also highlights the ensemble biases. The initial conditions ensemble mean shows a small positive bias from the observations, and all individual members are warmer than the observations. The perturbed parameters ensemble mean has a larger, negative bias, but it is still less that 1K below the observations, while the structural ensemble mean is closer to the observations. In both of the latter ensembles, the spread of the individual runs is large, and each includes runs that are both positively and very negatively biased.

Table 3.1: Raw run statistics per ensemble.

Ensemble	mean global temp (K)	global temp std. dev (K)
Observations	289.8	NA
Initial conditions	290.2	0.038
Perturbed parameters	288.8	2.29
Perturbed structure	288.6	2.74

The next step is to bias correct on the individual model runs by removing the difference between the global time mean of the run and the observations. This process is standard practice in climate change experiments (e.g. Macadam et al., 2010; Solomon et al., 2007), and is explained and defended in Section 2.3.6.

After bias correction of each model run, the inter-model variance in the perturbed parameters and structure ensembles is greatly reduced (Figure 3.2). This suggests that much of the variance in the un-bias corrected ensembles stems from drift over the 120 years prior to the sample period (1851-1970). However, even after bias correction, there is still more ensemble variance in the perturbed parameter and structure ensembles. Because of the bias corrections, the values equivalent to Table 3.1 would all be identical to the observations. The temporal mean of the standard deviation of the models' global temperature errors at each time step is

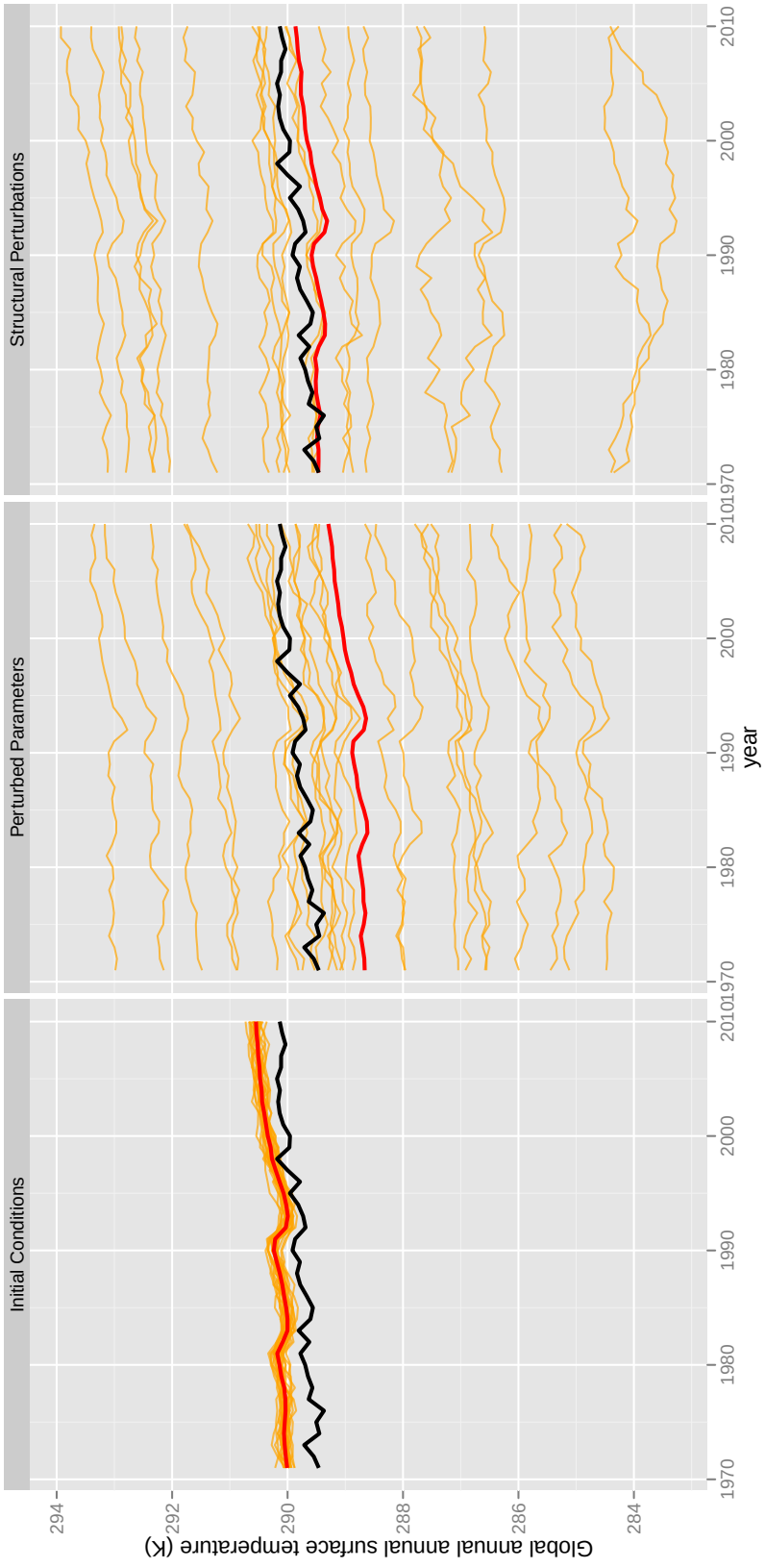

Figure 3.1: Global annual mean surface air temperatures (K) of model runs, grouped by generation method. A common y-axis scale is used to allow comparison. The model runs are in orange, the ensemble mean is in red, and the observations are in black.

given in Table 3.2. This is calculated by taking the global average temperature error (the difference between the run and the observations) for each run at each timestep, (giving a single global error time series per run), then taking the standard deviation across the ensemble at each timestep, and then finally taking the mean and standard deviation over the 480 months. This demonstrates that the structure ensemble has by far the greatest diversity of run behaviour, followed by the perturbed parameters ensemble, with the initial conditions ensemble having very little apparent behavioural diversity. The second column of Table 3.2 show that this variance is also fairly unchanging over time for the ICE and the PPE, but changes a lot over time in the PSE.

Table 3.2: Statistics for the instantaneous standard deviation of global temp error (K) of bias-corrected model data per ensemble (see explanation in text)

Ensemble	Time-mean	Standard deviation
Initial conditions	2.293	0.011
Perturbed parameters	2.815	0.347
Perturbed structure	4.275	1.872

With bias corrected data, we can implement Bishop and Abramowitz's weighting methodology. In the plots of bias corrected models (Figure 3.2), the three means are displayed: the unweighted multi-model mean (red), the performance weighted mean (purple), and the independence weighted mean (green). See Section 2.4.1 for how these are calculated. RMSE values are given in Table 3.3. Note that the weights used here are designed to optimise the independence weighted mean over per-cell monthly data – that is, to minimise the MSE between the model run and the observations across all grid cells – while we are plotting global annual data. The weights *can* be calculated over global annual data, however the resulting mean will almost certainly suffer from over-fitting, as there are 20-25 free variables (weights per model), and only 40 data points.

For the perturbed parameter ensemble, the performance weighted mean shows a slight improvement over the unweighted mean, while the independence weighted mean shows a much larger improvement. For the perturbed structure ensemble, the performance improvement of the two weighted means over the unweighted mean is again large, although the improvement due to the performance weighting relative to the independence weighting is larger than for the perturbed parameters ensemble. For the initial conditions ensemble, all three means are almost identical, and there is very little improvement due to either the performance or independence-weighting.

