COMPARISON OF BAYESIAN CALIBRATION METHODOLOGIES FOR CLIMATE SYSTEM MODELS

TRISTAN HAUSER

Comparison of Bayesian Calibration Methodologies for Climate System Models

by

© Tristan Hauser

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Master of Science

Department of Physical Oceanography

Memorial University of Newfoundland

August 2009

St. John's Newfoundland

Abstract

Earth Systems models that attempt to forecast equilibrium states or make long term predictions are sensitive to the unavoidable approximations they employ. It is therefore important for such models to be parameterized through objective and repeatable methods that quantify the uncertainties associated with the inexactness of these approximations. In this study Ensemble Kalman Filters and Neural Network Bayesian Models are used to investigate parameter sets for the Budyko Energy Balance Model and the more computationally demanding Planet Simulator of the University of Hamburg Meteorological Institute. These calibration methods employ observational data to generate posterior probability distributions for model parameter sets, allowing the determination of high-probability parameter sets and their confidence intervals. Being fully Bayesian, such approaches accurately propagate uncertainties in observational data into the posterior distributions. Comparing calibrated model results permits the two approaches to be assessed under varying levels of model complexity.

Acknowledgements

This project was conceived, orchestrated and tirelessly assisted by Dr. Lev Tarasov. The implementation of the experiments was made possible by Dr. Michelle Shaw, Stephen Condran, Christopher Stevenson, and Robert Briggs. Writing and editing was aided by Dr. Andrew Keats. This work has benefited from continued assistance and feedback from Dr. Martina Schäfer, Graig Sutherland, and Robert Briggs.

Contents

Abstract ii

Acknowledgements iii

List of Tables vi

List of Figures vii

1 Introduction 1

1.1 Motivation 1

1.2 The Closure Problem and Model Calibration 4

1.3 Bayesian Methodology 7

1.4 Outline of Experiments 9

2 Calibration Methodologies 13

2.1 The Ensemble Kalman Filter 13

2.1.1 Overview 13

2.1.2 Extension of algorithm to an ensemble method 15

2.1.3 Joint Parameter and State Estimation Problem 17

2.1.4 Method Extensions and Variations 17
2.2 Markov Chain Monte Carlo Sampling Using Neural Networks 19
2.2.1 Multilayer Perceptron Neural Networks 19
2.2.2 Solution Space Sampling Routine for Neural Networks 22
2.2.3 NN/MCMC Calibration Routine 25

3 Experiments with Budyko Energy Balance Model 28
3.1 Outline of Model 28
3.2 Calibration using the EnKF 32
3.3 Calibration using NN/MCMC 40
3.4 Discussion 43

4 Experiments with Planet Simulator 48
4.1 Overview of Experiment 48
4.2 Calibration using the EnKF 53
4.3 Calibration using NN/MCMC 57
4.3.1 Calibration Routine 57
4.3.2 Results Produced by the Calibration Routine 61
4.4 Discussion 67

5 Conclusion 73
5.1 Summary and Future Work 73

Bibliography 78

Appendix 82

List of Tables

3.1 Initial EBM Parametrization 31

3.2 EnKF Settings for EBM calibration 32

3.3 Prior distributions for EBM calibration 32

3.4 Architecture of neural network used for the EBM calibration routine . 40

4.1 Investigated Parameters and their Priors 51

4.2 EnKF Settings for Planet Simulator calibration 53

4.3 Architecture of Neural Networks used for the Planet Simulator calibration routine 57

4.4 Comparison of global (without poles) RMSE between the observed state and the weighted ensemble and default model winter fields . . . 66

A.1 Terms for basic Kalman Filter derivation. 82

List of Figures

2.1 Conceptual flow chart of the basic Kalman filter. 15

2.2 Conceptual Flow Chart of Calibration using the EnKF. 18

2.3 Conceptual Flow Chart of a Neural Network with three inputs, one hidden layer of size four, and two outputs. 19

2.4 Conceptual flow chart of calibration using NN/MCMC routine. 25

3.1 Model results (black dots) compared to observational data (green dots, dashed lines represent observational uncertainties). 31

3.2 State space results from final iteration of the EnKF routine, ensemble mean (black line) compared to observational data (green dots). 34

3.3 The evolution of ensemble parameters,while iterating the EnKF routine, displayed as ensemble mean (black line) and standard deviation (black dash). 35

3.4 The evolution of ensemble parameters,while iterating the EnKF routine, displayed as ensemble mean (black line) and standard deviation (black dash). 36

3.5 Prior (uniform distributions, shown as box and whisker plots) and final posterior (shown as Gaussian bell curves) for the EnKF calibration of the EBM. 37

3.6 Prior (uniform distributions, shown as box and whisker plots) and final posterior (shown as Gaussian bell curves) for the EnKF calibration of the EBM. 38

3.7 The mean (black line) and standard deviation (black dash) produced by creating a model ensemble by selecting parameter values from the EnKF created posterior distributions, compared to observational data (green dots). 39

3.8 Distributions of parameter sets for the EBM produced from the prior, then two sequential executions of MCMC posterior sampling. 42

3.9 The mean (black line) and standard deviation (black dash) produced by creating a model ensemble by selecting parameter values from the neural network emulator-derived posterior distributions, compared to observational data (green dots). 44

4.1 Observed difference in mean seasonal surface temperatures (in degrees Celsius) for DJF (upper left) MAM (upper right) JJA (lower left) and SON (lower right) between 2008-1999 and 1959 - 1968. 50

4.2 Sensitivity testing of parameter "*acllwr*". 52

4.3 Sensitivity testing of parameter "*tdissd*". 53

4.4 Sensitivity testing of parameter "*tswr*1". 54

4.5 Sensitivity testing of parameter "$vdiff_{lamm}$". 55

4.6 Distributions of individual parameter values from the prior and first iteration of the EnKF analysis routine. 56
4.7 Fits between training data (x-axis) and network predictions (y-axis) for the neural network simulating model output for the Siberian region for the winter and summer seasons. 59
4.8 Fits between training data (x-axis) and network predictions (y-axis) for the neural network simulating model output for the South Pacific region for the winter and summer seasons. 60
4.9 Fits between test data (x-axis) and network predictions (y-axis) for the neural network simulating model output for the South Atlantic region for the winter and summer seasons. 62
4.10 Fits between test data (x-axis) and network predictions (y-axis) for the neural network simulating model output for the North American region for the winter and summer seasons. 63
4.11 Distributions of individual parameter values from the prior and two iterations of the MCMC analysis routine. 64
4.12 Plot of log-likelihood values calculated from model output and observations for members of the original model ensemble, that of the ensembles produced by two iterations of the NN/MCMC routine, and the original default parameter settings. 66
4.13 Difference between observed and weighted ensemble mean winter surface temperatures for 1959-1968 (top left) and 1999-2008 (top right) with the standard deviation for the ensemble results (below). 68

4.14 Difference between observed and weighted ensemble mean winter surface temperatures for 1959-1968 (top left) and 1999-2008 (top right) and difference between observed and standard model mean winter surface temperatures for 1959-1968 (bottom left) and 1999-2008 (bottom right). 69

4.15 Difference between observed and ensemble anomaly between seasonal surface temperatures for 1959-1968 and 1999-2008. 70

Chapter 1

Introduction

1.1 Motivation

Models of earth systems are useful tools for investigating natural processes whose scale and/or complexity prevents them from being observed in their entirety or from being manipulated for physical experiments. However, these models pose challenges in their development and interpretation. By their nature models are limited and/or generalized descriptions designed in accordance to the scope of the investigations they facilitate. While the mathematical descriptions of the modeled processes are often very sophisticated and accurate they inevitably contain certain approximations. This is especially true when these models are to be used in forecasts and experiments pertaining to "real world" contexts, as in such applications it is desired to address as many elements of the total system as possible, rather than describing one particular physical process within the system. This introduces the issue of closure, which is discussed below. As models expand to include more components of the earth system,

the number of approximations invoked also tends to increase. These approximations generally require parameters whose values are not derivable from first principles. Such model parameters can not be said to have a correct value. Rather one is chosen that "works". As earth system models generally have nonlinear dependencies on these parameters, the determination of appropriate parameter values is a highly nontrivial task. This component of model construction is hardly, if ever, documented in published literature. As such there are two key sources of generally unquantified uncertainly induced; those associated with the formulation of the approximations, and those associated with the setting of the parameter values these approximations utilize.

This is a disconcerting situation to anyone who wishes to apply model results to practical applications. In geoscience and engineering applications models are often calibrated, as in they are tested against and modified to match data that as closely as possible represents situations they are to be used to forecast e.g (Nettuno 1995; Moradkhani et al. 2005; Khu and Henrik 2005). The goal of such calibration is both to improve prediction and to gain a quantitative estimation of model uncertainties. This practice is less common in the field of climate science. The earth system models employed in this field are much more complex and computationally expensive than those used in many other applications. The time and spatial scales considered in climate modeling make it difficult to prescribe with certainty appropriate calibration data. Also, the highly nonlinear nature of these models increases the difficulty of identifying correlations between parameters and model output. Thus the tendency of subjective parameter selection in climate model development. This is undesirable, as at present time climate projections based on earth systems models are being looked

to on many levels to inform responses to the issue of climate change. As similar models can often produce very different results and provide limited estimation of the uncertainties in their predictions, climate models function poorly as decision making tools. Due to their nonlinear nature, the evolution of climate systems are inherently impossible to describe in explicit deterministic fashion. Therefore, discussions regarding climate change issues are inherently ones about risk management. This is a process understood by anyone who has ever used a weather report to plan a future outdoor activity. The projections of current climate models are for the most part not presented in a form that allows for that type of decision making.

The above concerns have been expressed by many sources, including current reports by the Intergovernmental Panel on Climate Change (Solomon et al. 2007). Various methods for dealing with the aspects of climate modelling which make calibration difficult have also been presented. The ensemble Kalman filter (discussed below) has been proposed as a potential objective calibration tool for climate models (Annan and Hargreaves 2004). While the potential of this algorithm for parameter estimation has been addressed for certain contexts (Evensen 2005; Evensen 2009) it is not commonly seen in climate forecasting contexts. The use of Markov chain sampling methods for estimating model uncertainties has been investigated (Oakley and O'Hangan 2002; Jackson et al. 2004). The use of statistical emulators as a way of coping with the computational demands of current earth systems models has also been discussed (Rougier 2008) and their application to lessen the computational expense of Markov chain methods has previously been put into practice (Tarasov and Peltier 2005). While the components of these methods have been highly developed, there is little guidance available concerning their implementation with regard to com-

plex earth systems models. The following work is an exploratory examination of their practical application to such models.

1.2 The Closure Problem and Model Calibration

An inherent issue in the development of earth systems models is that of closure (Müller and von Storch 2004), i.e. that of defining the boundaries of a system under investigation. Natural processes occur within continuous, open systems. At some level, all elements of earth systems processes, as they exist within the global system, affect one another. For these processes there is nothing that can be considered large or small scale enough to be truly regarded as external. Earth systems models, however, are by necessity closed and discrete, and therefore incomplete in their construction. They must focus on specific scales, regions and phenomena within a larger system. Furthermore, even within the particular range of focus of a given numerical model it is impossible to describe the entire physical state of a given element. Instead, such models operate using discrete representations of distinct qualities at selected locations. The mathematical formulations of the physical laws on which numerical models are based are to various degrees approximated and parametrized in order to account for external forcing and sub-scale processes that are not explicitly calculated in the model (Müller and von Storch 2004).

Because of this fundamental divide between their numerical formulation and the reality they approximate, these models can not be assessed according to whether they generate "correct" or "incorrect" numerical solutions to benchmark problems. Rather, their effectiveness is determined by the degree of consistency maintained

with the phenomena they attempt to simulate, as well as how well their construction agrees with the current understanding of the processes they describe (Müller and von Storch 2004). A climate model does not explicitly output, for example, that Arctic summers will be on average warmer in the future than they are at present. Such a model instead computes temperature values of grid cells for discretized time-steps. Taken individually, these values are next to meaningless as they are almost certainly, by formal standards, "wrong". No modeler or weather forecaster would ever expect twenty years after making a prediction of a certain temperature for that future date to see their model "verified" by a recording of the exact value predicted. The same applies to the approximations (parameters) these models employ. Compiling data to calculate an exact average planetary albedo (for example) for use in an energy balance model would be impossible and the discrete result again would be almost meaningless. However, while the explicit output of numerical models cannot be seen as the defining criteria of their success, these individual elements can still be used to invalidate models. An Arctic winter temperature prediction of 313 Kelvin under present conditions, or a modern planetary albedo value set so high as to only be possible under conditions of extensive glaciation, are both clearly "wrong", and imply that the model being used is ineffective. Despite intrinsic uncertainties in their exact meaning and derivation, numerical values of input parameters must be prescribed that can be considered reasonable in the context of past experience and current understanding.

The parametrization of model equations, i.e. having to select and interpret numerical values for the simplified representation of complex processes, is a fundamental source of model uncertainty. In order to accurately interpret model results, this uncertainty needs to be quantified as far as possible. Also, in many cases numerical models

(in keeping with the behaviour of the non-linear systems they represent) can be very sensitive to these parametrized sub-processes, with small changes in parametrization causing large changes in model output. Therefore, numerical models must be objectively calibrated against relevant observational data with methodologies that accurately account for uncertainties. Assessing model results against observational data is not a trivial operation. Observations can very rarely be compared directly to model output and have their own associated uncertainties which must be taken into account when using them as validation criteria. Also, while most calibration methodologies are formulated based on the concept of optimizing the match between model output and some target state or data record, it can be dangerous to think of the procedure solely in these terms. A model that is tuned to a limited set of observations from a given climatic regime may have little predictive power for alternate climate regimes. Furthermore, the quantified forecast uncertainty is often not a measure of potential goodness of fit. Rather, in climate and earth systems modelling, computed uncertainties are often based on the range of behaviors forecast by the model. It is a priori unclear the extent to which model variability is a measure of the true range of potential system behaviors.

The calibration techniques used in this work are ensemble methods based on Bayesian methodology, as described below. These methods, to large though incomplete extent, address the issues raised above with respect to the inexactness of approximations and predictions as well as the uncertainties inherent in observational information used to initialize and calibrate the model. They also provide an objective and repeatable framework for approaching these issues.

1.3 Bayesian Methodology

The central concept of the Bayesian methodology is the use of probability as an expression of uncertainty (Neal 1996). This is common in everyday language. One is often comfortable using prior experience to "lay odds" on an outcome, even without previously observing the exact event. In a more traditional statistical context however, such an expression has no meaning. In such contexts, probability is used only as a description of frequency. An event has a "three out of ten" chance of occurring only if there is record to show that it has, on average, occurred three times for every ten identical instances. In Bayesian terms, though, the statement "three out of ten" is representative of a degree of confidence, i.e. that one is less confident in the occurrence of this event than one ranked by the odds "seven out of ten". This form of expression of uncertainty is utilized in Bayesian statistics because it provides an existing formal mathematical method for inference; the rules of probability. For example, if one event is granted "three out of ten" odds and another "seven out of ten" then in Bayesian methodology, provided the events are mutually exclusive, this is an expression of complete confidence that either the former or later event will occur, as $30\% + 70\% = 100\%$.

The other advantage of a probabilistic expression of uncertainty is that it can be translated directly from statistical data, allowing again the use of the formal rules of mathematical probability to improve prior beliefs as more information becomes available. This is accomplished through Bayes' rule (from which the method receives its name),

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

which express the probability of an event A given the occurrence of B in terms of the probability of the events occurring independently and in the conditional probability of B should A occur. This formula is derived as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)},\ P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A|B)P(B) = P(B|A)P(A)$$

from the standard definitions of conditional probability (Hogg and Tanis 2001).

The application of the above formula can be used to improve probabilistic models. One might attempt to describe the probability distribution $P(x)$ for a set of unknown quantities $\{x_1, x_2, \dots\}$ as a function of parameters θ, ie $P(x|\theta)$. For example, if $P(x)$ is a Gaussian distribution $N(\mu, \sigma^2)$, then $\theta = \{\mu, \sigma^2\}$. The model $P(x)$ with a set parametrization is used to learn about the probability of occurrence for certain values or events $\{x_i | i \in \mathbb{N}\}$ by computing $P(x_i|\theta)$. If, however, there exist a set of independent observations $\{x_1, x_2, \dots, x_n\}$ then these can be used to learn about the statistical confidence in a given parametrization θ. This is done by computing the likelihood function,

$$L(\theta|x_1, x_2, \dots, x_n) \propto P(x_1, x_2, \dots, x_n|\theta) = \prod_{i=1}^{n} P(x_i|\theta) \tag{1.2}$$

and so the likelihood of different parameter sets $\{\theta_j | j \in \mathbb{N}\}$ is investigated by computing the condition probability $P(x|\theta)$ of constraint quantities x_i with respect to the parameter vector θ (Hogg and Tanis 2001). In the Bayesian methodology, uncertainty about how to effectively parametrize the model is expressed probabilistically. Here the probability distribution $P(\theta)$ expresses the prior understanding of parameter selection. This allows the use of Bayes' rule from above as,

$$P(\theta|x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n|\theta)P(\theta)}{P(x_1, x_2, \dots, x_n)} \propto L(\theta|x_1, x_2, \dots, x_n)P(\theta) \tag{1.3}$$

where the posterior distribution $P(\theta|x_1, x_2, \ldots, x_n)$ represents the improved understanding of θ (Neal 1996). The outcome of this calculation can then be used to improve the future prediction of the observed quantity so that,

$$P(x_{n+1}|x_1, x_2, \ldots, x_n) = \int P(x_{n+1}|\theta)P(\theta|x_1, x_2, \ldots, x_n)d\theta \tag{1.4}$$

where $P(x_{n+1}|x_1, x_2, \ldots, x_n)$ is the predictive distribution (Neal 1996).

In the previous and following discussions the models under investigation are numerical models of deterministic dynamical systems. Where these models are uncertain is in the degree of mismatch between their outputs and the phenomena they are meant to describe. In keeping with the Bayesian methodology the mismatch between the model and a perceived true state is considered a forecast uncertainty to be treated probabilistically. This is done by considering the true system state, ψ_T, to be described by the model forecast, ψ_f, plus a stochastic process describing the model error/uncertainty, ρ, written as $\psi_T = \psi_f + \rho$. Furthermore, the information used to improve models of a given system, the $x_1, x_2, \ldots, x_n$ from above, are inexact observations. The uncertainty concerning the accuracy of the measurements is expressed in the same form, $\psi_T = \psi_{obs} + \epsilon$, where again the final term, ϵ represents the measurement error and is considered the result of a stochastic process. The implementation of this and the methodology for describing the resulting posterior distribution will vary from case to case and so will be discussed in more detail as warranted.

1.4 Outline of Experiments

This work documents experiments using two different calibration techniques with two different numerical models. All are discussed in detail in the following chapters.

The calibration techniques employed are ensemble methods, meaning they utilize large numbers of differently parametrized model runs to generate information about model response. As numerical models can be very computationally expensive to run, as well as potentially very nonlinear in their responses, it is often neither possible nor informative to undertake random or exhaustive samplings of model behavior. Here the Ensemble Kalman Filter (EnKF) and Neural Network assisted Markov Chain Monte Carlo sampling (NN/MCMC) are used to focus these samplings so as to extract useful data concerning model behavior. Both techniques are in essence routines to find approximate solutions to the Bayesian inference problem discussed above. Starting with a prior distribution of possible parameter sets, the goal is to use the above methods and available data concerning the observed state of the modeled system to create an informative (and ideally narrower) posterior distribution of parameter sets. Selecting parameters from the posterior distribution should allow the model to make improved predictions about the system it describes. Of the commonly used Bayesian methods for combining observational data and prior insight in a modeling context, the EnKF and NN/MCMC methods are the most directly applicable to the problem of model calibration (Wikle and Berliner 2007).

