Probabilistic uncertainty quantification and simulation for climate modelling

by

©Tristan Paul Hauser

A Thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Physics and Physical Oceanography

Memorial University of Newfoundland

May 2014

St. John's Newfoundland

Abstract

This thesis addresses probabilistic approaches to uncertainty quantification, within the context of climate science. For the results of climate studies to be appropriately understood and applied, it is necessary to quantify their relation to the observable world. Probability theory provides a formal approach that can be applied commonly to the encountered uncertainties. Three studies are presented within. The first addresses the Bayesian calibration of climate simulators. This method quantifies simulation uncertainties by taking into account inherent model and observation uncertainties. Here an alternative method for the fast statistical emulators of model parameter relationships is tested, as well as a rigorous approach to quantifying model limitations. The second examines probabilistic methods for identifying regional climatological features and quantifying the related uncertainties. Such features serve as a basis of comparison for climate simulations, as well as defining, to some extent, how we view evolution of the modern climate. Here typical patterns are recreated using an approach that quantifies uncertainty in the data analysis. As well, temporal shifts in the distribution of these features and their relation to ocean variability is explored. The third study experiments with approaches to regional stochastic weather generation. There is an inherent residual between climate simulations and large scale features, and regional variability seen on daily timescales. Weather generators provide an error model to quantify this uncertainty, and define features and variability underrepresented in global simulations. A method is developed which allows for regional, rather than site specific, simulation for the North Atlantic, a region of very active and varied atmospheric activity. In total, the work presented within covers the range of uncertainty types that must be considered by climate studies. The individual articles addresses contemporary questions concerning appropriate methods and implementation for their probabilistic quantification.

Acknowledgements

Financial support was provided from the Canadian Foundation for Innovation, the National Science and Engineering Research Council, and the Atlantic Computing Excellence Network. Without this support, it would not have been possible to pursue this research on a full-time basis.

The projects presented in this thesis were instigated and tirelessly assisted by my supervisors Lev Tarasov and Entcho Demirov. Further guidance, support, and instruction were generously provided by my other committee members, Brad deYoung and Joel Finnis.

This work has benefited greatly from the patient instruction provided by Robert Briggs and Andrew Keats. Also, by assistance from and discussions with Taimaz Bahadory, Wolfgang Banzhaf, John Foley, Arnault Lebris, Kevin Le Morzadec, Sarah Lundrigan, Radford Neal, Jonas Roberts, Alexander Slavin, Yi Sui, Christopher Stevenson, Graig Sutherland, David Thompson, and Roger White.

Table of Contents

List of Tables

List of Figures

Nomenclature

Abbreviations

AIC	Akaike Information Criterion
AMOC	Atlantic Meridional Overturning Circulation
ANN	Artificial Neural Network
AR	Atlantic Ridge
ARD	Automatic Relevance Detection
BANN	Bayesian Artificial Neural Network
BIC	Bayesian Information Criterion
EMIC	Earth systems Model of Intermediate Complexity
GCM	General Circulation Model
GMM	Gaussian Mixture Model
GPE	Gaussian Process Emulators
GS	Greenland-Scandinavian Dipole
GSA	Great Salinity Anomaly
H5	500mb geopotential height
HCM	Hybrid Coupled Model
LIM	Linear Inverse Model
LSW	Labrador Sea Water
LWR	Long Wave Radiation
MCMC	Markov Chain Monte Carlo
NAM	Northern hemisphere Annular Mode
NAO	North Atlantic Oscillation
NCAR	Nation Center for Atmospheric Research
NCEP	National Center for Environmental Prediction
NP	North Pacific
PC	Principal Component
PCA	Principal Component Analysis
PDF	Probability Density Function
PNA	Pacific - North American
RCM	Regional Climate Model
RMSE	Root Mean Square Error
SLP	Sea Level Pressure
SOM	Self Organising Map
SST	Sea Surface Temperature
SWR	Short Wave Radiation
WG	Weather Generator

Chapter 1

Introduction

This chapter provides an overview of and supporting material for the three articles that document the work performed for this thesis. As the body of the thesis is composed of stand-alone articles, there is overlap between these individual articles and with the material presented here. Some emphasis is placed here on providing details for subjects not directly addressed in later sections and on synopses of alternate approaches not pursued.