3.2.1 Model ensemble spread

Spread clearly varies widely across the three ensembles. We now compare ensemble spread relative to the spread in the observations, using two methods: the rank

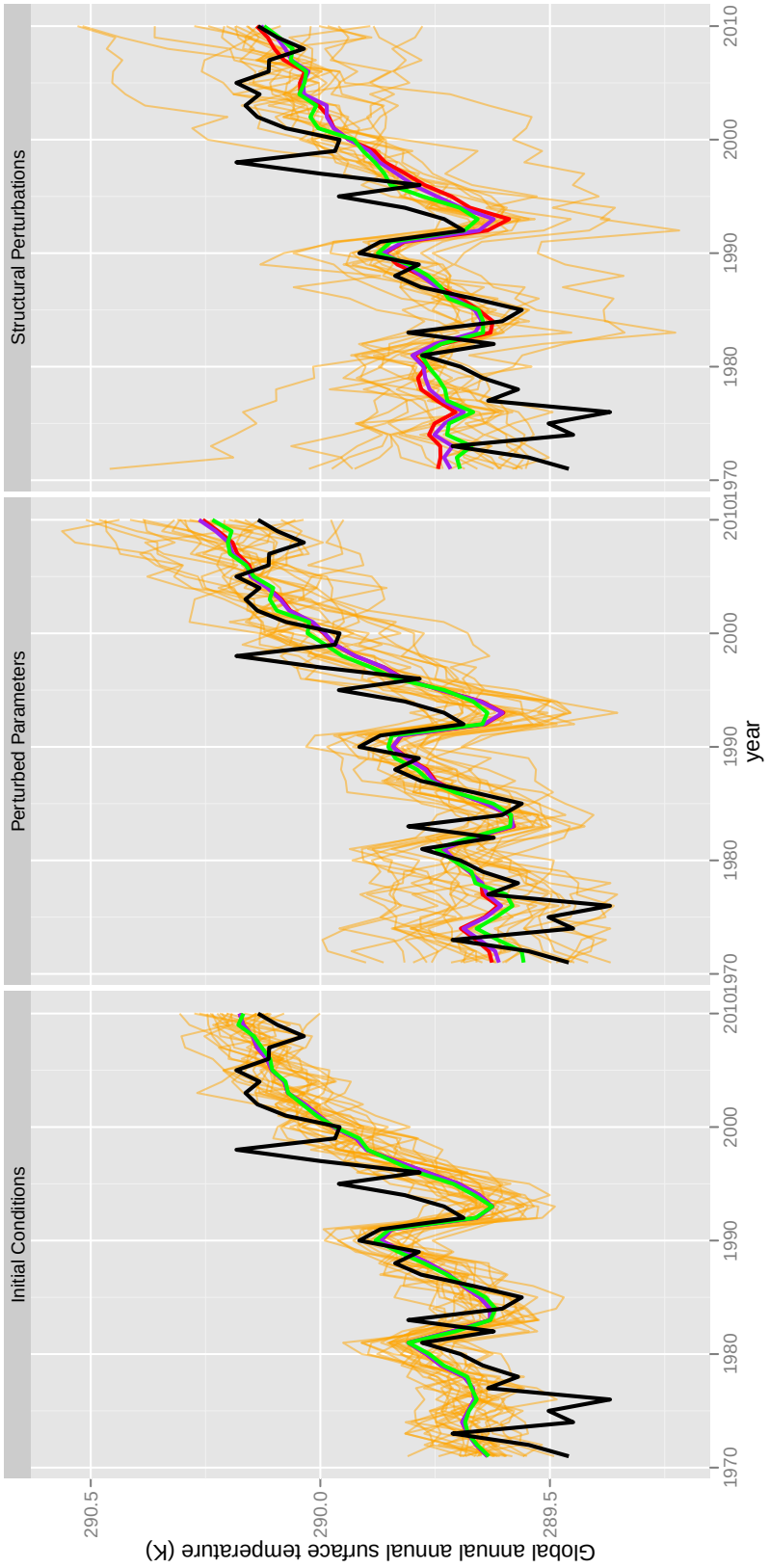

Figure 3.2: Global annual mean surface air temperatures (K) of bias corrected model runs, grouped by generation method. The unweighted mean is in red, the performance weighted mean is in purple, and the independence weighted mean is in green. Bias-corrected models are in orange, and observations are in black.

Table 3.3: RMSE values for each mean per ensemble, relative to the observations.

Ensemble	Unweighted mean	Performance weighted mean	Independence weighted mean
Initial conditions	2.0503	2.0502	2.0469
Perturbed parameters	2.2015	2.1689	2.1212
Perturbed structure	2.5752	2.1546	2.0385

histogram; and the QQ-plot. First, we examine rank histograms for both the per-cell data set (718560 data points), and the global data set (480 data points), then we examine QQ-plots of the per-cell data.

Figure 3.3 shows the rankings of global monthly observations values compared to the models. While there is some noise, it is clear that the initial conditions ensemble is under-dispersive, underestimating the variance in the observations, as indicated by the U-shaped histogram (see Section 2.4.2). The structural ensemble clearly overestimates the variance in the observations, as indicated by the strong bell-shape in the histogram. The perturbed parameter ensemble appears to have an almost flat distribution, indicating that the ensemble is approximating the variance in the observations reasonably well.

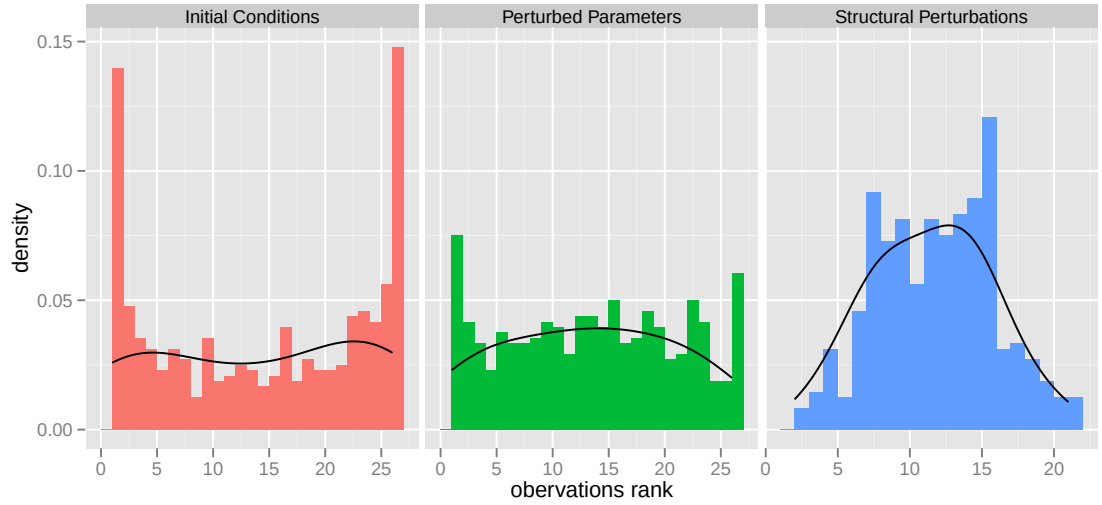

Figure 3.3: Rank of observations relative to bias-corrected model runs in each ensemble, based on global monthly means. The black line is a Gaussian kernel density approximation.

Examining the same graphs calculated per-cell (Figure 3.4), we see a similar, and clearer, result for both the initial conditions and structure ensembles – they are respectively underestimating and over estimating spread. The per-cell rank histogram of the perturbed parameters ensemble tells quite a different story to the global graph: it is now clear that the perturbed parameters ensemble is also

overestimating variance. This highlights the non-transitivity of performance metrics over different domains.