The EnKF over many iterations tracks the development of members of an ensemble of models. It actively adjusts the attributes and outputs of the individual ensemble members to provide a closer fit to observational data. This incorporation of observations in order to refine model predictions is known as data assimilation, and is the common application of the EnKF (Kalnay 2003; Evensen 2007). The EnKF formulation also lends itself to the weighting of parameter sets given model performance compared to observations (Evensen 2005). This technique has had some success in en-

gineering applications (Moradkhani et al. 2005) and initial attempts have been made to apply it to larger scale models (Annan et al. 2005). Data assimilation does not occur in NN/MCMC calibration methods. Neural networks use the results from an ensemble of model runs to form a complex statistical emulation that predicts model output given a new parametrization. This emulator is then used as part of a MCMC sampling routine to select parameter sets that are more likely to produce output that closely match observational data. The potential of using emulators of some form to decrease computational expense has been discussed (Annan and Hargreaves 2007) and have been utilized in model calibration (Tarasov and Peltier 2005) in the form of Bayesian neural networks.

While both methods provide approximate solutions to the same problem, there are some important differences in their respective approaches. The EnKF routine assumes that all uncertainties are expressible as Gaussian distributions, whereas the approach utilizing neural networks can handle any explicit uncertainty distribution. Therefore, it is rare that the resulting posterior distributions can be compared directly. Also, unlike neural networks, the EnKF routine is also a method for combining observational data and model forecasts into an improved prediction. The NN/MCMC approach calibrates model parameters but does not otherwise directly modify model predictions. The results of the direct manipulation of the model predictions in the EnKF should not be misinterpreted as result of the calibration routine.

For some simple test problems it is possible to attempt to tune models to a set "standard" model run with preset initial conditions and parameter values. This "standard" provides the "observational" data and the final assessment of success. Such experiments have been performed with various simple models, such as the Lorenz

equations (Annan and Hargreaves 2004), bimodal stochastic systems (Kim et al. 2003), highly simplified atmospheric-slab ocean model system (Jackson et al. 2004), and a low resolution coupled atmosphere-ocean model (Annan et al. 2005). This form of assessment provides some insight into the methods being investigated, but does not mimic a realistic scenario. In any realistic situation there would be no access to the "true" state of the system, and at times only limited understanding of the extent to which measurements differ from the real system being observed. Also when comparing a model to reality it makes no sense to think of there being "true" parameter values given the dynamical simplifications these parameters represent.

The methods presented in the next chapter were applied to a basic Energy Balance Model (EBM). The simplicity and relative transparency of the model will assist discussion. A parameter set that is appropriate for investigation will be described as well as a set of reanalysis data to be used as the observational information. This will serve as a demonstration of the setup and analysis of the calibration methods.

Insights gained about the effective use of the two calibrations procedures from experiments with the EBM will then be applied to a Global Circulation Model (GCM) forced with historical observations of atmospheric CO_2 from the past fifty years. This model will be calibrated against observed seasonal climatologies for the 50 to 40 years BP (before present) interval and for the last 10 years. The results of the calibration procedures will then be compared against the climatological record.

Chapter 2

Calibration Methodologies

2.1 The Ensemble Kalman Filter

2.1.1 Overview

The Kalman filter is a method that discretely samples from an approximation of the posterior PDF by assuming that errors in the models and observations can be expressed as Gaussian noise. Kalman filtering constructs a new (analysed) state by altering the result of a model's forecast of that state with respect to the difference between observational data and what observations would be expected if the forecast state were a true description of the reality being observed. This difference between forecast and observed state is called the innovation. When correcting the ensemble towards the observational measurements, the weight given to how much the innovation affects the difference between the forecast and analysed state is computed from information about uncertainties in the forecast model and observational measurements. This weighting scheme is referred to as the Kalman gain.

As outlined in Figure 2.1, the algorithm begins with a state estimate and a covariance matrix that expresses its estimated uncertainties. These elements are either prescribed as initial conditions ($t = 0$) or are a result of previous iterations of the process ($t-1$). The estimate is then used by the model to predict the next state. The innovation is calculated from the externally acquired observational data and the forecast. Often the observational data is not of the same system aspect being estimated by the model and a transition operation must be employed, i.e. one must calculate what observations are predicted by the model. Even if there are direct observations of the elements forecast by the model, often the observations will differ in quantity and spatial distribution and a transition operation will still need to be employed. The Kalman gain matrix is computed from the uncertainty estimates and is then used to compute the updated state estimate with a corresponding uncertainty estimate. These results provide the starting point for the next iteration of the algorithm ($t+1$). Each iteration could represent a time step as the process being modeled and observed progresses in time, or alternatively, iterations could be applied repeatedly on a steady state model in order to improve the estimate. The derivation of the Kalman filter algorithm is described in detail in the appendix. Key points of this derivation are as follows:

- That the Kalman gain matrix is "optimal" in that it minimizes the analysis error covariance. This minimization is equivalent to minimizing a cost function in the Bayesian formulation if for that problem the error terms are assumed to be derived from Normal distributions (Evensen 2007).

- The use of error covariance statistics avoids the problem found in spatial in-

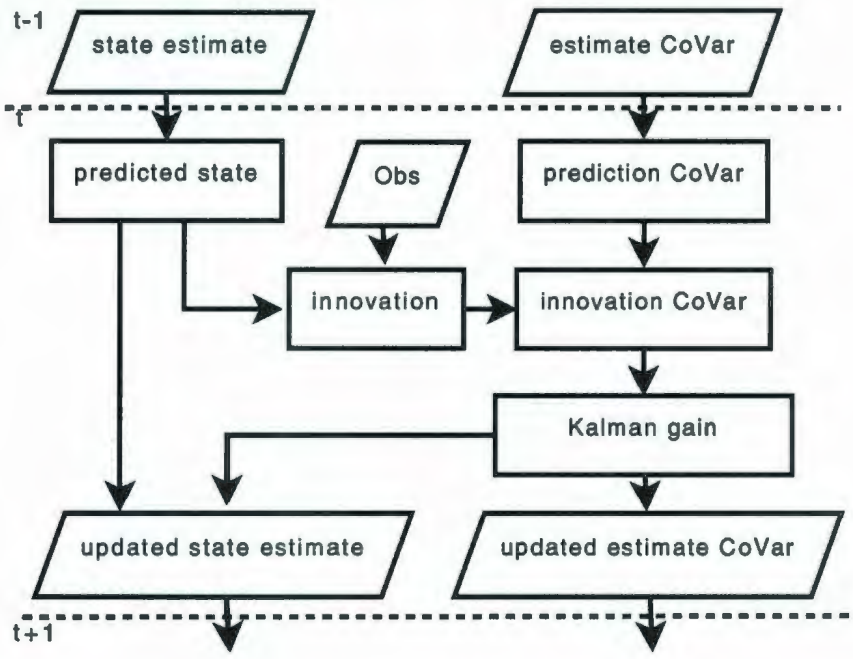

Figure 2.1: Conceptual flow chart of the basic Kalman filter.

terpolation data assimilation where similar (bunched) data points can weight the analysis, i.e. there is a function to express the significance of the data as opposed to its density (Kalnay 2003).

- There is an assumption that model and observation errors are unbiased, which for many practical applications may not be entirely realistic.

2.1.2 Extension of algorithm to an ensemble method

At its simplest, the process outlined in the section above (and described in more detail in the appendix) is designed as a linear process and therefore not applicable to nonlinear systems. One approach to cope with this problem is the construction of an Extended Kalman Filter (EKF), where the forecast and observation models

are created through local linearization. While these new equations can no longer be shown to optimize the result, this can be addressed on a practical level by adjusting the assumed noise statistics to compensate for vagueness in the model construction (Gershenfeld 1999). Still, for highly nonlinear systems the effectiveness of the EKF is limited (Evensen 2007). A more effective method involves creating an ensemble of runs (using nonlinear forecast and observation models) and using the comparison of these results to derive the statistics that propagate the filter routine, which is performed on each individual ensemble member. This is the Ensemble Kalman Filter (EnKF). The details of the extension of the Kalman Filter algorithm to an ensemble method, as well as the implementation scheme used in this project are given in the appendix. The key points of this extension and its implementation are as follows:

- The equations are essentially the same as before but error values are statistically generated from the ensemble. Thus, in practice, all the needed error terms become available and need not be set externally.

- As $N \rightarrow \infty$, where N is the number of ensemble members, this formulation approaches that of the standard filter. The assumption that the mean of the ensemble forecast approaches the true system state as $N \rightarrow \infty$ is harder to defend.

- As the EnKF is an analytical method for extracting information from the posterior PDF of the joint state-observation problem, each ensemble member represents a single sample drawn from this PDF. It is important to remember that these PDFs can be much more complex in nature than is apparent from the sampling algorithm. (Anderson 2001).

2.1.3 Joint Parameter and State Estimation Problem

The EnKF can be used for model parameter estimation through a trivial extension of the filter. The model state is augmented to contain the parameters of interest and they are treated as a part of the state space with no corresponding observations. This invokes no conceptual jump, as in most applications there will be many state space members that do not correspond directly to observational data. However, as the correspondence between the observed data and the model parametrization is often rather indirect and nonlinear the ability of the innovation and Kalman Gain estimates to efficiently direct the parameters towards more accurate values can be limited. Also, preventing the ensemble from converging at local minima can require nontrivial manipulations of the process (Annan et al. 2005). The procedure used in the following experiments is outlined in Figure 2.2. A prior distribution defines what parameter sets are used when generating an ensemble of model runs. The outputs of these runs are used along with observational data in the EnKF algorithm, which results in a new distribution of parameter sets. These sets are then used to create a new ensemble, and ideally the procedure continues until the updated distribution of parameter sets becomes invariant, i.e. the filter no longer adjusts the parameter values of the individual ensemble members. In practice, with computationally expensive models the number of iterations is limited by available computational resources.

2.1.4 Method Extensions and Variations

The above methodology has many variations designed to improve efficiency and accuracy. The most direct extensions following from the above method are de-

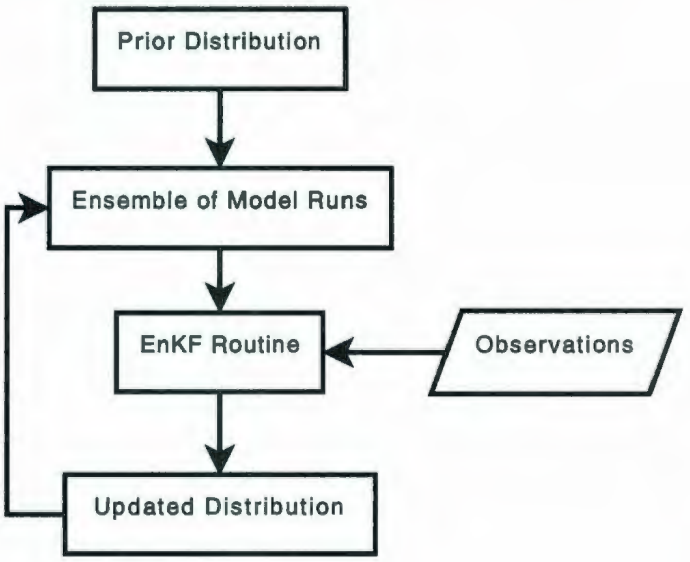

Figure 2.2: Conceptual Flow Chart of Calibration using the EnKF.

scribed in (Evensen 2004) and their implementation is currently available on-line at http://enkf.nersc.no/. These extensions are designed for improving the computational performance for applications where the dimensions of the forecast and observational state are very large. Other extensions involve methods for approximating the ensemble statistics in order to decrease computational time (Bishop et al. 2000; van der Merwe et al. 2000) or approaches which restrict the influence of observational data on the forecast elements (Anderson 2001). These extensions are of interest to data assimilation applications rather than calibration. Thus, in the following experiments only the method outlined in the appendix is used.

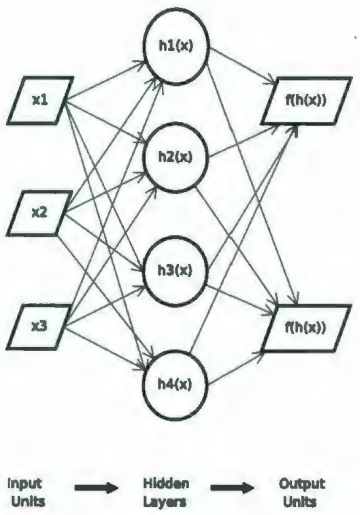

Figure 2.3: Conceptual Flow Chart of a Neural Network with three inputs, one hidden layer of size four, and two outputs.

2.2 Markov Chain Monte Carlo Sampling Using Neural Networks

2.2.1 Multilayer Perceptron Neural Networks

As calibration using MCMC methods would require the serial execution of a prohibitive number of model runs, statistical emulations are used as economical substitutes for the models themselves. The relationship between the parameter set utilized and the model output can be very complex and in general for numerical earth system models there is no simpler deterministic description of this relationship than the model itself. Therefore, for a computationally expensive model some form of statistical regression between parametrization and output is required. Here neural networks are used as non-linear regressors of some aspect of model output against model pa-

rameterizations. Neural networks can often perform to a sufficiently high degree of accuracy as to be useful in identifying new parameter sets that create good matches between model output and observational information. Different statistical emulation approaches have been used to capture the behavior of computationally demanding models. These include the description of atmospheric models by outer-product emulators (Oakley and O'Hangan 2002) and avalanche models by Bayes linear inference (Rougier 2008). In this work, multilayer perceptron neural networks are used. This decision is motivated by the assertion that these emulators are general and flexible enough that the training of such neural networks can be to a large extent automated, with a general network structure utilized to capture various model outputs (Neal 2004).

Multilayer perceptron networks are tools for the statistical emulation of complex systems. A network of functions described by prescribed weights and biases is used to map given input to an expected output. To avoid confusion with the earth system model parameters addressed in this work, these weights and biases will be referred to as Neural Network Parameters (NNPs). As the functions in the network are derived from statistical relations between previously observed data rather than through a system of descriptive equations these networks are more computationally efficient than a numerical model and can be implemented with much weaker understanding of the underlying system dynamics. These networks are more flexible than many statistical regression methods that are based on linear correlations as they contain so called hidden layers, composed of nonlinear functions. Furthermore, for the Bayesian neural networks employed herein, the NNPs are not single valued. Instead, they are defined by probability distributions numerically derived by training the network

against "observed" input-output maps (ie the training dataset). The resulting neural network is considered a nonparametric model, as it does not assume any type of statistical distribution when fitting to data. However, this does not mean that the network lacks defining parameters. Rather, the NNPs are far more numerous and less conceptually meaningful then those found in so labeled parametric models (Neal 1996). Figure 2.3 depicts the architecture of a simple multilayer perceptron network. The functions $h(\cdot)$ and $f(\cdot)$ compute the weighted sum of the values input to them, which is then adjusted by a bias. In the hidden units, $h(\cdot)$, a nonlinear operation is applied to this weighted sum as discussed above. In Figure 2.3 this process is used to compute two distinct output values.

Figure 2.3 displays a single example of a neural network. The architecture of individual networks can vary a great deal. While the number of inputs and outputs are determined by the problem being addressed, the number of hidden layers and their sizes are at the discretion of the user. The form of connections between layers is also adjustable. Figure 2.3 shows a simple system where the input information feeds only to the first hidden layer, which is then the only source of information for the final calculations. Direct connections between input and output can be invoked as well as networks where certain hidden layers are reserved for certain input elements. Also, as the NNPs of each network element are a result of Bayesian inference, prior distributions and further parameters (referred to as hyper-NNPs) must be assigned to determine how the network elements are affected by the training routine. These factors must all be adjusted either manually or automatically in order to accommodate the problem being investigated with the network.

The link between network architecture and the workings of the system or model

that is being emulated is often vague at best (Neal 1996). In this context, Bayesian multilayer perceptron networks have a key advantage in that they minimize the risk of "over-fitting" (explained below), no matter how large the network (Neal 1996). Given this feature and given that Bayesian networks can function well with vague or meaningless network priors, computational capacity often has a larger impact on network design than prior beliefs about a system. Often, network architecture is (re)arranged so as to improve the fit between the network and a set of test data. A common concern with statistical methods is over-fitting; the situation where the emulator creates too close a fit to the members of a given data set (and thereby to the noise in the data set) rather than the trends it displays, and so hampers its predictive ability. Bayesian networks are resistant to over-fitting, with the further practical advantage that all the available data can be used as training data. This is critical for the context herein where the generation of training data is computationally the limiting factor. The usual non-Bayesian approach is to split the constraint data into training and test data and stop training when fits to the test data set start to degrade. The fact that network construction can often be rather ad hoc is part of what makes the method flexible and approachable but makes it difficult to argue that the utilized network is optimal for the problem at hand.

2.2.2 Solution Space Sampling Routine for Neural Networks

The Bayesian learning routine used to determine NNPs and their associated hyper-NNPs results in a posterior distribution that is far too complex to be calculated explicitly. Rather, a description of the space is obtained by searching it using MCMC

methods. These methods sample the desired distribution at many distinct points and then use the statistics generated by this sampling to approximate the nature of the equation that could not be solved for directly. At their most basic, these methods are "random walks" through the space being examined; however, for a complex problem these will not produce a usable result within any feasible amount of computational time. The more involved methods used here attempt to direct the search to cover more of the sample space and/or be more focused on its informative areas in a lesser number of iterations. Following is a brief overview of the approach employed here for training the neural networks. This use of MCMC sampling in training the neural networks should not be confused with the MCMC sampling that is used to evaluate the posterior distribution for model parameters in the calibration routine. The MCMC procedure currently under discussion is only used to produce the neural networks that will be used in the model parameter space sampling discussed in the following sections.

First the hyper-NNPs are investigated by Gibbs sampling. This algorithm samples a given multivariate distribution by using this distribution as a conditional distribution for each variable in turn while the other variables remain fixed. For situations where this algorithm can be implemented it is very desirable as the search method parameters are determined by the sampled distribution and so it does not introduce any further parametrization schemes into the system (Neal 1996).

Gibbs sampling can not be used to directly sample the NNPs as the resulting conditional distributions are too complex to provide useful estimates (Neal 1996). Instead, after each new set of hyper-NNPs becomes available, a Hybrid Monte Carlo algorithm is employed. In this method a parameter set is considered to be the po-

sition vector of an allegorical particle, which has an associated "potential energy" computed from the likelihood of the set given the training data and the prior distribution, as well as a stochastic momentum vector (Neal 1996). Each "position" reached by a "particle" is considered to be a sample and is then used along with a new momentum vector to calculate the next position. Local minima are avoided by arranging the equations of "particle motion" so that the total "energy" of the system is conserved; therefore, as the "potential energy" decreases near an area of better fit to the training data, the "momentum" increases to widen the search area. However in the approximate discrete form, which requires finite time steps, the total energy does not stay completely constant (Neal 1996). This is accommodated for by including a Metropolis algorithm to constrain the search. If for each iteration the total "energy" of the particle is unchanged then this next step is accepted. If not, the probability of the new step being accepted is proportional to the difference that exists between the new energy value and the old. If the step is rejected then the previous sample is recorded as the value for the current iteration as well.