1.1 Types of Uncertainties

The central theme of the work presented here is the quantification of uncertainty for climate modelling applications. The need for improved quantification of forecast uncertainties has been expressed by many sources including the Intergovernmental Panel on Climate Change; e.g., Stainforth et al. (2005) ; Murphy et al. (2007a) ; Solomon et al. (2007a) ; Snyder et al. (2011). Without some measure of how climate simulation results relate to the observable world, these studies represent little more than best guesses or thought experiments. It is desired that, when possible, these uncertainties be expressed probabilistically (Webster et al., 2006) as, due to the inherent unpredictability of climate systems (Palmer, 2006), most end user decisions regarding climate issues are matters of risk management (Richardson, 2006; Beven, 2009). As well, end users are not only concerned with long-term global phenomena, but require information on sub-annual scales for prescribed regions, as this relates to a variety of civic, industrial and agricultural concerns; e.g, Beven (2009), Maraun et al. (2010), Brand (2011), and Mukherjee and Dutta (2011). As such, uncertainty quantification requires methods which describe the natural variability of the investigated system(s) and the potential errors of the simulation methods. The work presented here addresses both needs by creating probabilistic descriptions of various observed and simulated elements of the climate system.

Computational simulations, such as those used in climate studies, have many inherent sources of uncertainty, listed in Table 1.1. Most general is that they are typically applied to complex nonlinear systems. As such, even though these simulations are defined by formal relationships there is no *a priori* way to know the output that will result from given inputs. Nor is it possible to fully predict how changes in input will effect the resulting output. This is referred to as “code uncertainty” (Kennedy and O’Hagen, 2001), although often this situation is a result of the complexity of the system being investigated, hence the need for complex models, rather than only a property of the programmes used. This is especially relevant to climate simulations given the highly nonlinear and open nature of the climate system as well as the intricacies of many climate models.

In climate simulations, responses to subprocesses and external forcing are often represented by empirical parametrizations, defined by constants referred to as model parameters. In climate modelling, uncertainty regarding appropriate values for parameters which cannot be derived from first principles is known as parametric uncertainty. These approximate descriptions of indirect processes are unavoidable as climate is a continuous open system while numerical models are limited to discrete descriptions on predetermined scales; an issue known as the "closure problem" (Muller and Storch, 2004). The term parametric uncertainty is in other contexts used to refer to uncertainty regarding model inputs in general (Kennedy and O'Hagen, 2001).

For climate simulations, additional inputs are required to prescribe external forcing and initial conditions. External forcing includes solar cycles and greenhouse gas concentrations for global models as well as boundary conditions and other indicators for regional simulations. Solar cycles are highly predictable and paleo records of atmospheric concentrations are available. The future composition of the atmosphere, however, is dependent on many chemical, biological, and anthropogenic factors, few of which are well understood and all of which will be influenced by climate evolution (Snyder et al., 2011). Greenhouse gas concentrations are external in the sense that it is uncertain how to describe these feedback effects, although carbon cycle models are under development. Typically concentrations are prescribed as abstract experiments or as plausible scenarios (Nakicenovic et al., 2000a). For the near future (next fifty years) the difference between conservative and extreme proposed scenarios is small and other uncertainties dominate (Wilby and Harris, 2006; Deque et al., 2007). Over many (~ 10) decades, however, the range of plausible outcomes diverges and becomes a driver of the mean climate state (New and Hulme, 2000). An additional input uncertainty for climate simulations are initial conditions. Chaotic systems, such as those which describe atmospheric evolution, are highly sensitive to this uncertainty which dominates over meteorological time scales. Since initial conditions cannot be perfectly known, even a perfect simulation will diverge from the observed evolution in a matter of (simulated) days (Kalnay, 2002). However, unless the system is near a tipping point; i.e., an unstable state where a small deviation will cause the system to tend towards a new equilibrium, mean climate statistics are insensitive to initial conditions, over sufficiently long simulations. As such, climate simulations should be evaluated using long-term climate statistics, rather than the observed chronologies typically used to evaluate meteorological forecasts (Muller and Storch, 2004).

Note that the parametric uncertainty, as discussed, refers only to choosing appropriate parameter values, not the parametrization schemes that incorporate them. Errors resulting from differences between the true system dynamics and simulator design are referred to as structural errors. Since these errors would exist even if optimal forcings, initial conditions, and parameter values could be known, they are considered a separate source of uncertainty. As all simulations are limited representations of observed systems these discrepancies are inherent to the practice of modelling. Describing structural errors for climate simulations is complex due to the number of simulated variables, and the imprecision with which the relationships between these variables is known (Allen et al., 2006). As well, limited amounts of data for past climate can make it difficult to quantify the temporal nature of model biases; e.g., if a simulation under-represents an observed warm period, this could imply that the simulation is cold biased, or that due to uncertainty in initial conditions it is representing a different state of a natural cycle than that of the observed chronology cf., Deser et al. (2012).