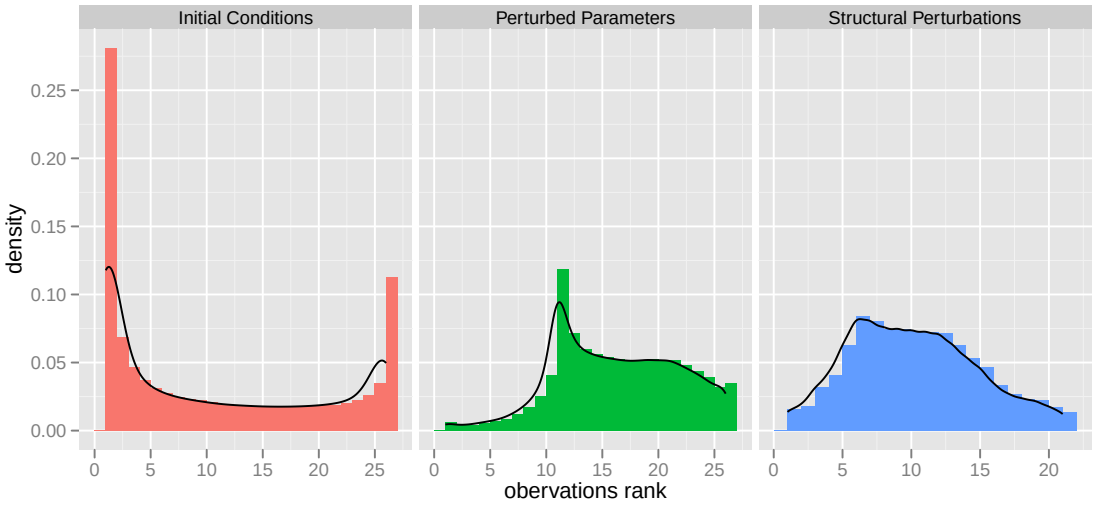

Figure 3.4: Rank of observations relative to bias-corrected model runs in each ensemble, based on monthly per-cell data. The black line is the density approximation as in Figure 3.3.

The large asymmetry in the perturbed parameters ensemble per-cell rank histogram indicates that the observations are ranking higher than the models more often than should be expected, given bias corrected data. This may be because the model means are based on a distribution that is highly skewed relative to the observations – e.g. the models are too hot in the tropics or polar regions. Because the ranks are not dealing with real values, the bias correction does not guarantee a balance of high and low ranks. We now attempt to address the problem by examining the same data using QQ-plots.

3.2.1.1 Quantile-Quantile plots

We can see that the QQ-plots (Figure 3.5) approximately reflect the per-cell rank histogram (Figure 3.4), in particular showing a steep rise toward the centre of the perturbed parameters data set, corresponding to the peak in the associated rank histogram. These plots also clearly show that the initial conditions ensemble is under-dispersive relative to the observations, while the other two ensembles, especially the structural ensemble, are over-dispersive.

3.3 Pair-wise error correlation

We now compare using pair-wise correlations between model run errors as a basic measure of independence within ensembles. Note that we used error covariance

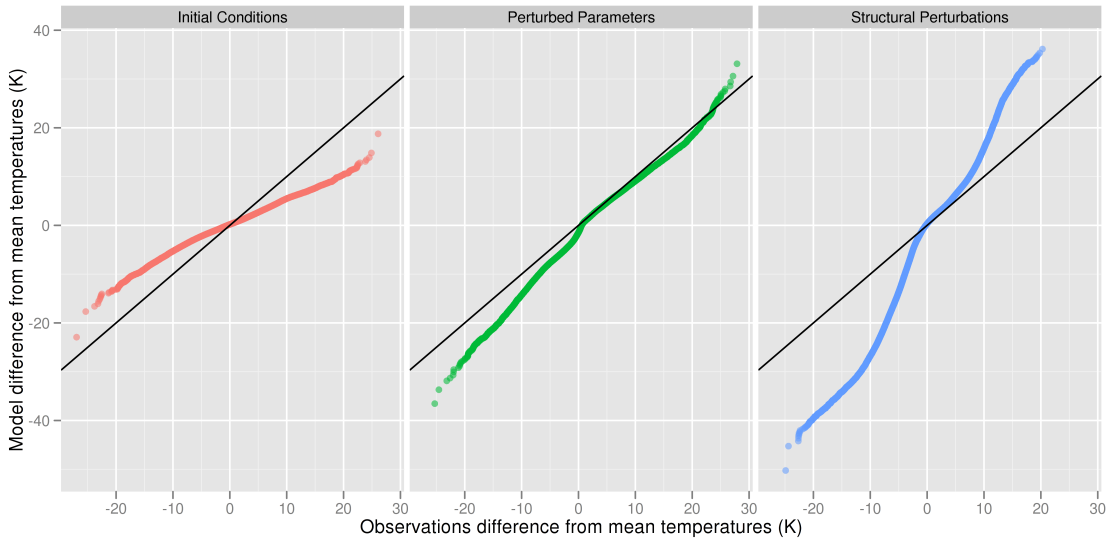

Figure 3.5: QQ-plots of observations relative to bias-corrected model runs in each ensemble, based on de-trended monthly per-cell data (mean of the ensemble removed from both models and observations).

rather than correlation in the independence calculation in Bishop and Abramowitz's methodology. Error correlation is, however, more intuitive than covariance when looking at the causes, and allows direct comparison between model ensembles.

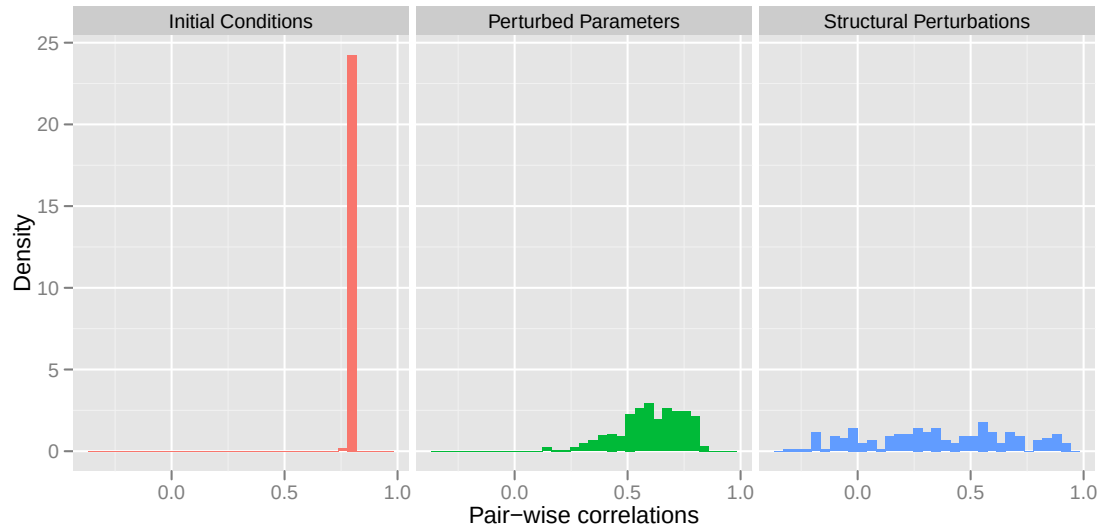

Figure 3.6: Density of pair-wise error correlations between runs in each ensemble. There are 300 pairs in the first two ensembles, and 190 in the structural ensemble.

Figure 3.6 shows the histograms of pair-wise correlations between the runs in each ensemble. There is a clear change in homogeneity between the ensembles, the

initial conditions run pairs all have very similar correlations, while the perturbed parameters ensemble pairs are much broader, and the structural ensemble pairs broader again. The average pair-wise error correlation for the initial conditions ensemble is 0.791, for the perturbed parameters ensemble 0.601, and for the structural ensemble 0.354.

The narrowness of the initial conditions ensemble may be expected from Figure 3.2, where we see that the variance between the runs is small enough to allow the observations to stick out. In particular, in the mid 1970s, the observations are lower than all the models, while in the mid-late 1990s, the opposite is true. This alone would add significantly to the correlations between run errors. There are likely similar patterns in seasonal and spatial trends that we do not see in Figure 3.2. This clearly points to strong dependence between the runs in the initial conditions runs.

Given those factors, the spread of the correlations between structural ensemble runs is somewhat surprising – there are even run pairs with negative correlation. This means that the patterns differ between the runs so much that the variance introduced by the observations is outweighed by the variance between the runs.