The use of finite step sizes in computing the Hybrid Monte Carlo algorithm introduces two new elements to the system which must be externally determined. These are step size and the number of steps to be taken between each sampling. A more detailed discussion of both of the MCMC methods described above can be found in (Neal 1993).

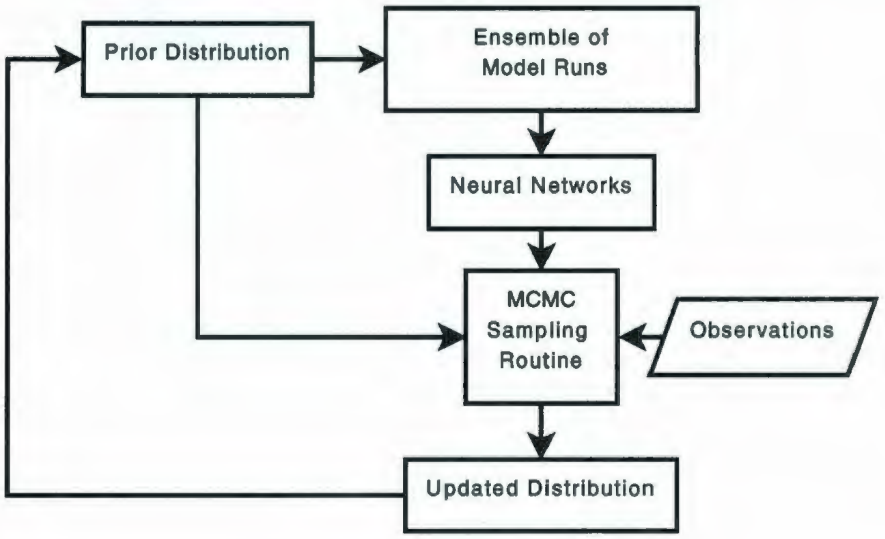

Figure 2.4: Conceptual flow chart of calibration using NN/MCMC routine.

2.2.3 NN/MCMC Calibration Routine

Figure 2.4 illustrates the methodological flow-chart for parameter calibration using the above described neural networks.

1. A prior distribution is defined for use in selecting possible sets of model parameters.

2. This prior distribution is used to create an initial ensemble of model runs.

3. Data from this initial ensemble is used to train neural networks to predict the resulting model output for an input parameter set.

4. MCMC methods are used to sample from the posterior distribution, which is calculated using a given likelihood function, the prior distribution, and observations. Neural networks are used to simulate model response to different parameter sets.

5. Once the Markov chains converge, i.e. remain in one general area of the search space, this portion of the search is sampled from in order to generate the posterior distribution for parameter selection. This information is used to create a new prior distribution and the process is iterated until the required (or computational feasible) degree of convergence of the posterior distribution is attained.

The original prior distribution is typically either chosen to be uniform over the possible range of parameter values, or else is described based on external knowledge of the system. The exact formulation of the likelihood function varies with respect to the amount of information available. If the observational data consists only of single measurements with predetermined noise levels, then the likelihood function may just be Gaussian. If more detailed statistics on the observational data are available, the likelihood function can be more complex.

In the following experiments the search routines employed for step (4) above are derived from a method known as slice sampling. The method samples an n-dimensional distribution by selecting an n dimensional slice from underneath the surface of its density function and then sampling uniformly from this slice. Because the shape of the density function will determine the size of slice sampled from (i.e. smaller slices will result from the more sharply peaked regions of the function) given enough iterations the density of the collected samples should converge to the density prescribed by the density function. Slice sampling posseses a similar advantage to Gibbs sampling in that the routine for sampling each element member is determined almost entirely by the other elements of the distribution and so requires minimal external tuning. Slice sampling, unlike Gibbs sampling, does require setting some ex-

ternal parameters that control the sampling of the slice, as the reason that sampling is being employed is that the shape of the slice is not explicitly known. However, unlike Gibbs sampling there is no need to prescribe conditional distributions between the elements of the distribution, allowing it to be applied to more general applications, such as the above described calibration routine. A thorough discussion on the theory and possible variations of slice sampling can be found in (Neal 2000).

Chapter 3

Experiments with Budyko Energy Balance Model

3.1 Outline of Model

The one dimensional Budyko Energy Balance Model approximates mean annual latitudinal temperatures by balancing incoming and outgoing radiation values for the planet. It does this by considering the earth to be a black body and by approximating albedo and energy transport on the earth's surface. The numerical model used for the following experiments is available on-line at http://www.phys.uu.nl/ nvdelden/EBM.html and further details about the model can be found in (van Delden 2008).

The earth is assumed to be in energy balance; i.e., the amount of energy entering the system as solar radiation, Q, must equal that being emitted by the system as long wave radiation, I. The model assumes equilibrium energy balance for each latitudinal

grid band, ϕ, with albedo $\alpha(\phi)$ and latitudinal energy transport $A(\phi)$ as:

$$Q(\phi)(1-\alpha(\phi)) = I(\phi) + A(\phi)$$

.

To express I as a function of temperature in degrees Celsius, T, one can use the Stefan-Boltzman law and the binomial theorem to write

$$I = \sigma(273.15+T)^4 \approx \sigma(273.15)^4 + 4\sigma(273.15)^3 T = I_0 + bT.$$

This would suggest that $I_0 = 320.64 W/m^2$ and $b = 4.69 W/m^2 C$ but empirical tests have suggested that the approximation is better served by $I_0 = 205 W/m^2$ and $b = 2.23 W/m^2 C$ (van Delden 2008). As I is the outgoing long wave radiation at the top of the atmosphere, and since it is most informative for T to represent temperature at sea level, the equation is further adjusted as

$$I = I_0 + b(T - h\Gamma)$$

where h is height above sea level and Γ is the temperature lapse rate.

To express A as a function of temperature the Budyko parametrization

$$A = \beta(T - T_p),\ T_p \equiv \frac{1}{2}\int_{-1}^{1} T\,dx,\ x \equiv \sin\phi$$

redistributes heat based on the difference between the temperature of a given location and that of the mean sea-level temperature, T_p. This is parametrized by a relaxation coefficient, β.

The albedo α is also given T dependence. At or below a set threshold temperature, T_0, it is assumed that the surface grid cell is ice covered and therefore assigned a high albedo α_0. For temperature T_1 and above, the grid cell ϕ is considered ice free and

assigned a low albedo, α_1. For temperatures between T_0 and T_1 the albedo is adjusted proportionally. This is implemented as follows:

$$\begin{aligned} \alpha &= \alpha_0 && \text{if } T \leq T_0 \\ \alpha &= \alpha_0 + \tfrac{T-T_0}{T_1-T_0}(\alpha_1 - \alpha_0) && \text{if } T_1 \geq T\, T_0 \\ \alpha &= \alpha_1 && \text{if } T \geq T_1. \end{aligned}$$

Based on the above discussion, the tunable parameters for the model include the terms I_0, b, α_0, α_1, T_0, T_1, and β. The term 'tunable' denotes these values as being approximations of unresolved systems, as discussed in the first chapter, rather than physical properties, and thus it is appropriate to calibrate them against observed conditions. Figure 3.1 shows the model output for the listed initial parametrization, as seen in Table 3.1, compared against the mean of zonally averaged temperatures from the past sixty-one years. The plotted observational data is taken from (Kalnay et al. 1996), and the model output is interpolated to match the locations of the observational data points. The uncertainties for the observational data set are approximated by combining the mean for each location of the inter-annual anomalies over the time span and the average difference between these and temperature anomalies obtained from a separate reanalysis project (Reynolds et al. 2002).

The general shape of the curves are similar, suggesting that the model is effective in describing some of the nature of latitudinal energy transport over the planet's surface. However the temperature values of the model are warm-biased relative to the observational climatology. Furthermore, in contrast to observations, the model computes the earth's temperature to be symmetric across the equator. This is an issue of model design that cannot be addressed through calibration. How the calibration methodologies cope with this "can't win" situation adds an interesting dimension to

Table 3.1: Initial EBM Parametrization

I_0	205.0
b	2.23
α_0	0.62
α_1	0.25
T_0	263.0
T_1	273.0
β	3.8

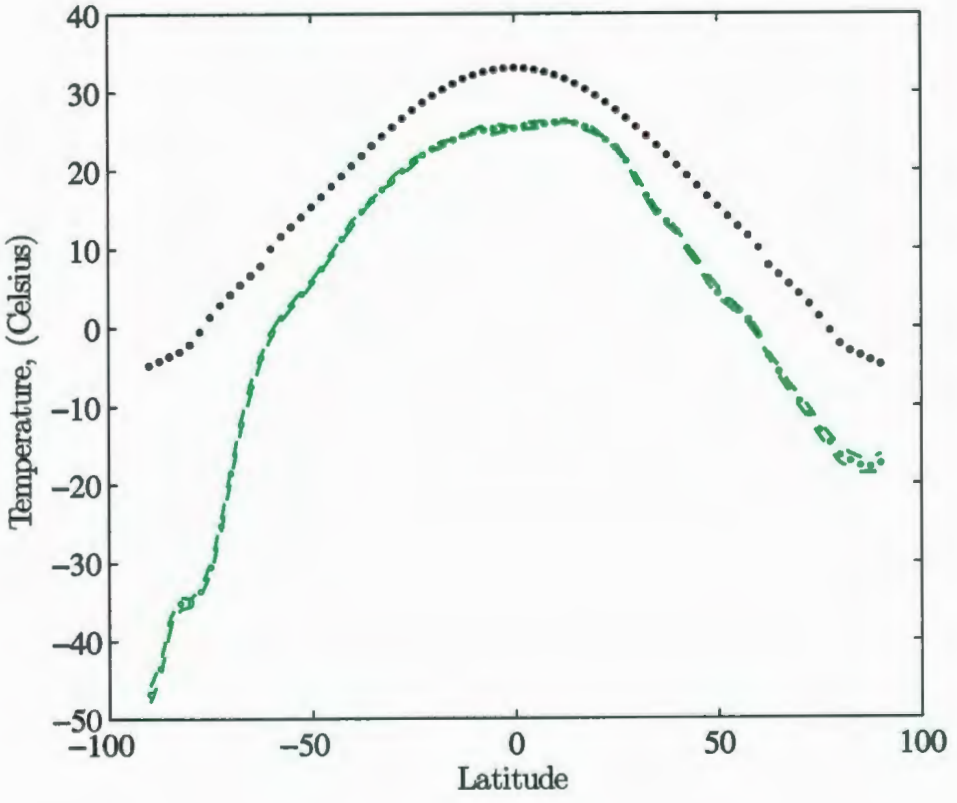

Figure 3.1: Model results (black dots) compared to observational data (green dots, dashed lines represent observational uncertainties).

the experiment.

3.2 Calibration using the EnKF

Table 3.2: EnKF Settings for EBM calibration

Number of iterations	30
Number of ensemble members	800
Size of forecast space	$161 + 7$
Size of observation space	73

Table 3.3: Prior distributions for EBM calibration

I_0	$U(104.0, 304)$
b	$U(0.03, 4.43)$
α_0	$U(0.45, 0.85)$
α_1	$U(0.05, 0.45)$
T_0	$U(248, 278)$
T_1	$U(263, 283)$
β	$U(1.8, 5.8)$

The settings for the EnKF routine used for the calibration of the EBM are given in Table 3.2. Further explanations of their function within the routine are given in the appendix. An ensemble of eight hundred members was needed to thoroughly explore the wide ranges allowed for the prior distribution, displayed in Table 3.3. This means

that in the course of the calibration routine the model was run 24800 times. This is not realistic for more computationally demanding applications, but possible here due to the short model run time. The forecast space consists of 161 data points defining the state space and is augmented to include the seven parameters. The model is calibrated against the seventy three points available from the observational data described above. The observation perturbation for each data point for each iteration is taken from the measurement uncertainty statistics also presented above. The routine is iterated thirty times, using the previous state as the initial state for the next iteration. This produces, counting the initial ensemble, thirty one separate manifestations of the forecast state.

The state space produced by the final iteration of the EnKF routine is shown in Figure 3.2. Aside from the polar region, the calibrated model produces a close fit to observations. The ensemble has converged (i.e. all ensemble members produce very similar outputs) to the point that the standard deviations are indistinguishable at the scale presented. This extensive convergence is also noticeable in Figures 3.3 and 3.4. By iteration ten the parameter values of all the ensemble members have become very similar; however, it is not until iteration twenty-five that the mean values of the ensemble become static. Figures 3.5 and 3.6 display the prior distributions from Table 3.3 compared with the final posterior distributions produced by the EnKF routine. As the EnKF assumes that all uncertainties are Gaussian the posteriors are presented as such.

As Figure 3.2 is the result of a combination of this evolution of parameter values and active data assimilation, it does not isolate the improvement gained by the generation of new parameter sets. To properly asses the improvement a new ensemble,

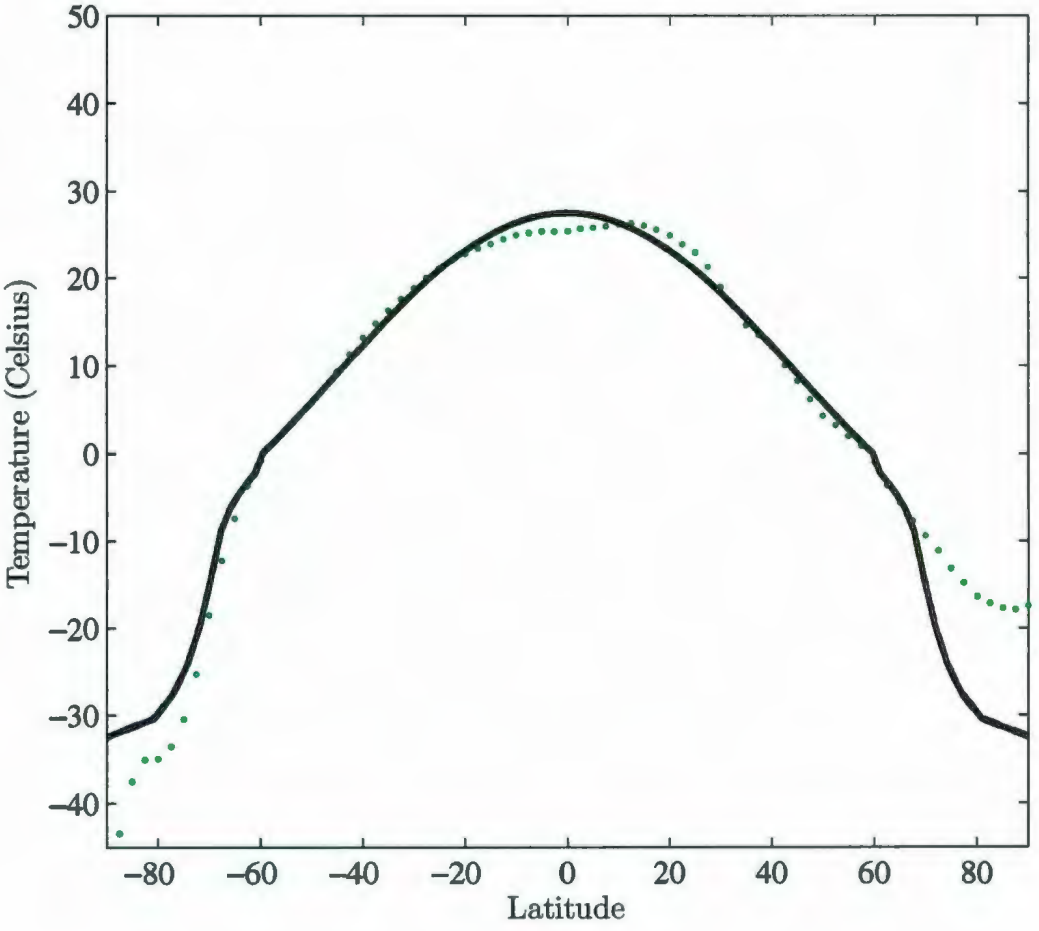

Figure 3.2: State space results from final iteration of the EnKF routine, ensemble mean (black line) compared to observational data (green dots).

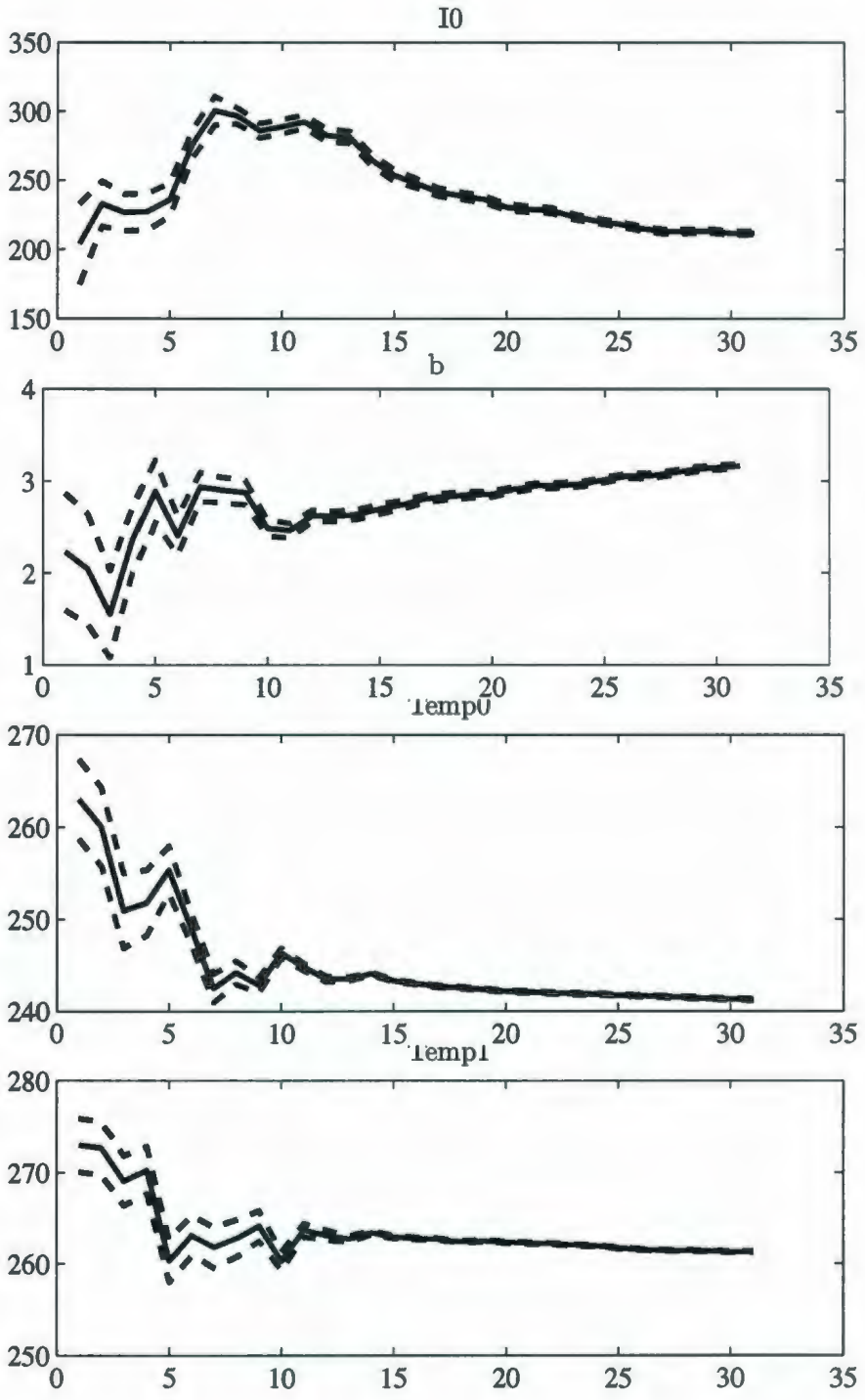

Figure 3.3: The evolution of ensemble parameters,while iterating the EnKF routine, displayed as ensemble mean (black line) and standard deviation (black dash).