3.4 Weighting for climate projections

We now examine the effect that different weighting methodologies have on the three ensembles, considering both mean and variance of the projections. We use the various weighting methodologies, discussed in Chapter 1 and Chapter 2, to produce projections over the final decade of the data (2001-2010), using bias correction and weights based on the first 30 years (1971-2000). We use a simple unweighted mean and unweighted standard deviation first. We then use error variance based performance weighting (using the diagonal elements of the A matrix, after Bishop and Abramowitz, 2012), and weighted standard deviation formula (see Section 2.4.1). Lastly, we generate a projection using Bishop and Abramowitz's (2012) independence transformation methodology, described in Section 2.4.1. The result of these three projections using the data from the three ensembles is shown in Figure 3.7, and RMSE values of the means are shown in Table 3.4.

Because the independence coefficient-based averaging used earlier in this chapter actually takes an average of model *errors*, there is no way to use the original independence coefficients to weight for projection variance using just the bias corrected model data. As we are making projections over 2001-2010, and only using the observations over that period for comparison, we cannot calculate model errors for this period. In order to calculate a variance for the independence projection, Bishop and Abramowitz (2012) use an ensemble transformation, described in Section 2.4.1. A projection variance can then be calculated from these transformed ensemble members.

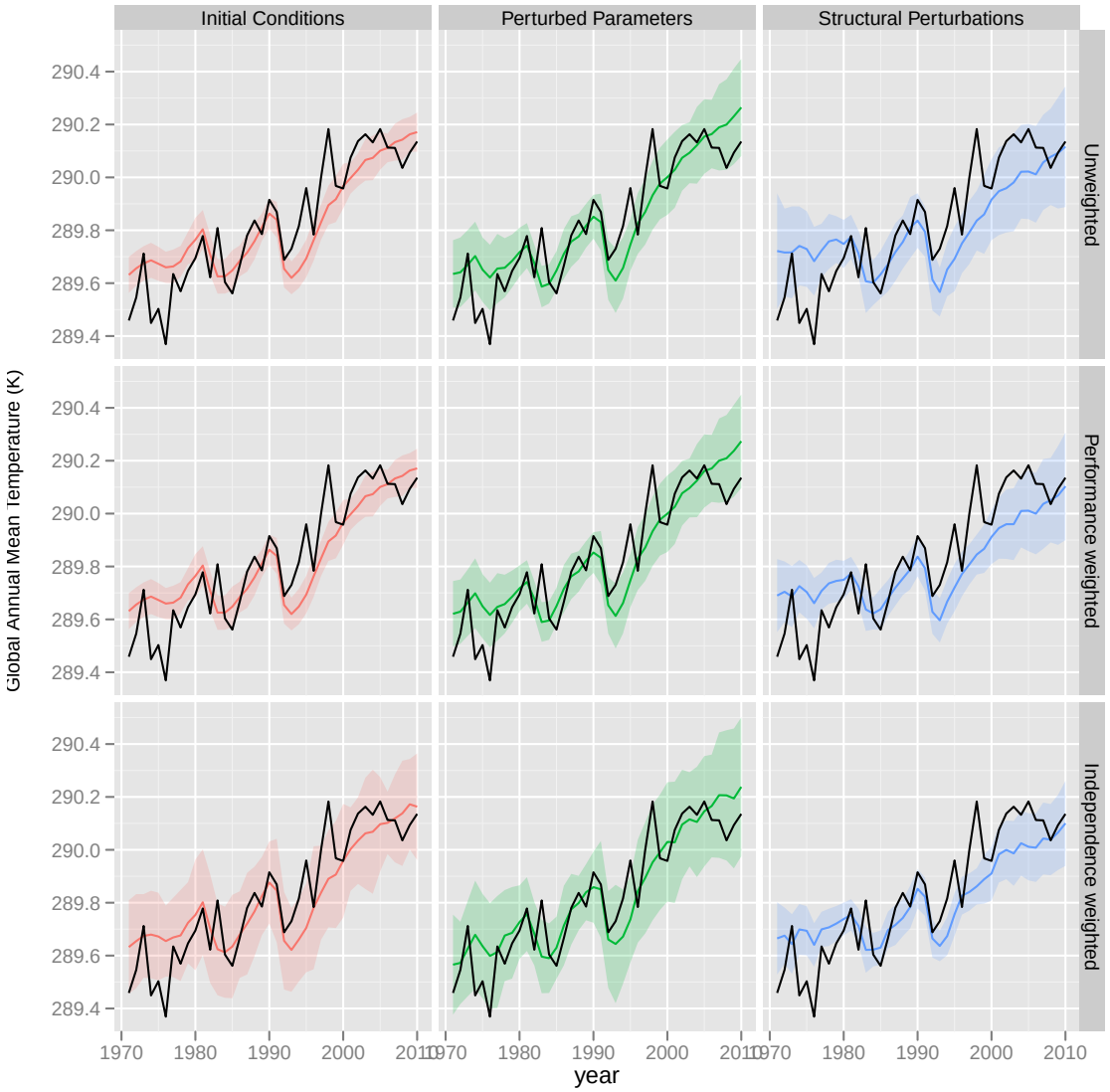

Figure 3.7: Climate estimates based on model data bias corrected and weighted over the period 1971-2000, and projected over 2001-2010. The three columns show the three generation ensembles. The first row is the unweighted mean of the bias corrected models, and their standard deviation. The middle row is the error variance based performance weighted mean and weighted standard deviation. The bottom row is the mean and standard deviation of the transformed ensemble members (*not* the original models, see Section 2.4.1 for details). The observations are shown in black.

3.4.1 Performance of the projection mean

The unweighted projection in the top row of Figure 3.7 shows the baseline estimate of climate projection, as is used in the IPCC AR4 (Solomon et al., 2007). Under this projection, we see that the initial conditions ensemble mean performs best compared to the other ensembles (see Table 3.4). The structural ensemble, which

has the highest variance over the entire period, has the worst performing mean.

Table 3.4: RMSE values for each ensemble mean under each weighting method over the out-of-sample projection period (2001-2010).

Ensemble	Unweighted mean	Performance weighted mean	Independence weighted mean
Initial Conditions	2.054	2.054	2.052
Perturbed Parameters	2.191	2.163	2.122
Perturbed Structure	2.575	2.157	2.039

Under error variance based performance weighting (in the second row of Figure 3.7), the performance of all three projected means improve somewhat, although the improvement in the initial conditions ensemble is minimal. The RMSE of the performance weighted mean projection for the perturbed parameters ensemble is also small, although larger than the initial conditions estimate. The structural perturbations ensemble improvement under performance weighting is substantial – 16% improvement over the unweighted mean.

Under the independence transformation, the structural perturbations ensemble again improves significantly – this time by about 20%. The perturbed parameters projection performance also improves, and is markedly better than under the performance-weighted mean. The initial conditions ensemble projection only improves marginally under the independence transformation.

3.4.2 Performance of the projection variance

We can also examine the spread of the projected ensembles by looking at the number of observations that fall within the projected variance range about the mean, and comparing that to the expected value. The percentage of observation data points that fall within one standard deviation of the projected ensemble mean are given in Table 3.5.

In a normal distribution, the value would be expected to approach 68.3%, but the distribution of surface temperatures in the observations data set is far from normal – the distribution has a long tail toward cooler temperatures. Over the sample period (the first 30 years), 84.14% of the observations data points fall within the range of ± 1 standard deviation of the global time mean. The same calculation on individual model runs gives a similar figure. However, this is not directly comparable to the percentages given in the table, as this value is calculated against a scalar observations mean, rather than a spatially and temporally varying ensemble mean. Under the replicate earth paradigm, we can assume that the independence weighted mean provides a good estimate of the CPDF mean, and use these to calculate the variance of the observations about the mean. Doing so gives values of 82.21%, 80.97%, and 81.45% for the three ensembles. The true value is likely slightly lower, as the

independence weighted means are only an approximation of the CPDF mean, and the true CPDF mean would be expected to reduce the MSE from the observations slightly.