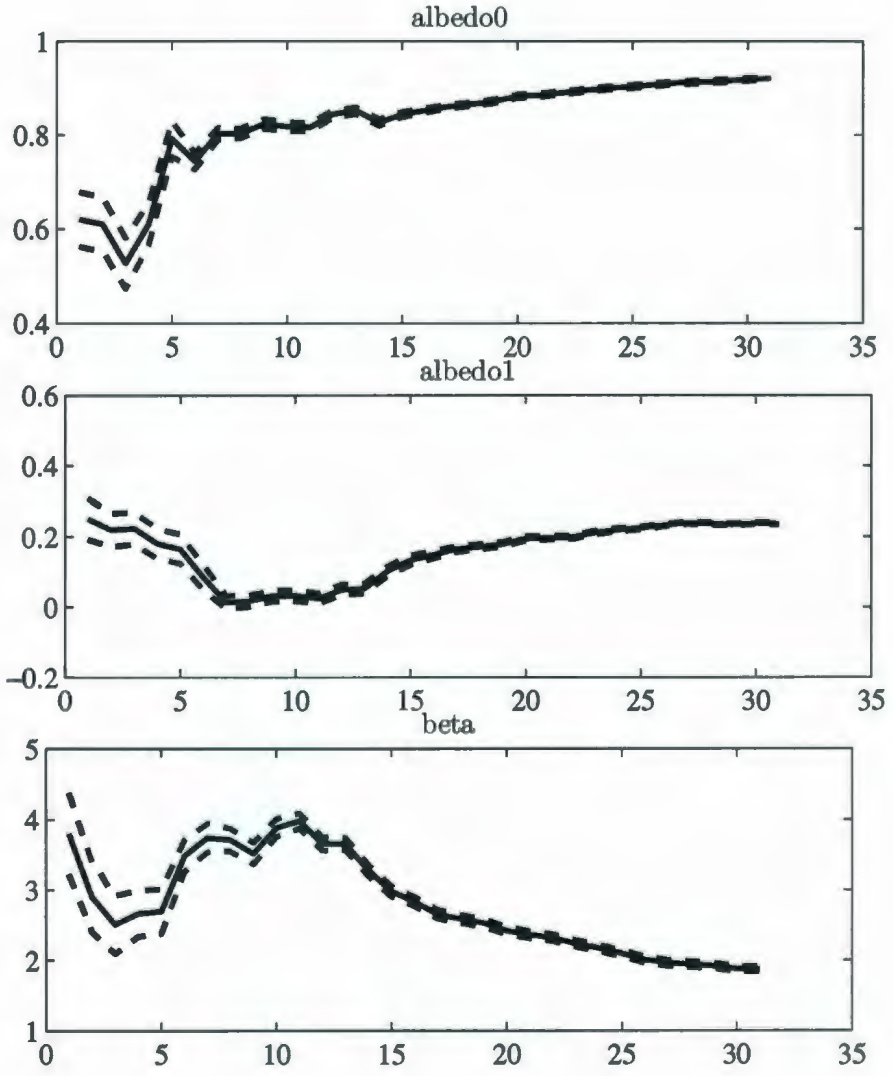

Figure 3.4: The evolution of ensemble parameters,while iterating the EnKF routine, displayed as ensemble mean (black line) and standard deviation (black dash).

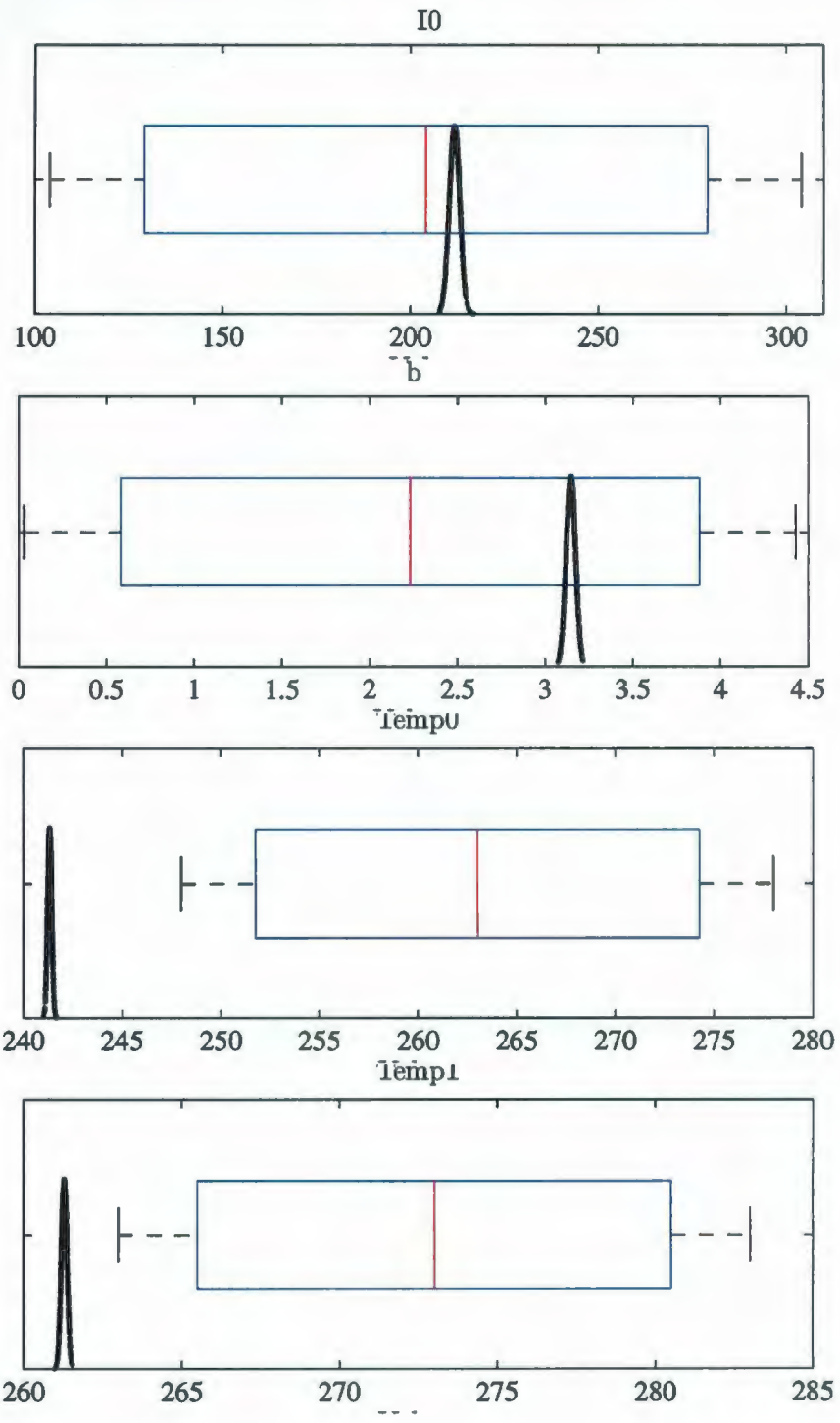

Figure 3.5: Prior (uniform distributions, shown as box and whisker plots) and final posterior (shown as Gaussian bell curves) for the EnKF calibration of the EBM.

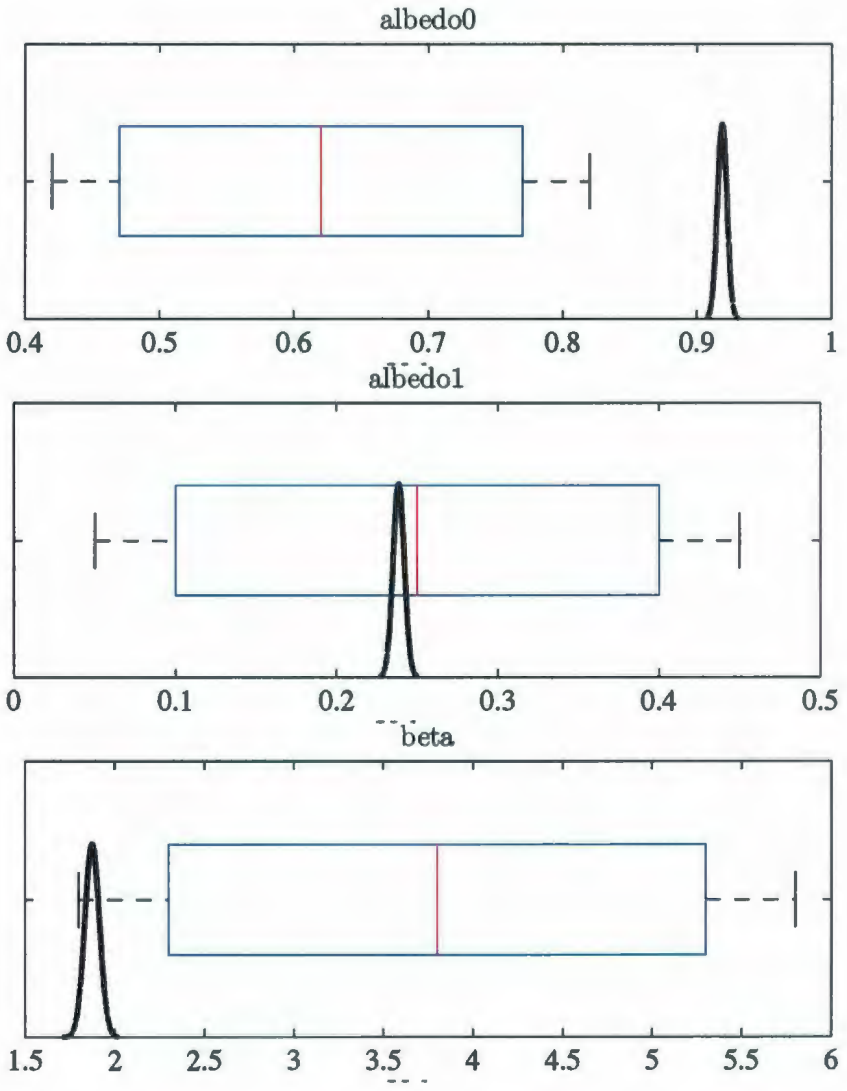

Figure 3.6: Prior (uniform distributions, shown as box and whisker plots) and final posterior (shown as Gaussian bell curves) for the EnKF calibration of the EBM.

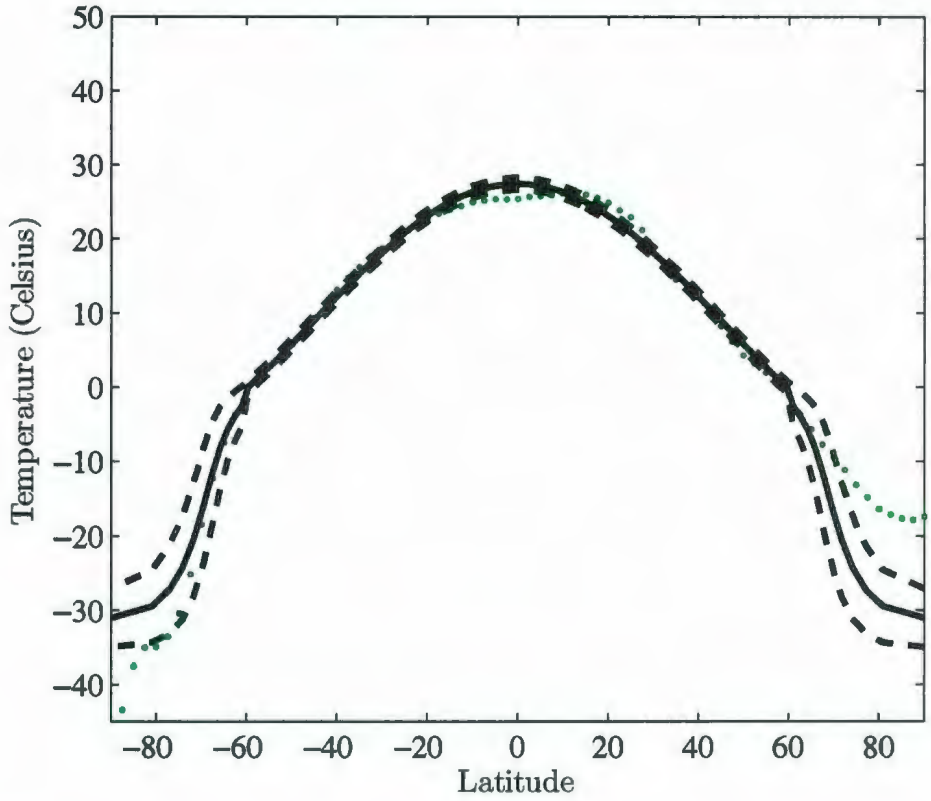

Figure 3.7: The mean (black line) and standard deviation (black dash) produced by creating a model ensemble by selecting parameter values from the EnKF created posterior distributions, compared to observational data (green dots).

also of size eight hundred, was produced by selecting parameter values from Gaussian posterior distributions prescribed by the final iteration of the EnKF routine. The results can be viewed in Figure 3.7. While the mean result is similar to that produced by the EnKF data assimilation, this figure gives a much clearer view of where vagueness in the parametrization translates into forecast uncertainty.

3.3 Calibration using NN/MCMC

Table 3.4: Architecture of neural network used for the EBM calibration routine

Input layer	size 7
Hidden layer 0	size 96
Output layer	size 7
Training ensemble	size 10000

In order to obtain training data on the model response to different parameter sets, an ensemble was generated using a collection of parameter sets created by a Latin Hypercube sampling of the space defined by the prior distribution. The architecture of the network used is given in Table 3.4. The network inputs were the seven parameter values used to generate each ensemble member. A network with a single hidden layer containing many units performed best in testing and so was selected to act as the emulator for the calibration routine. The choice of ninety-seven hidden units created a network with a total of 1344 weights to evaluate. This maintained the general practice of keeping the total number of weights below the quantity of available training data.

Training a network to emulate model predictions for all seventy-three data points involved very lengthy computation times when compared to the time requirements for the EnKF routine. Furthermore, given the high correlation between adjacent grid bands, it makes no sense to calibrate against the whole temperature field. The network emulator was therefore trained to only seven of these data points to produce a more even comparison of method performance under equal resource capacity. The grid bands selected were the locations 90 °S, 60 °S, 30 °S, 0 °, 30 °N, 60 °N, and 90 °N.

The trained network was incorporated into the likelihood function used to evaluate model performance for a given parametrization. For this experiment the likelihood was defined to be equal to $N(y|f(\theta), \sigma)$, where $f(\theta)$ is the neural network prediction for model output given a parameter set θ, and σ is defined to be the set of prescribed uncertainties in the data. The sigma values for the given data points are therefore the same observational uncertainties used for the EnKF routine described above. The likelihood function generates the posterior distribution for model fit to observations given the same uniform prior described in Table 3.3. The posterior distribution is explored through slice sampling, as discussed in Chapter 2. Here the initial step size for each parameter is given as one third of the range of the prior distribution range. Because Markov Chain Monte Carlo methods are susceptible to trapping in local minima, various sampling chains were run from different random initial points. When the behavior of these chains stabilized, the stable portion of the chains was sampled at an interval of every second element. These initial chains had stabilized by around fifty iterations. The statistics generated from these samples were used to generate the quantiles used to define prior distributions for the next round of sampling.

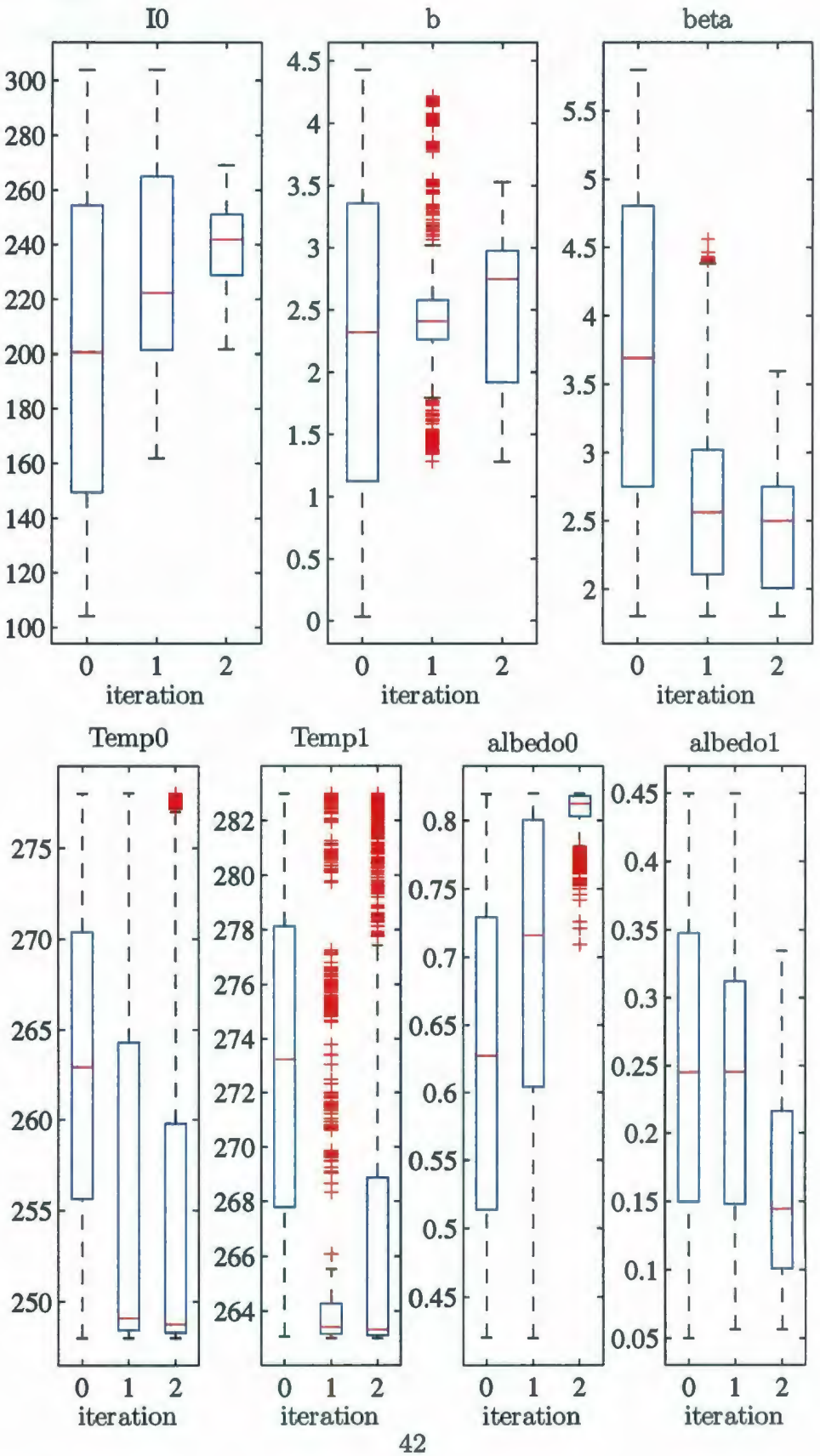

Figure 3.8: Distributions of parameter sets for the EBM produced from the prior, then two sequential executions of MCMC posterior sampling.

The acquired samples were then used to retrain the network for the second round of sampling. As this ensemble was smaller than the first, having 2500 members, the network size was reduced as well. The same architecture was maintained, except with a smaller hidden layer of seventy two units.

The initial slice sampling step sizes for each parameter were adjusted to be the standard deviation of their previous posteriors. The Markov chains produced by this arrangement required about two hundred iterations to stabilize and in many cases produced only a modest reduction of distribution length, if at all. The progression of the selection distributions from the prior to the final posterior is shown in Figure 3.8. In order to assess the results of this calibration routine, an ensemble was generated from the final posterior distribution in the same fashion as with the EnKF. The results are displayed in Figure 3.9.

3.4 Discussion

It is tempting to look Figures 3.7 and 3.9 and conclude quickly on the results of the experiment. The EnKF seems to have performed very well, with a close fit to the observations for the equatorial and middle latitudes, and an increase in prediction uncertainty near the poles, i.e. the regions the model was previously known to describe poorly. The NN/MCMC result certainly appears more "awkward" and indicates a much higher level of uncertainty in the generated forecasts. The resulting posterior distributions from the EnKF routine represent a clear and dramatic focusing of the prior space. Results for the NN/MCMC produced posteriors appear mixed.

There are many considerations that must be addressed, however, before drawing

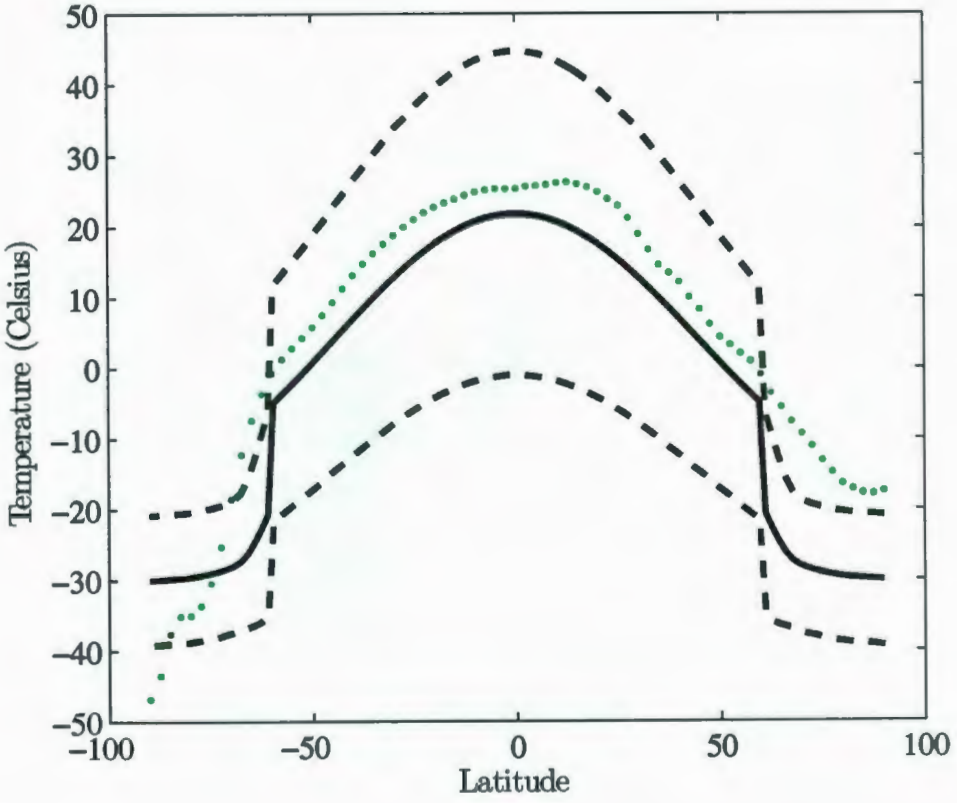

Figure 3.9: The mean (black line) and standard deviation (black dash) produced by creating a model ensemble by selecting parameter values from the neural network emulator-derived posterior distributions, compared to observational data (green dots).

conclusions from the results presented. Most striking are some of the resulting posterior distributions produced by the EnKF. A result indicating within a very small uncertainty range a greater than 90% albedo for snow is not physically accurate. The temperature thresholds for snow covered or non-snow covered surfaces are also far too low. All three of these parameter distributions fall beyond the range of their prior. This does not occur in the neural network calibration routines, as there the prior is mathematically combined with the likelihood function. This results in any value falling outside the prior receiving a null probability. For the EnKF the prior is simply used to generate the initial ensemble. Furthermore, the filter assumes the distribution of the initial ensemble members to be Gaussian in its subsequent calculations. Methods of rejecting the EnKF outputs outside the bounds prescribed by the prior are possible. Often however, these will only stall the routine unless the filter is employed with an ensemble size or observation permutation that is larger than is typically realistic. An artifact of this situation is the sharp temperature gradient around 55° North and South, observable in the NN/MCMC solution, Figure 3.9. The in both the EnKF and NN/MCMC calibrations extreme values for albedo were favoured (suggesting that the prior range was naïly set). In the EnKF solution this is compensated for by setting the temperatures at which land is considered ice covered unrealistically low (i.e. the parameters $Temp0$ and $Temp1$). This was not an option in the NN/MCMC solution due to restrictions set by the prior distribution for these parameters, although note that the posterior distribution favoured low values for these parameters, resulting in the observed physically unrealistic gradient in the resulting EBM output.

Also suspicious is how the EnKF calibration results in a model that favours the

data from the southern polar regions over the northern. As the model is symmetrical about the equator, it must be matched to one criterion or the other, either option creating the same degree of mismatch to the overall data. The bias of the calibration towards the southern data points indicates that the EnKF ensemble converged around a local minimum. The NN/MCMC result also provides a poor fit about the poles. However, the degree of misfit in north and south is more balanced and therefore more representative of the model's bimodal fit to the data. The misfit is also a partial result of the reduced number of observations provided to the NN/MCMC. Representing each polar region by a single point was clearly inadequate for the task of creating a complete view of model performance. What makes the choice of output subspace justifiable is the prior knowledge that the model can not resolve the difference between polar temperatures.

The forecast uncertainties produced by the EnKF are also very narrow. On average, the standard deviation of the ensemble state space is $0.02\,^{\circ}C$. This is much smaller than the natural variability of zonal temperatures, and also provides a much narrower error bar than would seem appropriate for as simplified a model as the one utilized. In contrast, by the same standards, the standard deviations of the NN/MCMC ensemble seem excessively wide. However, it can be seen that from 60 °S to 60 °N the ensemble mean is within one standard deviation of the observed state; this cannot be said for the EnKF forecast. The wider ensemble spread produced by the NN/MCMC routine reflects the (on average) wider posteriors generated by this method. These more complex distributions better reflect the general formulation of the mathematical model, and its bimodal fit to real world observations. As the EnKF is limited to assuming Gaussian distributions it is forced to create the

best fitting distribution of this form, rather than mirror the actual posterior. However, a contributing factor to the wide spread of the first standard deviation for the NN/MCMC result for the tropics and equatorial regions is likely an artifact of the ensemble containing members that have been biased by their parameter sets to warm or cold extremes so as to fit one or the other of the polar regions. This misrepresentation of the region that is poorly resolved by the model does not, at least qualitatively, appear in the EnKF solution.

This initial experiment with the EBM has given an interesting characterization to both methodologies. For a computationally undemanding model that produces a large number of outputs, the EnKF provides an efficient method to create good fits to the calibration data. However it appears that many of the EnKF results must be considered within the context of the method, rather than as definitive results about the model. The NN/MCMC sampling method shows a potential to be more informative, even for simpler models. The added computational investment of training the model emulator is rewarded by the ability to further investigate the posterior space. However, resource limitations can require the creation of an emulator of limited accuracy or else one trained to a subsection of the model output. In these circumstances care must be taken to ensure that the network provides a sufficiently robust emulation of the model behavior to justify a detailed investigation into the model responses it simulates.

Chapter 4

Experiments with Planet Simulator

4.1 Overview of Experiment

The Planet Simulator is an intermediate complexity General Circulation Model (GCM) developed by the Meteorological Institute of the University of Hamburg. This model is based on the primitive equations (i.e. the basic conservation laws) and incorporates a slab ocean model with sea ice (Lunkeit et al. 2007). For the purposes of this study, it is run at low T21 L5 resolution. The model is forced with observed annual atmospheric CO_2 concentrations from the years 1958 - 2008 (Tans 2009). The Planet Simulator is run for the full cycle of fifty model years with this forcing, plus a ten year initial spin up cycle, the results of which are discarded. Run times varied between platforms and model setup, requiring on average forty eight hours for each sixty year cycle. The model is calibrated against seasonal surface temperatures climatologies for the period of 1958 - 1968 and the period of 1999 - 2008. Figure 4.1 displays the calculated difference in surface temperature between the two periods.

This data is taken from the same source as in Chapter 3 and the uncertainties used in the calibration routines are calculated in the same manner. It is clear that the largest change between the two climatologies are in the polar regions. However, these regions are given low weighting in the model calibration as observations are limited for these regions, making the reanalysis results more suspect than is perhaps captured by the uncertainty calculations. Also it is believed that the Planet Simulator does not accurately describe the southern pole due to the simplicity of its ocean transport model.

The Planet Simulator incorporates many parametrized sub-processes with tunable constants. For the initial experiments presented here, five parameters were chosen for use in the calibration procedures. An effort was made to select parameters representative of a variety of physical processes. This is consistent with the view that calibration is not being used here to refine a particular area of model physics, but rather as an attempt to view unresolved processes as interdependent elements of a non-linear dynamic system. Sensitivity tests were used to select calibration parameters and set their respective prior ranges (i.e. for a uniform prior distribution). These parameters and ranges are listed in Table 4.1 and are referred to throughout this work by the labels given in the Planet Simulator code and documentation (Lunkeit et al. 2007) to facilitate independent investigation. The parameter *acllwr* is a coefficient representing liquid mass absorption in clouds, in the context of an equation approximating the Long Wave Radiation (LWR) flux permitted by different levels of cloud cover. The parameter $vdiffk$ is used in the calculation of the ocean vertical diffusion coefficient for the three layer slab ocean model. In the parametrization of atmospheric horizontal diffusion the damping time scale is linked to a time scale for

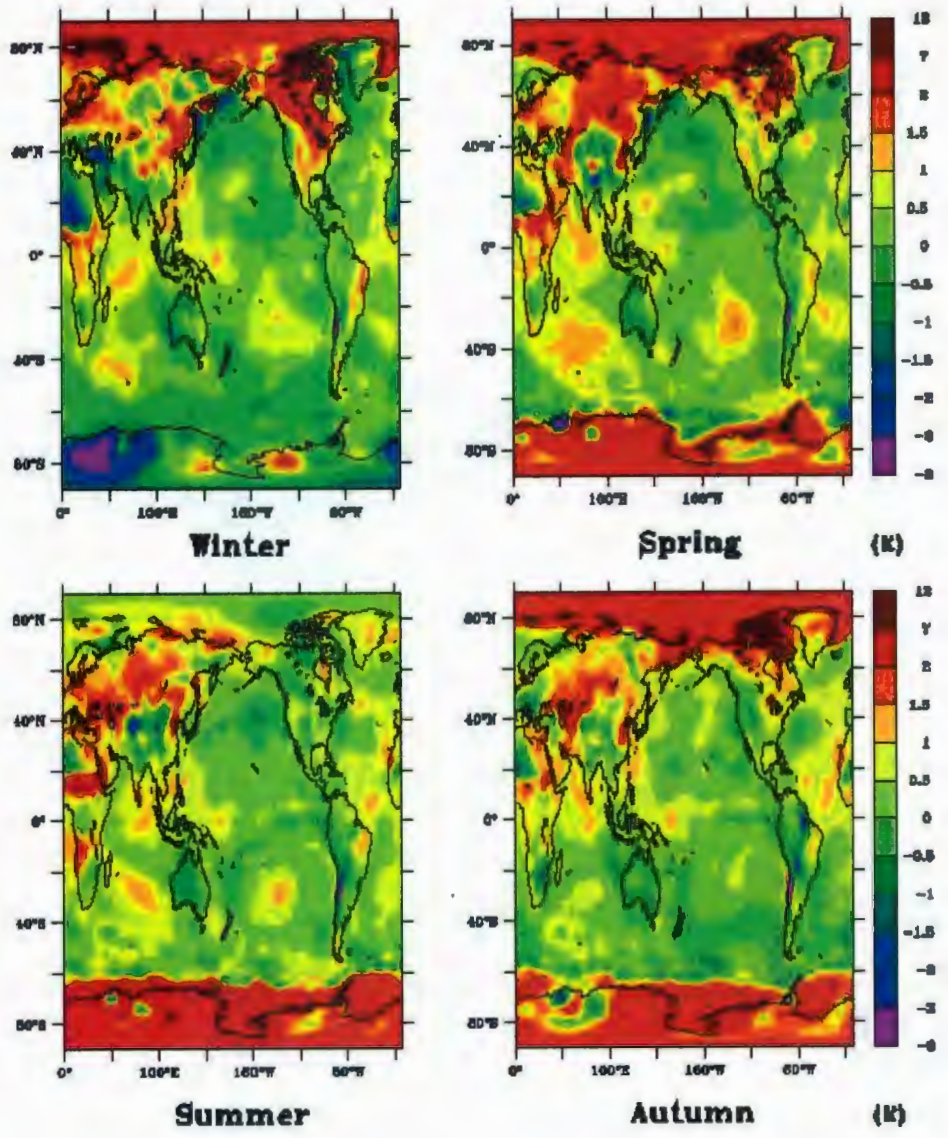

Figure 4.1: Observed difference in mean seasonal surface temperatures (in degrees Celsius) for DJF (upper left) MAM (upper right) JJA (lower left) and SON (lower right) between 2008-1999 and 1959 - 1968.

divergence *tdissd*. The calculation of cloud transmissivity for visible and ultraviolet Short Wave Radiation (SWR) accounts for back scatter through the product of the solar zenith angle and a constant $tswr1$. Atmospheric turbulent exchange is approximated as vertical diffusion relating to wind, temperature, and specific humidity. The equations for calculating exchange coefficients for momentum and heat include the parameter $vdiff_{lam}$ which adjusts the mixing length terms in these equations. A technical overview of the role of the particular constants investigated in the model equations can be found in (Lunkeit et al. 2007).

Table 4.1: Investigated Parameters and their Priors

acllwr	$U(0.05, 0.2)$
$vdiffk$	$U(1e10^{-5}), 1e10^{-3})$
tdissd	$U(0.05, 0.8)$
$tswr1$	$U(0.02, 0.08)$
$vdiff_{lam}$	$U(80, 320)$

Figures 4.2 - 4.5 display results of model sensitivity tests for what became the selected parameters excluding the ocean diffusivity term. These maps show the difference between the mean annual surface temperature of a three year model run and the results of the same model run with the parameter value set to the upper and lower extremes of its prior distribution. For the sensitivity tests only one parameter was varied at a time. The temperature differences in figures 4.2 - 4.5 were considered large enough to make the parameters viable candidates for the calibration experiments. Ocean diffusivity was excluded from sensitivity testing as it was accepted into

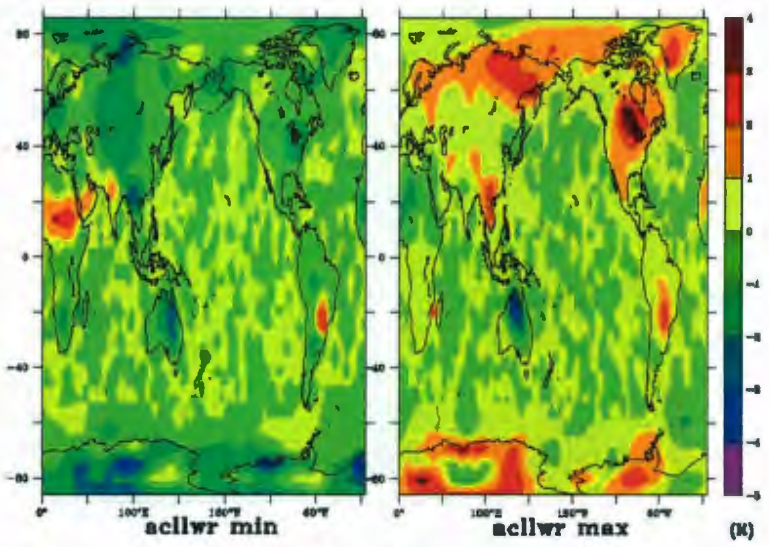

Figure 4.2: Sensitivity testing of parameter "*acllwr*".

the set of parameters on the basis of it being the only acceptable ocean related model parameter. Initial tests did not reveal discernible patterns of regional variance for this parameter. It was assigned the wide prior necessary for it to capture a degree of mean temperature field variation near to that of the other parameters.

Even at course resolution performing either calibration method utilizing every data point produced by the model is not feasible nor sensible. Instead calibration data was constructed by taking average seasonal surface temperature for regions of approximately 1000km × 1000km, centered on the geographic centers of the North Atlantic, South Atlantic, Indian Ocean, North Pacific, South Pacific, North America, South America, Europe, and Siberia. Models were calibrated against the mean seasonal temperature climatologies at each region for the first and last ten years of

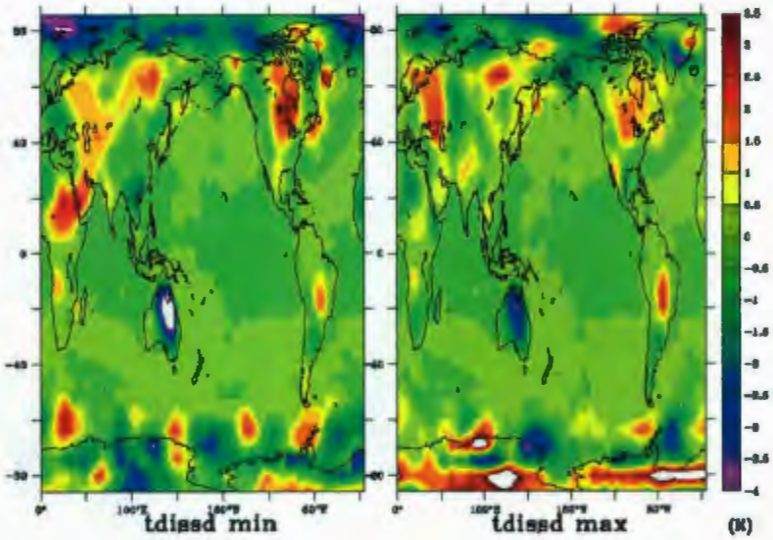

Figure 4.3: Sensitivity testing of parameter "*tdissd*".

the fifty year model run, therefore requiring a calibration data set of seventy two elements. A 40 year climatology separation is a short interval for a transient calibration, but arguably still better than the traditional target of a single climate state.