From Table 3.5, we see that the improvement in the projected variance under the performance weighting is far less consistent than the improvement to the projected mean: both the initial conditions and perturbed parameters ensembles variance estimates actually degrade under the performance weighting. Since we don't know the variance of the observations around the true CPDF mean, it is difficult to say whether the structural ensemble variance is improved or degraded under the performance weighted mean. It is possible that performance weighting improves variance estimates for ensembles that are very over-dispersive (such as our structural ensemble), however this data does not provide compelling evidence. It is also important to note here that this only applies to model-observations covariance-based performance weights. It is possible that, under other cost functions, performance-weighted projection variance does not perform so poorly.

Table 3.5: Percentage of observations that fall within one standard deviation of the projected mean.

Ensemble	Unweighted mean	Performance weighted mean	Independence weighted mean
Initial Conditions	37.407	37.404	75.891
Perturbed Parameters	53.047	52.324	76.187
Perturbed Structure	81.703	76.997	77.101

Under the independence transformation, all variance projections improve dramatically. The variance estimate for the initial conditions ensemble is still the lowest, but the range of values is only just larger than 1%. This strong convergence may suggest that the percentage of observations that fall within one standard deviation of the true CPDF mean is around these values.

3.5 Summary

This chapter has shown that the three generation methodologies result in vastly different ensembles. The initial conditions ensemble provides the least variability, and as a consequence may underestimate the variability in the climate when used to make projections. The perturbed parameters ensemble provides more variability, and the structural ensemble provide more again – to the point that they can overestimate the variability in the observations.

However, these ensembles' projections can be rectified to maximise their predictive power by using a weighting methodology. Under all ensemble generation methods, the performance weighting improves the mean projection performance, working especially well when there is high heterogeneity in the individual models'

performances. The independence weighting also improves the mean projection performance when there is high heterogeneity in the weights (i.e. varying dependence between models in the ensemble), and more so than under performance weighting.

Projected variance estimates are heavily dependent on ensemble generation method, and reflect the variance estimates in the ensembles themselves – initial conditions ensembles are too low, while perturbed parameters and structural ensembles are too high. It appears that performance weighting does not improve projected variance, although there may be some cases where it does. Independence weighting, on the other hand, can maximise the accuracy of projected variance estimates for all ensembles.

Chapter 4

Discussion and Conclusions

The aims of this thesis were:

1. To examine the effects of different ensemble generation techniques, using existing ensemble analysis techniques,
2. To examine the effect of different weighting methodologies on the ensembles produced in 1, and
3. To develop new ensemble analysis tools and methodologies, and use them to compare different ensemble interpretation paradigms.

This chapter examines the results with reference to these aims, and the ramifications for the field of climate modelling, before presenting some conclusions.

4.1 Properties of the ensembles

The three different ensemble generation techniques explored produce vastly different ensembles. In particular, the ICE was under-dispersive for surface temperature, while the PPE and PSE were both over-dispersive. Time constraints prevented the testing of other variables (such as sea level pressure or precipitation), but we should not necessarily expect these variables to behave in the same way as surface temperature (see Annan and Hargreaves, 2010).

4.1.1 Are the ensembles representative of the CPDF?

We have only one true sample of the CPDF – the underlying distribution of probable earth states over time, given known boundary conditions – with which to estimate what the entire CPDF looks like. Hence it is difficult to state categorically whether any of these ensembles are truly representative of the CPDF. We can not know for certain that our observations are not a stark outlier – that most other replicate earths would not be quite different. If they are, then our CPDF estimates will clearly be biased.

Statistically, it would be highly unlikely for 718560 data points to be outliers. However, the observation data points are not independent: neighbouring data points are strongly related, both spatially and temporally. If we make the assumption that the observations are not heavily biased, we still have to contend with the question of whether the model ensemble adequately represents the CPDF.

4.1.2 Can these results be generalised?

It is unclear whether or not these results can be generalised to other models. In particular, other models may provide more or less variability under initial condition or physical parameter perturbations. It seems highly likely that a true perturbed structure ensemble, using diverse models, would also have higher variability.

Barnett (1999) shows that inter-model variability is higher than intra-model variability, by comparing 11 of the first CMIP ensemble model runs. Barnett also shows that variability operates on different scales for different models, however, they make no comparison of the differences in general intra-model variability between models. It is possible that other models exhibit greater variability under initial conditions perturbation, such that ICEs from these models have a broad enough spread and *can* adequately represent the CPDF variance. Mk3l is low resolution and relatively simple compared to many other modern AOGCMs. However, it is not clear that a model that is more complex or has a higher resolution should necessarily exhibit greater behavioural diversity under a fixed physics parameterisation.

It is unlikely that the results for our PPE would be able to be generalised to other PPE experiments – the combination of parameters is too specific. Nevertheless, the methods used in this thesis can be applied to a perturbed parameter experiment, in order to get an indication of the fit of the ensemble spread. This would be especially valuable for large PPE experiments used for making projections, such as the Climate*Prediction*.net experiments. PPEs are also often used for ascertaining a reasonable default value or range for parameterisations: in such experiments, these tools would also be particularly useful, as they can allow estimates of parameter ranges to be qualified, or re-assessed.

It is almost certain that the PSE presented here is not representative of a true multi-model structural ensemble. We have perturbed the structure of only a small sub-set of the Mk3L model components. While it is true that some modern AOGCMs share some components (the theoretical background, if not the numerical implementation), these models have many more components than the seven that we have perturbed here. None of the Mk3L model switches were irrational choices – the alternative schemes have all be used successfully in models in the past (for example, the Mk3L predecessors). This suggests that true multi-model structural ensembles are likely to have far broader spread than our PSE does. Bishop and Abramowitz's (2012) transformation methodology allows us to recalculate variance estimates based on independence, and may be particularly useful here.

4.1.3 Ramifications

If we assume that these results *are*, at least to some degree, able to be generalised, then this has some important ramifications. Firstly, it is clear that ICEs are not behaviourally diverse enough – that is, too interdependent with respect to the CPDF – to provide an adequate estimate of spread. This is problematic because the most politically important projections – those presented in the reports of the IPCC – are based on ensembles that include smaller ICEs. However, it seems likely that any reduced variance due to these sub-ensembles is off-set by the fact that these ensembles are also partly PSEs, which tend to have far too broad a spread.

Because of this large spread exhibited by PSEs (and, to a lesser extent to PPEs, which are also used for projection), we must find ways of combining multi-model ensembles that account for this. This leads us to performance and independence-based ensemble weighting and transformation as a potential solution.

4.2 Impacts of weighting methodologies

As well as the large differences in estimates and projections for different ensemble generation methods, there were also significant differences in projection accuracy under different weighting regimes. Relative to the unweighted mean, performance weighting improved the projection mean, but only slightly for ensembles where there was less behavioural diversity. The improvement was most marked in the PSE, where some ensemble members performed quite poorly. But performance weighting drastically weakened the projected variance, for both the ICE and the PPE.

In contrast projections means and variance both improved under the independence transformation, and projections were far more consistent from each ensemble. The convergence of projections under independence transformation (Bishop and Abramowitz, 2012) seems almost too good to be true. It begs the question how poorly performing, or how dependent, would models in an ensemble have to be before the independence transformation could not produce an accurate projection? It is unclear how much of this effect is confirmation of the validity of the model itself, or whether there is some kind of over-fitting occurring.

We did perform some preliminary tests of the validity of the independence transformation over different periods, using a model similar to the toy model used by Bishop and Abramowitz (2012). These tests suggested that the independence weighting only produced useful projections if the variability seen in the in-sample period was related to the variability in the testing period. While this provides a tentative suggestion that these results are *not* largely due to over-fitting, there is certainly room for more exploration here.

4.2.1 Problems with performance weighting

Performance weighting is in widespread use, and is an active area in model combination research, yet almost all weighting procedures consider only the weighted mean, and ignore the effect on variance (e.g. Giorgi, 2005; Krishnamurti et al., 2000). Recent reviews of model combination methodologies and issues related to weighting also appear to have largely ignored the application of any kind of weighting to projection variance (Weigel et al., 2010; Knutti et al., 2010b). We have shown fairly conclusively in this thesis that this can be very problematic: performance weighing can drastically reduce the quality of projections of variance. As a consequence, performance-weighted projections may severely underestimate the variability we should expect to see in the future climate. For small ensembles, this would likely only be compounded by using naive variance calculations in the place of appropriate weighted variance calculations, as explained in Section 2.4.1.