4.2 Calibration using the EnKF

Table 4.2: EnKF Settings for Planet Simulator calibration

Number of Iterations	1
Number of Ensemble Members	200
Size of Forecast Space	$72 + 5$
Size of Observation Space	72

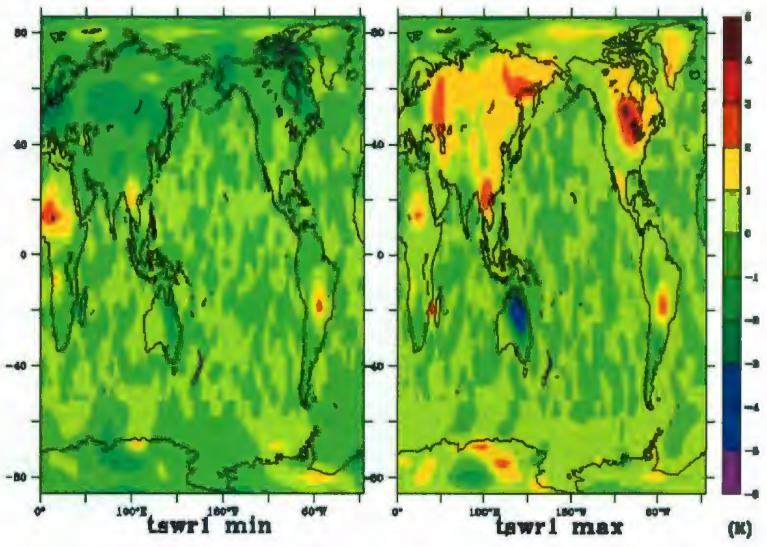

Figure 4.4: Sensitivity testing of parameter "$tswr1$".

The settings for the EnKF routine used for the calibration of the Planet Simulator are given in Table 4.2. An ensemble of two hundred members is the order of magnitude upper limit of the number of model runs that could be performed for a current generation GCM with the resources typical of a modern modelling center. Early experiments with ensembles of quarter or halve this size did not produce viable results. These resulted in a "collapse" of the ensemble to a single manifestation. Because of these concerns as well as computational demands of the model the iteration of the routine was not automated. Rather after each iteration the result was analysed to determine the desirability of performing a further iteration.

Figure 4.6 displays the distribution of the original parameter sets selected from the prior along with the distribution of parameter sets created by iteration of the

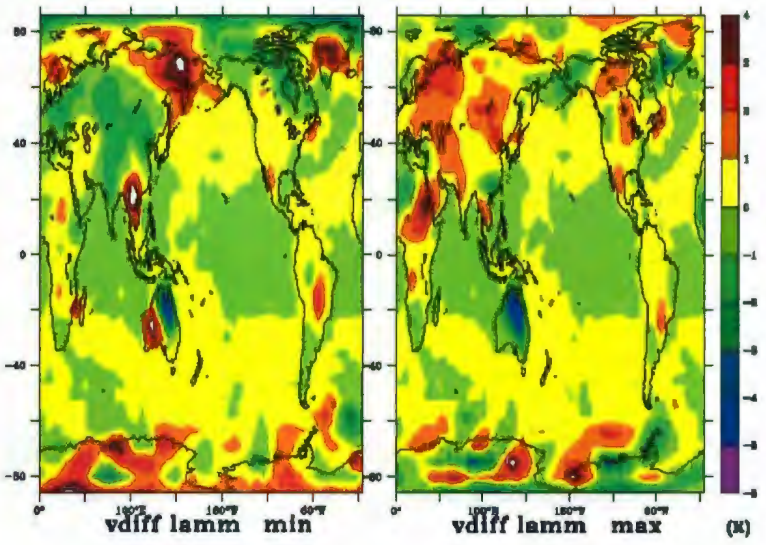

Figure 4.5: Sensitivity testing of parameter "$vdiff_{lamm}$".

EnKF analysis routine. It is immediately visible that the distribution of the "acllwr" parameter within the sampled parameter sets has all but collapsed to a null value. Over a quarter of the sets have placed the "vdiffk"to null as well. As null or negative parameter values are nonsensical for this application the analysis routine is artificially instructed to reject such values, resulting in a very limited number of usable parameter sets.

That the observed collapse is to a lesser extent than to those of smaller ensembles, is suggestive that the increasing of ensemble size is improving performance. Comparison of the analysed state vector to that of the calibration data shows a larger mismatch than that displayed by the original state vector. A situation where the ensemble statistics create a weighting scheme that heavily favours the model predictions

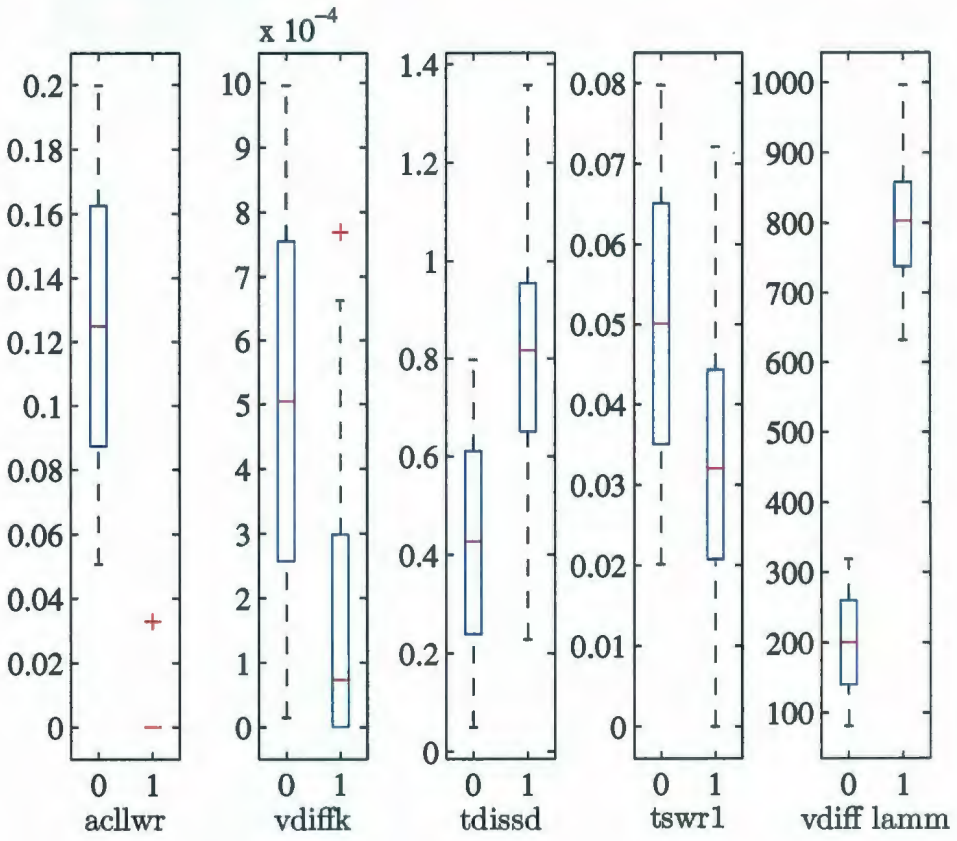

Figure 4.6: Distributions of individual parameter values from the prior and first iteration of the EnKF analysis routine.

over the influence of the observations suggests that the initial ensemble variance was too small. However, considering the model's computational demands and the performance of the EnKF in a previous experiment documented in Chapter 3, a further increase of ensemble size could not be justified. No further experimentation has been performed with the EnKF for this application.

4.3 Calibration using NN/MCMC

4.3.1 Calibration Routine

As in the experiment outlined in Chapter 3, the first step in this calibration routine is to create an ensemble of model runs with parameter sets selected from the prior through Latin Hypercube sampling. This ensemble provides the initial training data for the neural networks. An ensemble of two hundred members was used for this experiment, which is notably smaller than the ensemble used in Chapter 3.

Table 4.3: Architecture of Neural Networks used for the Planet Simulator calibration routine

Network label	A	B	C
Input layer	size 5		
Hidden layer 0	size 13	size 19	size 17
Hidden layer 1	size 6	–	size 4
Output layer	size 8		
Training ensemble	size 200		
Number utilized	2	3	4

The experiment with the Planet Simulator GCM creates a far more complex relationship between model parameters and model output than was the case of the EBM of Chapter 3. Here multiple networks were needed to successfully approximate model response. Nine networks were used, one for each calibration data location. Initially three networks were trained to each location, their architectures are labeled A,B, and

C and Table 4.3. For each location the network architecture that resulted in the best emulation of the training data was selected to represent the location in the calibration procedure. As previously discussed the resistance of the Bayesian networks to over fitting makes this an acceptable if not ideal selection criteria. Future work will examine in detail the predictive value of various geometries. All the networks have as their inputs the five parameters discussed above. As each network is trained to express all the data of a specific region each produces an eight member output vector, i.e. the two target model temperature values for each of the four seasons. Figure 4.7 depicts the degree of fit obtained by the poorest performing network, which modeled the Siberian region, while Figure 4.8 depicts the degree of fit obtained by the best performing network, which modeled the South Pacific region. The mean correlation value between the produced networks and their training data was 0.998. Networks with hidden layers up to three times larger than those displayed in Table 4.3 were tested, but these only increased computation time without improving fit.

The trained neural networks were incorporated into the same MCMC sampling routine outlined in Chapter 3. In keeping with the outlined NN/MCMC calibration procedure, the results of this sampling were used to create a new model ensemble. As the posterior distribution obtained from the first iteration of calibration routine had narrowed from the prior it was judged appropriate to limit this ensemble to one hundred members. Data from this ensemble was then used to provide an independent test set for the neural networks from the first iteration. Figure 4.9 depicts the degree of fit obtained by the poorest performing network, which modeled the South Atlantic region, while Figure 4.10 depicts the degree of fit obtained by the best performing network, which modeled the North American region. It is immediately obvious that

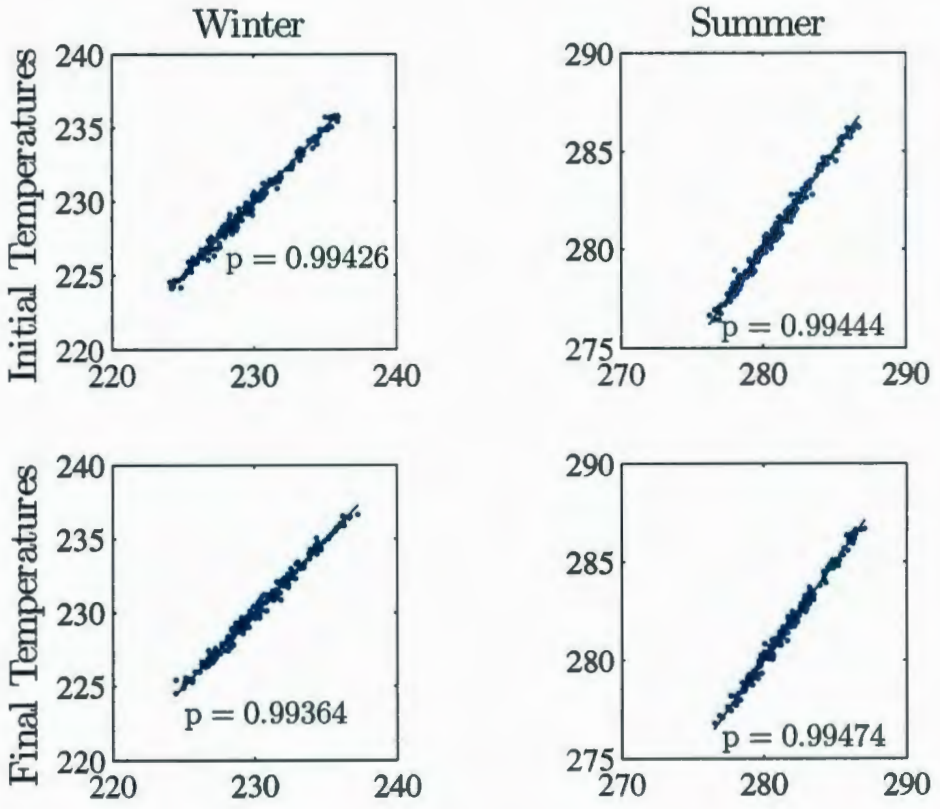

Figure 4.7: Fits between training data (x-axis) and network predictions (y-axis) for the neural network simulating model output for the Siberian region for the winter and summer seasons.

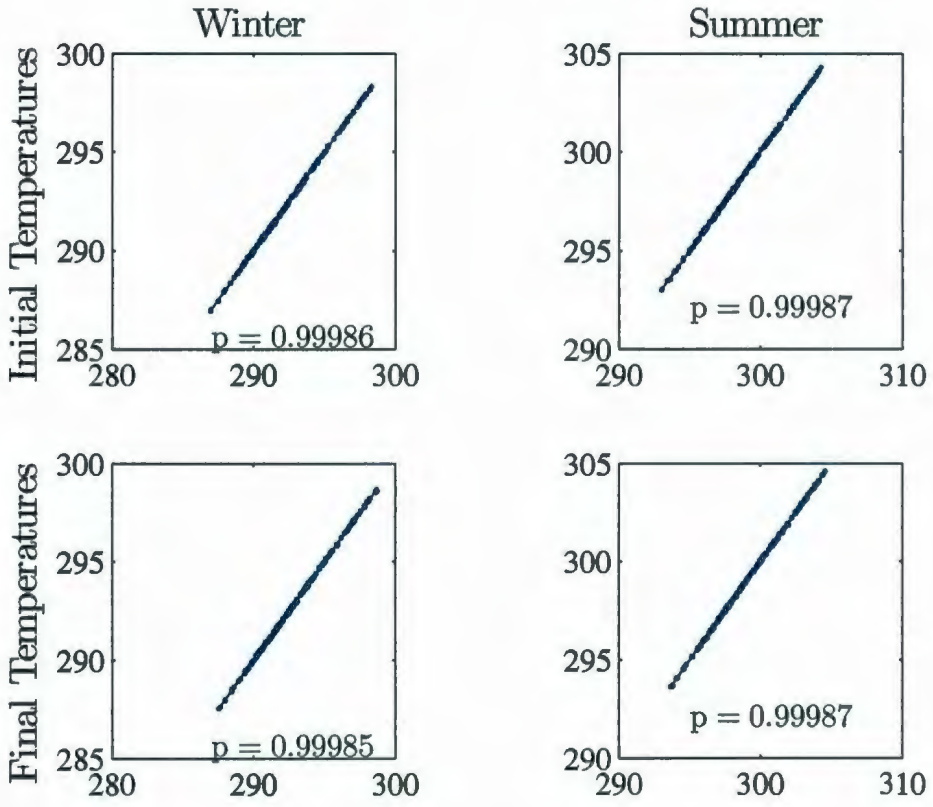

Figure 4.8: Fits between training data (x-axis) and network predictions (y-axis) for the neural network simulating model output for the South Pacific region for the winter and summer seasons.

the emulation skill of these networks is weak. The mean correlation value between the produced networks and test data from this sub-region of the original sample space data was 0.6267. For the second iteration, the original network training data is augmented with the data from the new ensemble, and the networks are retrained. A second sampling iteration is performed with the retrained networks utilized in the MCMC routine. The prior distribution for the sampling routine is also altered to be the distribution produced by the previous iteration. The evolution of the posterior distributions (shown with respect to the individual parameters, rather than the actual 5-dimensional space) from the prior through the two MCMC iterations is displayed in Figure 4.11.

4.3.2 Results Produced by the Calibration Routine

Figure 4.12 displays the benefit of creating neural networks to allow much broader sampling of model response. Here the minus log-likelihood of each ensemble member given the calibration data has been computed for both the initial ensemble and the ensembles produced by each iterations of the MCMC sampling. The log-likelihood for the output produced by the default model parametrization is also included. These values are calculated with the same likelihood function used in the calculation of the Bayesian posterior during the MCMC sampling. Values are given for each ensemble member, of which the initial ensemble had two hundred, while the ensembles produced from the results of the MCMC routines were limited to sizes of one hundred and of seventy four. These are ordered from best to worst ranked, smaller values being the more desirable result. The large improvement in the fit of the selected model runs to

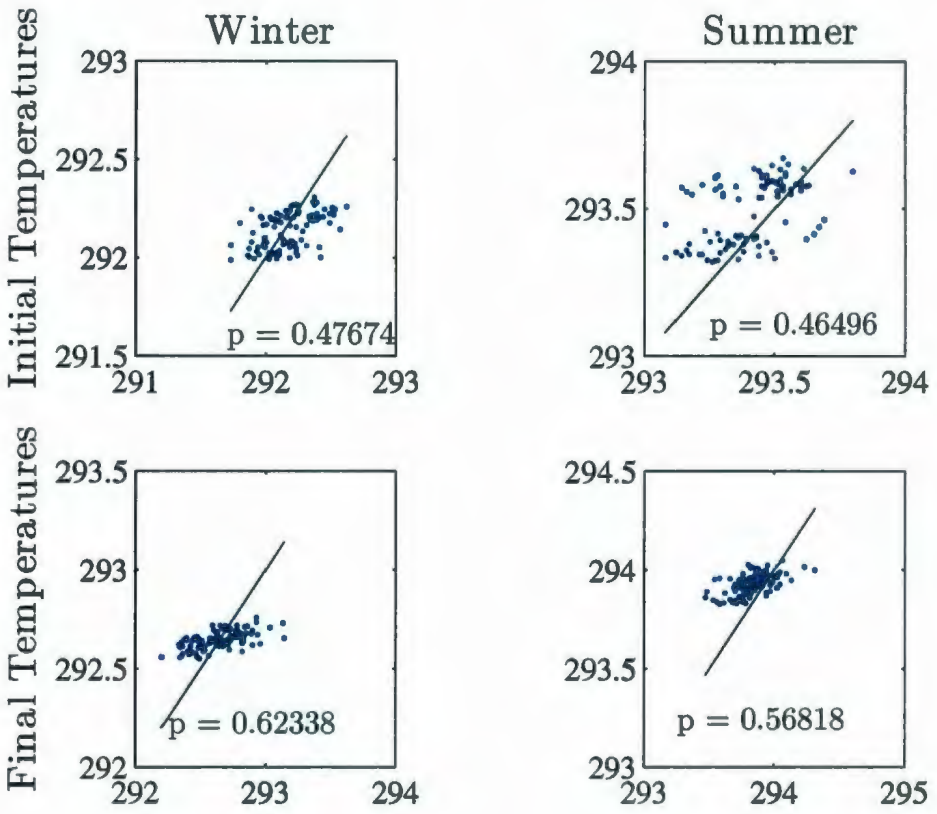

Figure 4.9: Fits between test data (x-axis) and network predictions (y-axis) for the neural network simulating model output for the South Atlantic region for the winter and summer seasons.