As we only considered error variance based weighting, it is theoretically possible that other cost functions do no suffer as severely from this problem. However, the very nature of performance weighting reduces the impact of the more extreme samples, and in doing so, likely reduces the variance of projections.

4.2.2 Effect of the CPDF mean estimate on independence-transformed projections

In the replicate earth transformation, the distance of the CPDF mean estimate from the observations is critical in determining how broad the CPDF variance estimate should be. The CPDF variance estimate is just the time-averaged variance between the observations and the CPDF mean estimate. This means that if the CPDF mean tracks the observations very closely, the difference between the mean and the observations, and hence the CPDF variance, will be very small. If, on the other hand, the CPDF mean estimate is very smooth, the natural variability in the observations will ensure that the CPDF variance estimate is higher.

The obvious implication here is that an ensemble of simulations from models which have over-fitted parameterisations will likely significantly underestimate the variance. On the other hand, an ensemble of poorly performing models is likely to overestimate variance. It is also worth noting here that the CPDF variance is, to some degree, a function of the number of models. The more models that are added to the ensemble, the more tightly the CPDF mean estimate can be fitted to the observations. However, this is not likely to be a problem except for very large ensembles, with a very short period, or low resolution. In our ensembles, we have at most 25 models, and a corresponding number of free variables (the model weights), with which to fit hundreds of thousands of data points, so over-fitting is not likely to be a problem for these results.

The CPDF mean, ultimately, represents the mean response to all large-scale forcings that all replicate earths would share. So the CPDF mean should respond to, for example, changes in CO_2, solar forcings, and volcanic and anthropogenic aerosols,

which are shared input to models. It should not respond directly to chaotic fluctuations in internal model processes, such as El-Niño Southern Oscillation (ENSO) cycles, or large ocean and atmospheric eddies. It *should* capture changes in the patterns of those chaotic fluctuations, for example, a state-shift that shuts down the North Atlantic thermohaline circulation, or a shift to a permanent El Niño state, if those changes are inevitable with the given boundary conditions. The difficulty then lies in determining which changes are important, affecting every replicate earth, and which are replicate-specific.

4.2.3 Comparison of paradigms

The results shown in Section 3.3 highlight a striking difference between the truth plus error and replicate earth paradigms (introduced in Section 1.2). Under the truth plus error paradigm we expect the observations to be the centre of the model distribution, with extra random noise in the models. With this understanding, we should expect the model error to be randomly distributed around the observations, and hence expect a set of independent models to have a mean error correlation of 0.

In contrast, under the indistinguishable paradigm we expect that the observations are similar to model runs, as both are drawn from the same distribution. Under the replicate earth paradigm, the same is true, but only if the model runs adequately represent replicate earths. In both cases, "model errors" are actually a linear combination of two samples (the model, minus the observations), and some of the variance between a pair of model errors is contributed by the observations. Thus, if the models and observations are independently drawn from the same distribution, as in the indistinguishable paradigm, or are true replicate earths, then the expected correlation between error pairs is actually 0.5 (Bishop and Abramowitz, 2012).

Under the replicate earth paradigm, if the models and observations are *not* drawn from the same distribution – for example, if the models' distribution has less variance – then the observations contribute more variance to the linear combination, and we should expect higher error correlations. Likewise, if the models' spread is higher than the observations, we should expect error correlations to be lower.

Figure 3.6 shows that the truth plus error paradigm would be hard to justify with any of these ensembles. The confidence interval (CI) interval for the ICE and PPE both exclude 0. The CIs intervals of the initial conditions ensemble (0.782, 0.80) also excludes 0.5, the value expected by the indistinguishable paradigm. The CI for the PSE is far broader, partly due to the sample size, with a CI of (-0.2699, 0.9783), and includes both 0 and 0.5. The results for the ICE would be hard to support even under the indistinguishable paradigm, where we would *always* expect 0.5 error correlation. But under the replicate earth paradigm, if models are *not* replicate earth-like, we actually expect the error correlations to vary depending on the variance in the observations relative to the variance in the models, and due

to the very small spread of the initial conditions ensemble, a mean pair-wise error correlation of 0.79 seems entirely understandable.

4.3 Tools used, and potential new directions

We have used and developed a number of ensemble analysis tools which have rarely, or never, been used before in the field of climate modelling. Rank histograms were introduced into the field by Anderson (1996), but have rarely been used until recently (Annan and Hargreaves, 2010; Bishop and Abramowitz, 2012). As far as we are aware, QQ-plots have never been used for this purpose. The pair-wise error correlation histograms that we use in Section 3.3 have been alluded to in Annan and Hargreaves (2010); Bishop and Abramowitz (2012), but have never before been explicitly produced.

4.3.1 Rank Histograms

Rank histograms are useful tools for model verification. The methodology provides a means to compare any ensemble to the observations to determine whether spread is close the CPDF variance. This methodology can be applied to any model ensemble. It would be possible, for example, to generate standardised ICEs for any number of models, using default parameterisations, and then to compare the internal variability of different models by comparing rank histograms. Unfortunately, different models do not always contain comparable numerical schemes and parameterisations, so this can not be used for simple comparison between different models for PPE experiments. It could, however, be used to compare variability of a model with different implementations of a particular component (e.g. the same atmosphere and ocean model, with two different land surface schemes).

But there are a number of drawbacks to the rank histogram that may apply in some situations. The resolution of the histogram is specified by the number of models plus one, so it can be difficult to get a good understanding of the differences in spread and distribution between models and observations for small ensembles. Conversely, for large ensembles over short sample periods, the number of bins is high, and there can be considerable noise in the histogram. Another problem is that because the histogram is calculated on ranks, there is no possibility of catching outliers – if the observations are a long way from the models, the rank will simply be 1, or n+1, where n is the number of models.

4.3.2 QQ-plots

We used the QQ-plot as a means of overcoming these problems: the QQ-plot uses real values, and so we potentially have the ability to spot large-scale temperature anomalies. The resolution is also independent of the number of models, and is

restricted only by the number of data points in the observational data set. The QQ-plot can also be used for inter-model comparison, and using something as simple as a linear regression on the QQ-plot data would allow us to numerically quantify over- or under-dispersion.

However, our QQ-plot methodology is not without problems. Firstly, the problem of de-trending: if we simply use the raw bias corrected model data in comparison with the observations, we are comparing data sets that contain large, correlated trends. These trends contain more variability than the natural grid-scale variability, and, because of the high correlation, overpower the QQ-plot trend line. We used a simple de-trending process involving subtracting multi-model mean from all the models and the observations. Unfortunately this introduces some ambiguity: we are now looking at the comparison of the distributions of differences between the model and observations temperatures from the mean. This is certainly de-trended, but meaning is quite hard to extract from this graph.

A potential solution lies in a more refined de-trending processes. It should be possible to use annual, seasonal, and zonal de-trending; averaging over all models or all models and observations, before computing the QQ-plot. This would allow a simpler interpretation of the QQ-plot trend: the data plotted is the difference from major trends. It may also be possible to de-trend based on some, but not all major trends. For example, the data could be de-trended using only the annual and seasonal trends, in order to compare the distribution of all spatial data, including the zonal trend. This would allow a comparison of the zonal trend. The QQ-plot is certainly an under-utilised tool in climate model verification, and there is large scope for development here.

4.3.3 Error correlation histograms

Error correlation histograms are a very simple tool, yet give a good initial indication of independence of the models in the ensemble, especially when combined with rank histograms or QQ-plots. The exact interpretation of these diagrams is highly dependant on ensemble interpretation paradigm, but, as we argue in Section 4.2.3, the results provide a strong argument for rejecting the truth plus error paradigm, and possibly the indistinguishable paradigm as well. If we adopt the replicate earth paradigm, these diagrams provide very useful information, giving an indication of the overall model variability relative to the observations. The ease with which these diagrams are generated allow us to recommend them highly for future work on ensemble analysis.

4.3.4 Comparison of observations to ensemble spread

In the last part of our results (Section 3.4.2), we present a method of comparing observations to projections, by comparing how many of the observations fall into the domain defined by the mean and the variance of the projection. We are not aware of this having been done before. There is certainly some improvement that

could be made to this methodology, for instance, the expected value for a perfect ensemble is not known. However, this tool is useful even in its present form, and could be used for verification and validation of previous projections. It could also be used to guide future projections: if used in a similar way to this thesis, to create a pseudo-projection over the last decade or so of our in-sample period, it may be possible to get an estimate of the predictive power of projections over future periods.

4.4 Conclusions and Future Work

The results presented in this thesis form another stepping stone on the path to understanding the impact of dependence on climate models. We have focussed on three main areas: The effect of different ensemble generation methods, the impact of different weighting methodologies on ensemble-based projections, and the development of new ensemble analysis methods and tools.

We have shown that there are stark differences between ensembles generated in different ways: initial conditions ensembles tend to have too narrow spread, and hence tend to underestimate the variability in projections. There is no strong reason to believe that this would be significantly different models other than Mk3L. In our results, the perturbed parameter ensemble over estimated that variance, however, this is likely highly dependant on the parameters perturbed, and the scale of those perturbations, and is difficult to generalise to other perturbed parameters ensembles. Our perturbed structure ensemble also exhibited a spread that was far too broad. This is likely to be even more exaggerated in a true multi-model ensemble, where there would usually be more diversity of model components within the ensemble.

These results have quite important ramifications for the field of climate modelling. In particular, it should be expected that while mean projections from these ensembles may be comparable, the range of uncertainty in the projections, as represented by ensemble spread, will be vastly different depending on what ensemble generation technique is used. How this interacts with ensembles of opportunity (such as the CMIP ensembles), or grand ensembles (mixed initial conditions and perturbed parameter ensembles, such as the Climate*Prediction*.net experiments) is uncertain, but certainly deserves further investigation.

We have also shown that different weighting methodologies have large effects on projections based on these ensembles. In particular, although mean projections will improve under performance-based weighting, we have shown that projection variance may not, and may in fact dramatically worsen. Considering there is significant effort being put into research in this area, these results have potentially far-reaching consequences. There remains the possibility that this result is particular to error variance-based results, and may be quite different for projections based on other cost functions. Without evidence to support that conjecture, however, it seems unwise to assume so without further exploration, and care should be taken when using performance based weighting for climate prediction.

Our results show that independence-based weighting does not suffer from this

problem, and that both the projection mean and variance improve notably. The improvement to the projection mean is notably better than that under performance weighting, and the improvement to the projection variance is both better, and far more consistent. This indicates that independence weighting could provide large gains in projection accuracy, which could be hugely beneficial to society in general, by reducing uncertainty around actions needed to avert the worst of climate change.

Because independence weighting is such a new area, there are certain to be problems that we have not managed to uncover in this thesis. More work needs to be done in testing this process. In particular, it would be useful to conduct similar experimentation using different cost functions as the basis for the independence measure. It would also be useful to apply the methodology to existing ensembles, and compare the results to other weighting methodologies, as this may uncover problems not apparent in the single-model ensembles that we have produced. Lastly, it would be worth doing more in-depth study using toy models in order to classify the conditions that produce pathological behaviour in the independence transformation process, and to try to clarify if those pathologies might apply to real climate data and models.

The ensemble analysis tools and methods that we have developed in this thesis are fairly rough, and more development and testing is needed. QQ-plots in particular, while crudely used in this thesis, show potential to produce some quite nuanced results. All of these tools provide good indicative results of ensemble performance, and are relatively easy to apply. There appears to be no reason why their use shouldn't become more widespread.

The ultimate goal that this thesis begins to pave the way for is the ability to design optimal ensembles, with the properties required, such as model independence, and appropriate spread to produce the most accurate projections possible. There is still significant work that needs to be done, much of which probably is unforseeable at this stage. This goal is not going to be reached tomorrow. However, it is on the horizon, and we hope that the work presented here has brought it closer.

Abbreviations

AOGCM coupled atmosphere-ocean general circulation model. iii, 8, 44

AR3 3rd Assesment Report. 19

AR4 4th Assesment Report. 3, 6, 12, 25, 40

ARIMA Auto-Regressive Integrated Moving Average. 2

CCRC Climate Change Research Centre. v

CMIP3 Coupled Model Intercomparison Project Phase 3. iii, 3, 7, 13, 14

CMIP5 Coupled Model Intercomparison Project Phase 5. 3, 7, 13

CPDF climate probability density function. vii, 4–7, 10, 41, 43–48

CSIRO Commonwealth Science and Industrial Research Organisation. 18

ENSO El-Niño Southern Oscillation. 46

GCM general circulation model. 1

GISS Goddard Institute for Space Studies. 19

HadCRUT3 UK Met Office Hadley Centre/University of East Anglia Climate Research Unit observations dataset. 17–19, 29

ICE initial conditions ensemble. 12, 13, 17, 19, 43–45, 47, 48

IPCC Intergovernmental Panel on Climate Change. 3, 6, 12, 19, 40, 44

IPO input-process-output. 10

Mk3L Mk3L Climate System model. v, 18, 19, 22, 50

MSE mean square error. 26, 32, 41

PCMDI Project for Climate Model Diagnosis and Intercomparison. 7

PDF probability density function. 10

PPE perturbed physical parameters ensemble. 13, 17, 19, 43–45, 47, 48

PSE perturbed structure ensemble. 13, 17, 19, 23, 43–45, 47

RMSE root mean square error. 8, 26, 34, 38, 40

TSI total solar irradiance. 19, 29

UNSW University of New South Wales. v

WCRP World Climate Research Programme. 3

Bibliography

Abramowitz, G. (2010). "Model independence in multi-model ensemble prediction". *Australian Meteorological and Oceanographic Journal* 59, pp. 3–6.

Allen, M. R. and W. J. Ingram (2002). "Constraints on future changes in climate and the hydrologic cycle". en. *Nature* 419 (6903), pp. 224–232. DOI: 10.1038/nature01092.

Anderson, J. L. (1996). "A method for producing and evaluating probabilistic forecasts from ensemble model integrations". *Journal of Climate* 9 (7), 1518–1530.

Annan, J. D. and J. C. Hargreaves (2010). "Reliability of the CMIP3 ensemble". *Geophysical Research Letters* 37, 5 PP. DOI: 201010.1029/2009GL041994.

Annan, J. (2010). *"IPCC Experts" New Clothes*. James' Empty Blog. URL: http://julesandjames.blogspot.com.au/2010/08/ipcc-experts-new-clothes.html (visited on 10/02/2012).

Barnett, T. P. (1999). "Comparison of Near-Surface Air Temperature Variability in 11 Coupled Global Climate Models". *Journal of Climate* 12 (2), pp. 511–518. DOI: 10.1175/1520-0442(1999)012<0511:CONSAT>2.0.CO;2.

Bishop, C. H. and G. Abramowitz (2012). "Climate model dependence and the replicate Earth paradigm". *Climate Dynamics*. In Review.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones (2006). "Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850". *Journal of Geophysical Research* 111 (D12). DOI: 10.1029/2005JD006548.

Chatfield, C. (2004). *The analysis of time series: an introduction*. 6th ed. Vol. 59. CRC press.

Ducharne, A. and K. Laval (2000). "Influence of the Realistic Description of Soil Water-Holding Capacity on the Global Water Cycle in a GCM". *Journal of Climate* 13 (24), pp. 4393–4413. DOI: 10.1175/1520-0442(2000)013<4393:IOTRDO>2.0.CO;2.