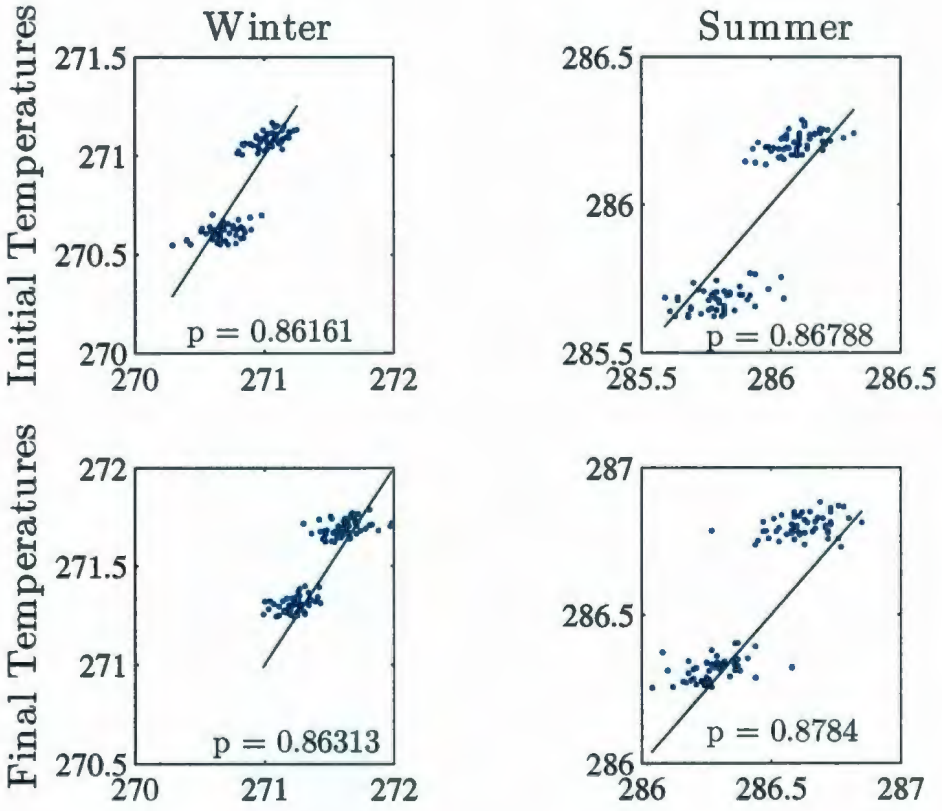

Figure 4.10: Fits between test data (x-axis) and network predictions (y-axis) for the neural network simulating model output for the North American region for the winter and summer seasons.

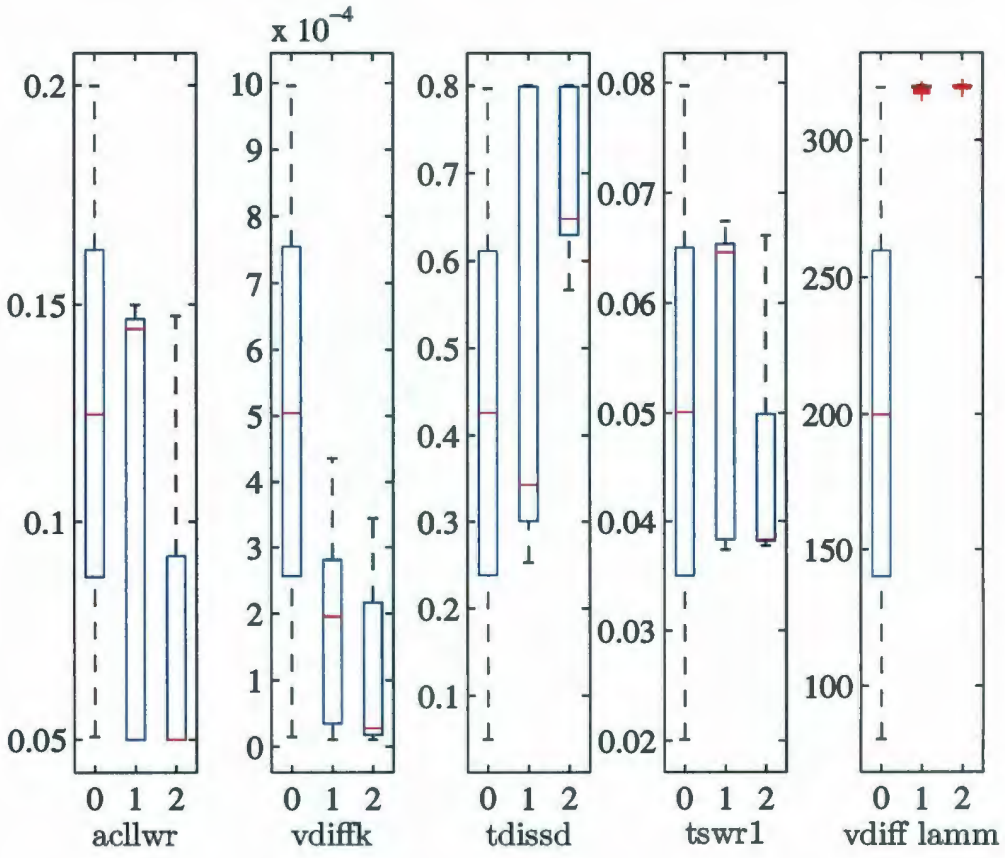

Figure 4.11: Distributions of individual parameter values from the prior and two iterations of the MCMC analysis routine.

the calibration data, between the initial ensemble and the first iteration is striking. The second iteration offered no apparent further improvement in the likelihood.

Model match to the calibration data is not necessarily indicative of match to the observation field. Figure 4.13 displays the difference between the observed initial and final winter temperatures and those from a weighted mean of the top twenty ensembles from the final iteration. As the final posteriors did not show evidence of containing multiple modes, a weighted mean was judged an adequate representation of the result. Polar regions are excluded from the assessment as they are poorly captured by the model, and were not accounted for in the calibration procedure. The winter season only is displayed here. The analysis showed no significant differences in results across seasons, and so the winter season was chosen to be representative of the general results. The overall mismatch is well beyond the ensemble standard deviation which is also provided in the figure. Figure 4.14 again displays the temperature difference between observation and ensemble, in this case comparing the ensemble result to that from the model run at its default parameter settings. The nature of the model fits to observation are distinct, but it is difficult to discern, given the magnitude of variability, if the ensemble mean represents an improved fit to the observed climate state. Root Mean Square Error (RMSE) statistics are shown in Table 4.4. By this metric the calibration appears to have produced a subtle improvement in the models ability to replicate the present climate state.

Figure 4.15 shows the difference between observed seasonal temperature change, i.e. initial climate state minus final climate state, and that from the calibrated ensemble. The climate change signal over the time span observed is small, at magnitudes of under one degree, and the regions where the degree of mismatch between forecast

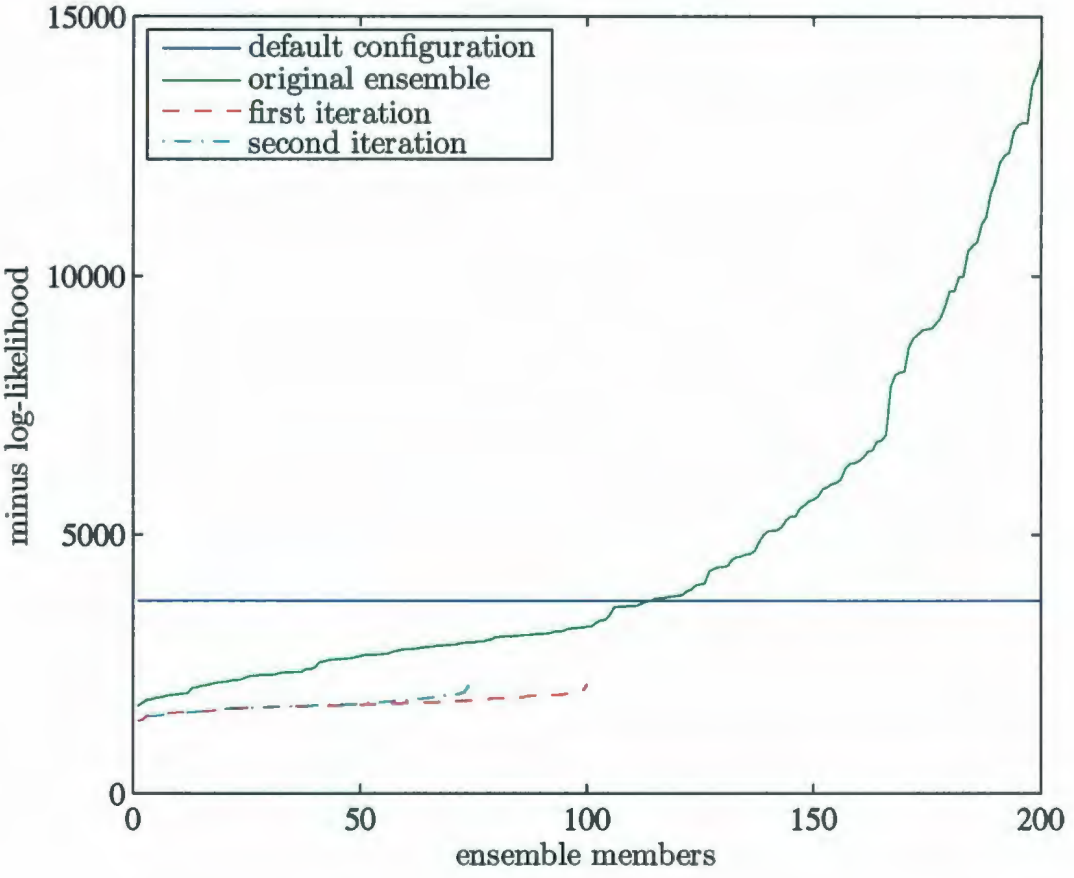

Figure 4.12: Plot of log-likelihood values calculated from model output and observations for members of the original model ensemble, that of the ensembles produced by two iterations of the NN/MCMC routine, and the original default parameter settings.

Table 4.4: Comparison of global (without poles) RMSE between the observed state and the weighted ensemble and default model winter fields

	initial state	final state
observation - weighted ensemble mean	3.599	3.615
observation - default model	3.787	3.819

and observation is not of a larger magnitude than this signal are limited. The best observation forecast fits occur over the oceans, which at the time scale used would demonstrate the smallest amount of variability, and so are of the least interest. That the climate change signal is obscured by the imprecision with which the model describes observed reality serves as a caution against using this calibrated model as a forecast tool for the time scale employed.

4.4 Discussion

For the Planet simulator, it has been shown that the NN/MCMC approach produced a notable improvement over the default settings in fit to the data used for calibration, and a lesser degree of improvement to the overall observed state. However, further calculations showed that the model ensemble was unable to replicate the observations within the given uncertainties of the ensemble and observations (observational uncertainty has not been displayed). To what extent this is due to the calibration methodology versus the limited dynamical response of the model to the chosen ensemble parameters is at this stage unclear. A model that is unable to cover the observed dynamical phase space with even the unknown optimal model parameter set will always have its ensemble standard deviations at least partially disjoint from the observed data. The construction of a complete error model for the calibrated ensemble that fully takes into account limitations in model fit will be explored in future work.

The calibrated model ensemble was unable to replicate the climate anomaly that occurs in the time scale considered. This is not surprising given the low signal strength

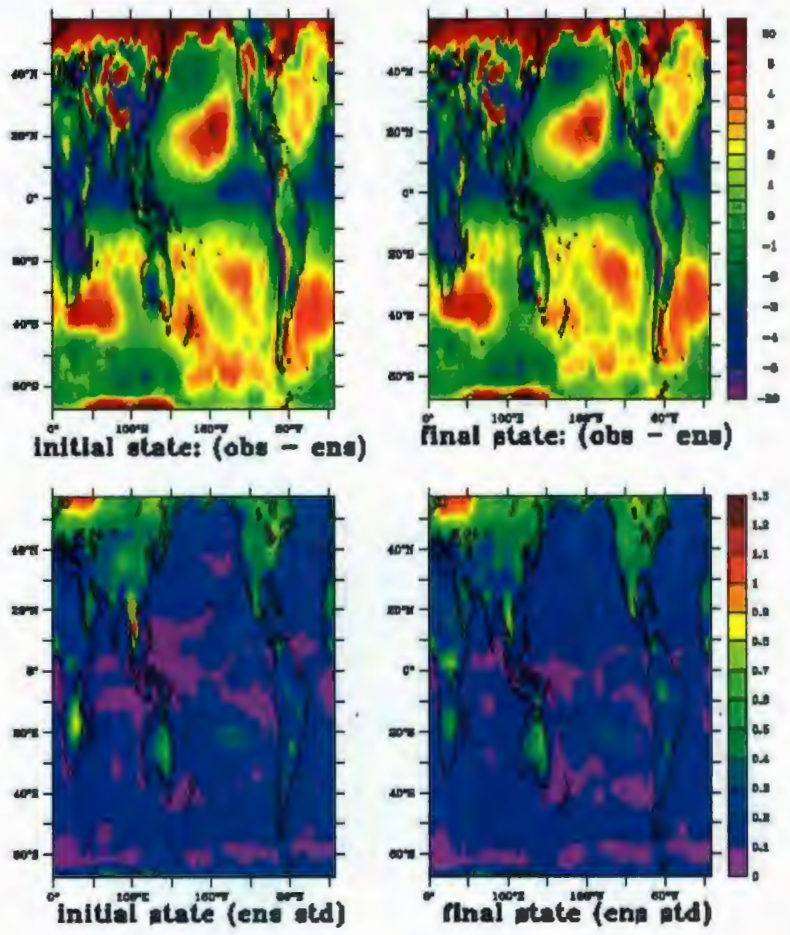

Figure 4.13: Difference between observed and weighted ensemble mean winter surface temperatures for 1959-1968 (top left) and 1999-2008 (top right) with the standard deviation for the ensemble results (below).

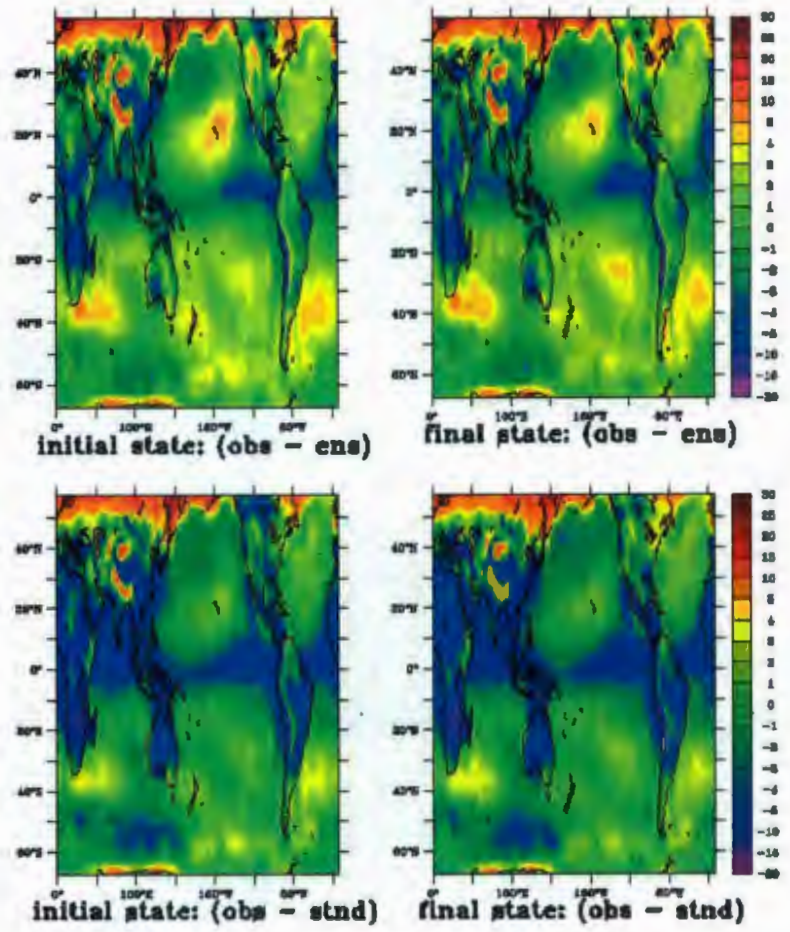

Figure 4.14: Difference between observed and weighted ensemble mean winter surface temperatures for 1959-1968 (top left) and 1999-2008 (top right) and difference between observed and standard model mean winter surface temperatures for 1959-1968 (bottom left) and 1999-2008 (bottom right).

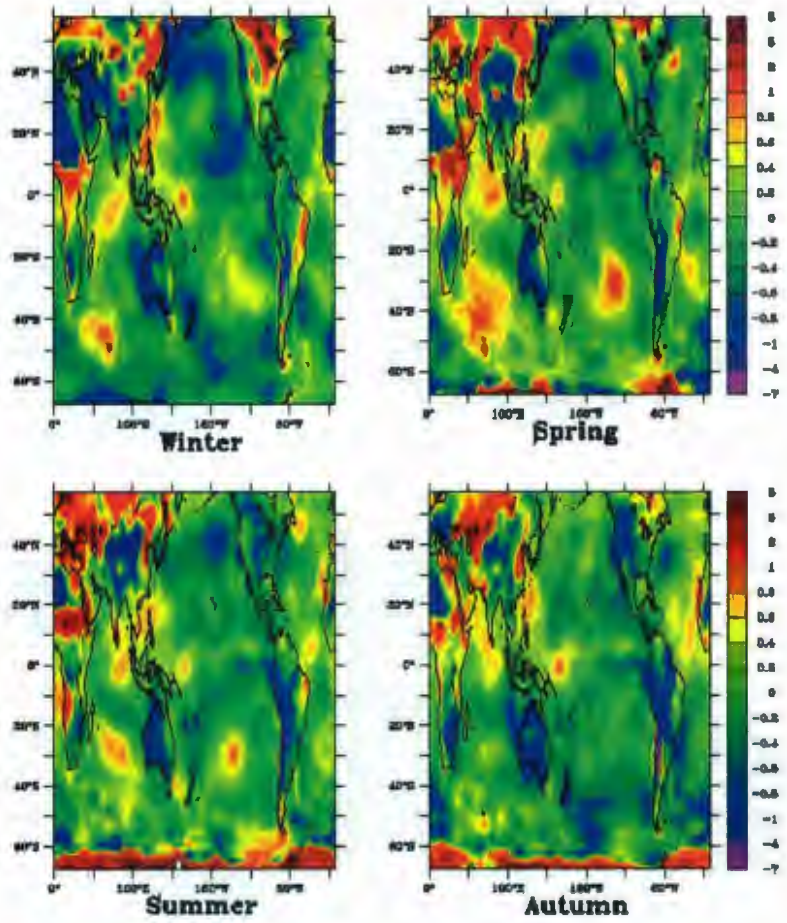

Figure 4.15: Difference between observed and ensemble anomaly between seasonal surface temperatures for 1959-1968 and 1999-2008.

of the anomaly, the simplicity and low resolution of the model, and limited set of calibration parameters. This lack of replication suggests that such models may offer no predictive value for such relatively short term transient climate forecasting.

Another significant factor is the degree to which the calibration data is representative of the observed state. In order to reduce the computational complexity of the calibration for this exploratory analysis, the model was calibrated to a much reduced representation of the observed system. Given the significant impact that the calibration achieved with respect to the constraint data shown by Figure 4.12, it conceivable that the selection of constraint data could have a large influence on the outcome of the routine. Another possible limiting factor in the NN/MCMC procedure is the ability of the networks to emulate the model being calibrated. Emulation performance can be improved by providing more training data, or by increasing the complexity of the networks. Increasing the complexity of the calibration metric above also puts increased computational demands on the network training, and may require additional training sets as well in order to realize the potential benefits of the more involved network.

It is also interesting to note that similar trends are present for both the EnKF and NN/MCMC concerning the evolution of the posterior distributions of parameter sets over the iterations, as depicted in Figures 4.6 and 4.11. This is not surprising considering the similarity in the likelihood criteria of both methods. The difference is in the usefulness of the results. The NN/MCMC produced a selection of parameter sets that could feasibly be used to construct a model ensemble. The EnKF however, produced a selection of parameter sets that were too distant from the prior to be realistically utilized. This can be considered a result of the size of the calibration set

and of the ensemble size. The updated parameter sets generated by the EnKF are produced by a nudging scheme informed by covariance statistics from the model ensemble. These statistics can identify trends that result in improved fits. For complex problems large amounts of observational data or ensemble members are required to refine these statistics to the point where the nudging will occur within an appropriate scale. The other issue is, as in Chapter 3, that the EnKF is not constrained by the prior in the same way as in the MCMC sampling. In the later the prior is a part of the calculation that determines the selection of every element sampled. In the statistics that propagate the EnKF algorithm the prior just provides the distribution of the initial ensemble. In the EnKF algorithm all distributions are assumed to be Gaussian so the uniform prior that informed the creation of the initial ensemble is perceived by the algorithm as being a normal distribution with a wide variance. To correct this issue prior ensembles of the EnKF must be constructed with this in mind. This creates a difficult issue to resolve as there are limited applications where the initial understanding of a system is precise enough to justify the use of a Gaussian prior.

Chapter 5

Conclusion

5.1 Summary and Future Work

In the experiments discussed in this work the NN/MCMC routines have outperformed the EnKF routines in several aspects. However, the experiments have highlighted issues in implementation of both methods that must be addressed in the future if they are to be used effectively for model calibration and uncertainty estimation.

The poor performance of the EnKF is not entirely surprising. This method was originally developed as a data assimilation tool for updating the initial conditions for forecasts of time evolving chaotic systems. It is the ability of the EnKF to calculate error statistics, and to use them to suggest new states rather than simply weighting current ones, that makes it of interest to the calibration problem. Limits on ensemble size, iterations performed, and calibration data can affect the ability of the algorithm to perform these functions optimally, as was observed in Chapter 4. That the limitations imposed in Chapter 4 are realistic for many earth system model calibration

scenarios suggest that the EnKF may not be a functional tool for these applications. The other concern regarding the EnKF is the Gaussian formulation. The algorithm requires the assumption of the existence of a best (yet noisy) fit to reality. As seen in Chapter 3, this causes the method to be prone to becoming trapped in local minimums, and to give inaccurate error estimates. For complex non-linear models with the potential for multiple modes of fit this method is inappropriate.

The NN/MCMC routines did meet some important benchmarks. In Chapter 3 the calibration result was responsive to the bimodal solution to the EBM calibration problem. In Chapter 4, The NN/MCMC method was able to identify (with respect to calibration data) higher likelihood parameter sets beyond those in the initial ensemble. Also, a weighted ensemble produced through the NN/MCMC calibration improved GCM fit to present day conditions. Like the EnKF the NN/MCMC routine is an ensemble method and so allows for the calculation of forecast uncertainties.

The degree of misfit between the ensemble mean and the observed state (taking into account the observational uncertainty) was calculated to be well beyond one standard deviation of the calculated forecast uncertainty in the case of Chapter 4. One unaccounted for source of uncertainty is the degree to which the calibration data misrepresents the full observed state. Uncertainty statistics that account for the error of interpolation and not just the error of observation would result in a wider ensemble. Also, in the future, more sophisticated methods of analyzing the ensemble data must be used to give a more complete view of ensemble forecast uncertainty. It will also be desirable to improve the ability to qualify model uncertainty statistics, i.e. to discern if the calculated forecast uncertainty is representative of system behavior and phase space, or a gauge of model precision. Also, in Chapter 4 the ensemble mean did not

capture the more subtle signals present in the data. This may have resulted from inherent model limitations, but may also be linked to the accuracy with which the neural networks were able to determine model response. The NN/MCMC approach will always be limited by the degree to which the neural networks are able to emulate the model behavior. The optimization of the emulation accuracy and the detection of situations where the emulators produce inaccurate and thus misleading results are key implementation issues. An important component of future work on this method will be to develop general procedures for this implementation. It may also be productive to attempt identifying possible alternative emulation schemes, such as Rougier (2008).

This exploratory study was confined to a low spatial resolution calibration of the Planet Simulator GCM. Calibration of GCMs/Earth systems models for current research and climate "forecasting" contexts using the NN/MCMC approach will require higher spatial resolution, improved emulation, and a likelihood model capable of accounting for correlations within the constraint data. Higher spatial resolution could be obtained by expanding the network to accommodate input vectors consisting of spatial coordinates as well as model parameters. Network output would represent the model state at this location. This approach is in keeping with the work of Tarasov and Peltier (2005). A network capable of performing such an emulation at a resolution close to that of the model (i.e. that was trained to make predictions for most of the grid points described in the model output space) would likely have to be much more complex than those described in the experiments above to attain a similar level of accuracy. Also, the data sets employed in training would be much larger if such a network were trained with data from a similar number of runs as used in these experiments. These factors equate to a large increase in network training time. In

initial experiments with this approach training time went from being in the order of hours (as in the experiments of Chapters 3 and 4) to days. However this is still a small increase in computational demands compared to that occurred by increasing the quantity of GCM runs. With a single network responsible for generating a wider variety of outputs, assessing the skill of the emulator is a more complex task. Discerning the range of emulator misfit to data, and identifying regions where the emulator skill is relatively high or low will be important tasks in deciding how to use the networks for the purpose of calibration. Increasing the spatial resolution of the calibration routine would require the use of a more sophisticated likelihood function, i.e. one that is capable of taking into account the high spatial correlations that would be present in the utilized fields. In general, assessing the effect of alternate likelihood functions and ways to select relevant functions that take full advantage of the information available in the observational data, will be important avenues of future work with the NN/MCMC method.

In the exploratory experiments presented here, the definition of priors was done in a very general fashion. Further investigation is needed to discern the sensitivity of calibration methods to the prescribed priors. For example, if verified, the possible need for a Gaussian prior when using the EnKF would be a significant limiting factor towards its application. It is desirable to be sure that what prior knowledge is available can be used accurately by the calibration routine employed. Also, the more vague it is possible for the prior to be, the more widely applicable the method.

The selection of parameters to be calibrated was conducted in an ad hoc fashion in the work presented here. Calibration will not accurately reflect model uncertainties and will have limited effect on model performance if parameters relevant to the output

used for calibration are not used in the procedure. As the parametrization of earth systems models can be quite extensive, and sensitivity testing can be computationally demanding, more sophisticated ways of determining appropriate parameter sets for calibration will need to be addressed in the future (such as Automatic Relevance Detection, Neal 1996).

To conclude, the NN/MCMC approach shows promise as a calibration routine for computationally demanding earth systems models. It will be important to develop a further understanding of how the individual components of the method influence its performance, so as both to refine the method and to better assess it applicability.

Bibliography

Anderson, J. L. (2001). An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review 129*, 2884–2903.

Annan, J. and J. Hargreaves (2004). Efficient parameter estimation for a highly chaotic system. *Tellus* (56A), 520–526.

Annan, J., J. Hargreaves, N. Edwards, and R. Marsh (2005). Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter. *Ocean Modelling* (8), 135–154.

Annan, J. D. and J. C. Hargreaves (2007). Efficient estimation and ensemble generation in climate modelling.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar (2000). Adaptive sampling with the ensemble transform Kalman filter. part i: Theoretical aspects. *Monthly Weather Review 129*, 420–436.

Evensen, G. (2003). The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics 53*.

Evensen, G. (2004). Sampling strategies and square root analysis. *Ocean Dynamics 54*.

Evensen, G. (2005). The parameter estimation problem revisited. Slides for SIAM Conference on Mathematical and Computational Issues in the Geosciences.

Evensen, G. (2007). *Data Assimilation. The Ensemble Kalman Filter.* Springer-Verlag Berlin Heidelberg.

Evensen, G. (2009). The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control Systems Magazine 83.*

Gershenfeld, N. A. (1999). *The Nature of Mathematical Modeling.* Cambridge University Press.

Hogg, R. and E. Tanis (2001). *Probability and Statistical Inference* (6th ed.). Prentice-Hall, New Jersey.

Jackson, C., M. K. Sen, and P. L. Stoffa (2004). An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions. *Journal of Climate 17.*

Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, C. M., W. Ebisuzaki, W. Higgins, K. Janowiak, J. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, and D. Jenne, R. Joseph (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteoroligcal Society 77.*

Khu, S. T. and M. Henrik (2005). Multiobjective calibration with Pareto preference ordering: An application to rainfall-runoff model calibration. *Water Resources*

Research 41(WO3004).

Kim, S., G. Eyink, J. Restrepo, F. Alexander, and G. Johnson (2003). Ensemble filtering for nonlinear dynamics. *Monthly Weather Review 131.*

Lunkeit, F., M. Böttinger, K. Fredrich, H. Jansen, E. Kirk, A. Kleidon, and U. Luksch (2007). *Planet Simulator Reference Manual* (15 ed.). University of Hamburg.

Lunkeit, F. Blessing, S., K. Fraerich, H. Jansen, E. Kirk, U. Luksch, and F. Sielmann (2007). *Planet Simulator User's Guide* (15 ed.). University of Hamburg.

Moradkhani, H., S. Sorooshian, H. Gupta, and P. Houser (2005). Dual state-parameter estimation of hydrological models using ensemble Kalman filter. *Advances in Water Resources 28.*

Müller, P. and H. von Storch (2004). *Computer Modelling in Atmospheric and Oceanic Sciences, Building Knowledge.* Springer-Verlag Berlin Heidelberg.

Neal, R. (1993). Probabilistic inference using Markov Chain Monte Carlo methods. Technical Report CRG-TR-93-1.

Neal, R. (1996). *Bayesian Learning for Neural Networks.* Springer-Verlag, New York.

Neal, R. (2000). Slice sampling. Technical Report No. 2005.

Neal, R. (2004). Slides on calibration for ice-sheet modeling.

Nettuno, L. (1995). Field measurements and model calibration in avalanche dynamics. *Surveys in Geophysics 16*(5-6), 635–648.

Oakley, J. and A. O'Hangan (2002). Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika 89*(4), 769–784.

Reynolds, R., N. Rayner, D. Smith, and W. Wang (2002). An improved in situ and satellite sst analysis for climate. *Jornal of Climate 15*, 1609–1625.

Rougier, J. (2008). Efficient emulators for multivariate deterministic functions. *Jornal of Computational and Graphical Statistics 17*(4), 827–843.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tignor, and H. Miller (2007). *Contribution of Working Group 1 to the Fourth Assesment Reoport of the Intergovermental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Tans, P. (2009). Mauna Loa CO2 annual mean data. www.esrl.noaa.gov/gmd/ccgg/trends/. NOAA/ESRL.

Tarasov, L. and W. Peltier (2005). Arctic freshwater forcing of the Younger Dryas cold reversal. *Nature 435*, 662–665.

van Delden, A. (2008). *Atmospheric Dynamics*. Institute for Marine and Atmospheric Research Utrecht University.

van der Merwe, R., A. Doucet, N. de Freitas, and E. Wan (2000). The unscented particle filter. Technical Report CUED/F-INFENG/TR 380.

Wikle, C. and L. Berliner (2007). A Bayesian tutorial for data assimilation. *Physica D 230*(1), 1–6.

Appendix

Derivation of the The Kalman Filter

The following is an overview of the derivation of the basic Kalman Filter, as presented in (Evensen 2007).

Table A.1: Terms for basic Kalman Filter derivation.

ψ	system state
ψ^t	true system state
ψ^f	forecast state
ψ^a	analysis state
ρ^f	forecast error
ϵ	measurement error
$C^a_{\psi\psi}$	error covariance of state estimate
$C^f_{\psi\psi}$	error covariance of process noise
d	observations
H	observation transition matrix
$d - H\psi^f$	innovation
$C_{\epsilon\epsilon}$	error covariance of observation noise
$HC^f_{\psi\psi}H^T + C_{\epsilon\epsilon}$	error covariance of innovation
K	optimal Kalman gain matrix

Assume the existence of a true state: ψ^t. Estimate a forecast of ψ as the true

state plus the error of forecast and a measurement of ψ as the true state plus the measurement error:

$$\psi^f = \psi^t + \rho^f$$

$$d = \psi^t + \epsilon$$

Assuming: $\overline{\rho^f} = 0$, $\overline{\epsilon} = 0$, $\overline{\epsilon\rho^f} = 0$, $\overline{(\rho^f)^2} = C^f_{\psi\psi}$, $\overline{(\epsilon)^2} = C_{\epsilon\epsilon}$

Want an analysis ψ^a s.t.

$$\psi^a = \psi^t + \rho^a = \alpha_1\psi^f + \alpha_2 d$$

Let: $\overline{\rho^a} = 0$, $\overline{(\rho^a)^2} = C^a_{\psi\psi} = 0$, i.e. ρ^a is unbiased.

This gives that:

$$\psi^t + \rho^a = \alpha_1(\psi^t + \rho^f) + \alpha_2(\psi^t + \epsilon)$$

Looking at the expectation of the above (where all error terms equal their mean) gives:

$$\psi^t = (\alpha_1 + \alpha_2)\psi^t \rightarrow \alpha_1 = 1 - \alpha_2$$

Which gives:

$$\psi^a = \psi^f + \alpha_2(d - \psi^f)$$

i.e. the analysis equals the forecast plus the weighted difference between the measurement and the forecast.

And it follows that:

$$\rho^a = \rho^f + \alpha_2(\epsilon - \rho^f)$$

So it follows from above assumptions that:

$$\overline{(\rho^a)^2} = C^a_{\psi\psi} = \overline{\rho^f + \alpha_2(\epsilon - \rho^f))^2}$$

Expanding to:

$$\overline{(\rho^a)^2} = C^f_{\psi\psi} - 2\alpha_2 C^f_{\psi\psi} + 2\alpha_2 (C_{\epsilon\epsilon} + C^f_{\psi\psi})$$

Solving for the minimum variance gives:

$$\alpha_2 = \frac{C^f_{\psi\psi}}{C_{\epsilon\epsilon} + C^f_{\psi\psi}}$$

So it follows that:

$$\psi^a = \psi^f + \frac{C^f_{\psi\psi}}{C_{\epsilon\epsilon} + C^f_{\psi\psi}}(d - \psi^f)$$

And that:

$$C^a_{\psi\psi} = C^f_{\psi\psi}\left(1 - \frac{C^f_{\psi\psi}}{C_{\epsilon\epsilon} + C^f_{\psi\psi}}\right)$$

Setting the equations in a discrete form with the same statistical hypotheses and including a measurement matrix H such that: $d = H\psi^t + \epsilon$, produces the same minimization problem with the resulting equations commonly written:

$$\psi^a = \psi^f + K(d - H\psi^f)$$

$$C^a_{\psi\psi} = (I - KH)C^f_{\psi\psi}$$

where:

$$K = C^f_{\psi\psi} H^T (H C^f_{\psi\psi} H^T + C_{\epsilon\epsilon})^{-1}$$

Details of the Ensemble Kalman Filter

Use of the Kalman Filter routine with ensembles

Following is an overview of the extension of the basic Kalman Filter algorithm to an ensemble method as presented in (Evensen 2007).

Here consider the ensemble averages and say:

$$(C^e_{\psi\psi})^f = \overline{(\psi^f - \psi^t)(\psi^f - \psi^t)^T}$$

$$(C^e_{\psi\psi})^a = \overline{(\psi^a - \psi^t)(\psi^a - \psi^t)^T}$$

But as ψ^t is unknown assume that $\psi^t = \overline{\psi^f}$ giving:

$$(C^e_{\psi\psi})^f = \overline{(\psi^f - \overline{\psi^f})(\psi^f - \overline{\psi^f})^T}$$

$$(C^e_{\psi\psi})^a = \overline{(\psi^a - \overline{\psi^a})(\psi^a - \overline{\psi^a})^T}$$

Define an ensemble of observations $d_j = d + \epsilon_j$, $\bar{\epsilon} = 0$, $C^e_{\epsilon\epsilon} = \overline{\epsilon\epsilon^T}$. Where ϵ_j is a vector of observation noise for each 1,...,j ensemble member.

Then the equations from above read as:

$$\psi^a_j = \psi^f_j + (C^e_{\psi\psi})^f H^T (H(C^e_{\psi\psi})^f H^T + C^e_{\epsilon\epsilon})^{-1}(d_j - H\psi^f_j)$$

With the ensemble mean given by:

$$\overline{\psi^a} = \overline{\psi^f} + (C^e_{\psi\psi})^f H^T (H(C^e_{\psi\psi})^f H^T + C^e_{\epsilon\epsilon})^{-1}(\bar{d} - H\overline{\psi^f})$$

So write K^e as:

$$K^e = (C^e_{\psi\psi})^f H^T (H(C^e_{\psi\psi})^f H^T + C^e_{\epsilon\epsilon})^{-1}$$

And $(C^e_{\psi\psi})^a$ can be reduced to:

$$(C^e_{\psi\psi})^a = (I - K^e H)(C^e_{\psi\psi})^f$$

Practical Formulation/Implementation of EnKF

This section outlines the formulation of the EnKF used in the experiments presented in this work, as outlined by (Evensen 2003). The final portion of the formulation is designed to allow the involved matrix calculations to be performed in the most computationally efficient way possible.

Standard Analysis Equation

$$A^a = A + P_e H^T (H P_e H^t + R_e)^{-1}(D - HA) \tag{A.1}$$

where:

$A \in \mathbb{R}^{n \times N}$ is the matrix of ensemble members $\psi_i \in \mathbb{R}^n$ such that

$A = (\psi_1, \psi_2, \ldots, \psi_N)$

$P_e \in \mathbb{R}^{n \times n}$ is the ensemble covariance matrix:

$P_e = \frac{A'(A')^T}{N-1}$, with $A' = A - \overline{A}$

$D \in \mathbb{R}^{m \times N}$ is the matrix of perturbed observations:

$D = (d_1 + \epsilon_1, d_2 + \epsilon_2, \ldots d_N + \epsilon_N)$,

$R_e \in \mathbb{R}^{m \times m}$ is the observation perturbation covariance matrix:

$R_e = \frac{\Upsilon'(\Upsilon)^T}{N-1}$, where:

$\Upsilon \in \mathbb{R}^{m \times N}$ is the matrix of perturbations:

$\Upsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_N)$

H translates A into the observable variables it forecasts

can rewrite Standard Analysis Equation

$$A^a = A + A'A'^T H^T (HA'A'^T H^t + \Upsilon\Upsilon^T)^{-1} D' \tag{A.2}$$

where: $D' = D - HA$

to solve the above equation

requires the computation of:

$$HA'A'^T H^t + \Upsilon\Upsilon^T = Z\Lambda Z^T \tag{A.3}$$

or

choose measurement perturbations such that:

$HA'\Upsilon^T \equiv 0$, i.e. forecast and measurement errors are uncorrelated

then can write: $HA'A'^T H^t + \Upsilon\Upsilon^T = (HA'\Upsilon)(HA'\Upsilon)^T$

compute SVD: $HA'\Upsilon = U\Sigma V^T$

so have:

$$HA'A'^T H^t + \Upsilon\Upsilon^T = U\Sigma\Sigma^T V^T \tag{A.4}$$

where:

$\Sigma\Sigma^T$ = the N nonzero eigenvalues of Λ

singular vectors in U = the first N eigenvectors of Z

so have

$$A^a = A + A'(HA')^T U \Lambda^{-1} U^T D' \tag{A.5}$$

let:

$$\begin{aligned}
X_1 &= \Lambda^{-1} U^T \\
X_2 &= X_1 D' \\
X_3 &= U X_2 \\
X_4 &= (HA')^T X_3 \\
&\Rightarrow \\
A^a &= A + A' X_4 \\
&= (A - \overline{A}) X_4 \\
&= A(I - 1_N) X_4 \\
&\textit{let} \quad 1_N X_4 \equiv 0 \\
&\Rightarrow \\
A^a &= A(I - X_4)
\end{aligned}$$

let:

$$\begin{aligned}
X_5 &= I + X_4 \\
&\Rightarrow \\
A^a &= A X_5
\end{aligned}$$