Dunne, K. and C. Willmott (1996). "Global distribution of plant-extractable water capacity of soil". *International Journal of Climatology* 16 (8), pp. 841–859.

Fischer, E. M., D. M. Lawrence, and B. M. Sanderson (2010). "Quantifying uncertainties in projections of extremes—a perturbed land surface parameter experiment". *Climate Dynamics* 37 (7-8), pp. 1381–1398. DOI: 10.1007/s00382-010-0915-y.

Giorgi, F. (2005). "Climate Change Prediction". English. *Climatic Change* 73 (3), pp. 239–265. DOI: http://dx.doi.org.wwwproxy0.library.unsw.edu.au/10.1007/s10584-005-6857-4.

Gleckler, P., K. Taylor, and C. Doutriaux (2008). "Performance metrics for climate models". English. *Journal of Geophysical Research* 113 (D6), p. D06104. DOI: 10.1029/2007JD008972.

Gordon, H. B., L. D. Rotstayn, J. L. McGregor, M. R. Dix, E. A. Kowalczyk, S. P. O'Farrell, L. J. Waterman, A. C. Hirst, S. G. Wilson, M. A. Collier, and others (2002). *The CSIRO Mk3 climate system model*. Vol. 130. CSIRO Atmospheric Research.

Hamill, T. M. (2001). "Interpretation of Rank Histograms for Verifying Ensemble Forecasts". *Monthly Weather Review* 129 (3), pp. 550–560. DOI: 10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Hegerl, G. C., F. W. Zwiers, P. Braconnot, N. P. Gillett, Y. Lou, J. A. Marengo Orsini, N. Nicholls, J. E. Penner, and P. A. Stott (2007). "Understanding and Attributing Climate Change". *Understanding and Attributing Climate Change*. Cambridge, United Kingdom: Cambridge University Press.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2010a). "Challenges in Combining Projections from Multiple Climate Models". *Journal of Climate* 23 (10), pp. 2739–2758. DOI: 10.1175/2009JCLI3361.1.

Knutti, R., G. Abramowitz, M. Collins, V. Eyering, P. J. Glecker, B. Hewitson, and L. O. Mearns (2010b). "Good practice guidance paper on assessing and combining multi model climate projections". *IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections*, p. 1.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran (2000). "Multimodel Ensemble Forecasts for Weather and Seasonal Climate". *Journal of Climate* 13 (23), pp. 4196–4216. DOI: 10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

Macadam, I., A. J. Pitman, P. H. Whetton, and G. Abramowitz (2010). "Ranking climate models by performance using actual values and anomalies: Implications for climate change impact assessments". *Geophysical Research Letters* 37 (16). DOI: 10.1029/2010GL043877.

Maltrud, M. E. and J. L. McClean (2005). "An eddy resolving global 1/10°ocean simulation". *Ocean Modelling* 8 (1–2), pp. 31–54. DOI: 10.1016/j.ocemod.2003.12.001.

Masson, D. and R. Knutti (2011). "Climate model genealogy". *Geophys. Res. Lett* 38 (8), p. L08703.

Meehl, G., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. Mitchell, R. Stouffer, and K. Taylor (2007). "The WCRP CMIP3 multi-model dataset: A new era in climate change research". *Bulletin of the American Meteorological Society* 88, 1383–1394.

Met Office (2010). *Met Office climate prediction model: HadGEM3 family*. en. The HadGEM3 family of climate models represents the third generation of HadGEM configurations and includes the NEMO ocean model and CICE sea-ice model components. URL: http://www.metoffice.gov.uk/research/modelling-systems/unified-model/climate-models/hadgem3 (visited on 10/24/2012).

Milly, P. C. D. and K. A. Dunne (1994). "Sensitivity of the Global Water Cycle to the Water-Holding Capacity of Land". *Journal of Climate* 7 (4), pp. 506–526. DOI: 10.1175/1520-0442(1994)007<0506:SOTGWC>2.0.CO;2.

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

Phipps, S. J., L. D. Rotstayn, H. B. Gordon, J. L. Roberts, A. C. Hirst, and W. F. Budd (2012a). "The CSIRO Mk3L climate system model version 1.0 – Part 2: Response to external forcings". *Geoscientific Model Development* 5 (3), pp. 649–682. DOI: 10.5194/gmd-5-649-2012.

Phipps, S. (2011). *The CSIRO Mk3L climate system model v1.2, Technical Report No. 4*. Hobart, Tasmania, Australia: The Antarctic Climate & Ecosystems CRC. 121 pp.

Phipps, S. J., H. V. McGregor, J. Gergis, A. J. E. Gallant, R. Neukom, Samantha Stevenson, Duncan Ackerley, Josephine R. Brown, Matt J. Fischer, and Tas D. van Ommen (2012b). "Paleoclimate data–model comparison and the role of climate forcings over the past 1500 years". *Journal of Climate*. Submitted.

Pitman, A. (2003). "The evolution of, and revolution in, land surface schemes designed for climate models". *International Journal of Climatology* 23 (5), 479–510.

Ramaswamy, V., O. Boucher, J. Haigh, D. Hauglustaine, J. Haywood, G. Myhre, T. Nakajima, G. Shi, and S. Solomon (2001). "Radiative Forcing of Climate Change". *Climate change 2001: Synthesis report: Third assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Reichert, P., M. Schervish, and M. J. Small (2002). "An Efficient Sampling Technique for Bayesian Inference With Computationally Demanding Models". *Technometrics* 44 (4), pp. 318–327. DOI: 10.1198/004017002188618518.

Reifen, C. and R. Toumi (2009). "Climate projections: Past performance no guarantee of future skill?" *Geophysical Research Letters* 36 (13). DOI: 10.1029/2009GL038082.

Räisänen, J. and T. N. Palmer (2001). "A Probability and Decision-Model Analysis of a Multimodel Ensemble of Climate Change Simulations". *Journal of Climate* 14 (15), pp. 3212–3226. DOI: 10.1175/1520-0442(2001)014<3212:APADMA>2.0.CO;2.

Schmidt, G. A., J. H. Jungclaus, C. M. Ammann, E. Bard, P. Braconnot, T. J. Crowley, G. Delaygue, F. Joos, N. A. Krivova, R. Muscheler, B. L. Otto-Bliesner, J. Pongratz, D. T. Shindell, S. K. Solanki, F. Steinhilber, and L. E. A. Vieira (2012). "Climate forcing reconstructions for use in PMIP simulations of the Last Millennium (v1.1)". *Geoscientific Model Development* 5 (1), pp. 185–191. DOI: 10.5194/gmd-5-185-2012.

Sellers, P. J. and K. E. Trenberth (1992). "Biophysical models of land surface processes". *Climate system modeling*. Vol. 14. Cambridge University Press, 451–490.

Solomon, S, D Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, and M. Tignor, eds. (2007). *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC Fourth Assessment Report: Climate Change 2007 (AR4). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. 996 pp.

Stainforth, D., T. Aina, C. Christensen, M. Collins, N. Faull, D. Frame, J. Kettleborough, S. Knight, A. Martin, J. Murphy, and others (2005). "Uncertainty in predictions of the climate response to rising levels of greenhouse gases". en. *Nature* 433 (7024), 403–406. DOI: 10.1038/nature03301.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012). "An Overview of CMIP5 and the Experiment Design". *Bulletin of the American Meteorological Society* 93 (4), pp. 485–498. DOI: 10.1175/BAMS-D-11-00094.1.

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

Washington, W. M. and G. A. Meehl (1989). "Climate sensitivity due to increased CO2 experiments with a coupled atmosphere and ocean general circulation model". *Climate Dynamics* 4 (1), pp. 1–38. DOI: 10.1007/BF00207397.

Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller (2010). "Risks of model weighting in multimodel climate projections". *Journal of Climate* 23 (15), 4175–4191.

Zou, H. and Y. Yang (2004). "Combining time series models for forecasting". *International Journal of Forecasting* 20, pp. 69–84. DOI: 10.1016/S0169-2070(03)00004-9.