Related to structural uncertainty is what Kennedy and O'Hagen (2001) call "residual variability". This refers to the situation where a simulation makes a prediction for a given set of inputs, for which there are many possible observed states. In some cases this is an example of structural error; a more detailed description of the system might more narrowly define the relationship between inputs and possible events. For weather forecasting, however, the chaotic nature of the involved systems make current states unpredictable past the range of a few days. As such, climate models can only reliably describe long-term climate conditions; statistical properties which would be shared by many possible weather sequences. For this context, structural error is considered to be uncertainty about a simulation's depiction of climate variables, while residual variability refers to uncertainty as to how these climatic conditions will manifest themselves as weather events.

Descriptions of the uncertainties assumes a known state for the simulation to be compared against. Observational uncertainties however are unavoidable and need to be taken into account when estimating other uncertainties. Additionally, often for climate studies direct measurements are not possible and so variables must be inferred from proxies using potentially imprecise relationships (Snyder et al., 2011). Identified climate features, such as trends or cycles, also have associated uncertainties relating to the statistical methods used to identify and/or define them. One major source of observational uncertainty for climate studies is the temporally and spatially limited supply of historical data (van der Veen, 2001). This makes it difficult to verify inferences based on proxy data and to determine the significance of statistical features.

Due to the interconnected nature of the climate system, these uncertainties compound in forecasts (Snyder et al., 2011). This can be thought of as there being multiple possible external forcing scenarios and initial conditions that can be interpreted through multiple imperfect climate simulations. These predictions, in turn, actually imply a continuum of possible current states, which themselves can only be imprecisely observed and so represent many possible realities. In practice uncertainties are described using such discrete ensembles of multiple possible trajectories. However, the situation is actually even more complex as the use of “multiple” and “many” in the preceding sentences would be more accurately replaced with “infinite possibilities of differing plausibility”. As such, these ensembles must represent these plausibilities in such a way, that their total effect on certainty can be determined from the finite set of projections included in the ensemble. How to best determine and quantify uncertainty is an open question and a subject of broad philosophical discussion. As these descriptions are a quantification of the amount of information available to particular observers concerning an unknowable true state/outcome/history they are unavoidably approximate and subjective. Different approaches have been applied in climate studies. Some focus on establishing upper and lower bounds of possible behaviours or quantifying linguistic expressions of belief (Matthies, 2007). Most often in practice, probabilistic models are used as they are able to express the highest level of mathematical detail (Matthies, 2007). It is largely agreed that this is the most desirable way to express uncertainty, although it is sometimes contentious as to whether it is possible in select situations to provide all the information needed to make use of this level of formalism (Beven, 2009). General statements can be made in probabilistic form; e.g., uniform distributions which allow equal probability within a predetermined range or Cauchy distributions which describe median tendencies but still allow significant probability for any magnitude of outlier, as well as non-parametric descriptions. However, since formal distributions define plausibility across an entire range of outcomes; i.e., not just upper and lower bounds for what is possible, careless use can overstate or misrepresent the level of information available.

Table 1.1: Summary of uncertainty types

Uncertainty type	Description
Structural	Approximations: conceptual and numeric
Parametric	Selection of constants used in parametrisation of sub-scale processes
Code	Non-linearity limits ability to predict model output from inputs
Residual	Difference between simulation and full range of potential events
Observational	Limitations in measurement and classification ability

1.2 Bayesian inference

Considering probability as a measure of uncertainty is referred to as the Bayesian interpretation. This differs from the more common interpretation of probability as the measure of the expected frequency of different types of random events, but is built on the same mathematical principles (Jaynes, 2003a). As such, in the Bayesian interpretation, the probability distribution is interpreted as a quantification of the amount of information available rather than an inherent property of the investigated phenomena (Jaynes, 2003a). This allows for Bayesian inference which uses the propositions of probability theory as rules of logical inference for uncertain quantities. As such it provides a formal framework for expressing and combining uncertainties, as well as using additional information to constrain initial beliefs.

Bayesian inference is applied many times in the work presented here. Although based on the ideas of Pierre-Simon Laplace (1749–1827), (Struik, 1987), it was popularised and formalised in the 20th Century (Cox and Hinkley, 1974). Jaynes (2003a) provides a complete overview of the concepts and methods, as well as some colourful commentary on scientific practise. More applied descriptions are given by MacKay (2003) and Sivia and Skilling (2006b).

Typically analysis is carried out using Bayes' Theorem (hence the name) :

$$P(H|D,I) = \frac{P(D|H,I)P(H|I)}{P(D|I)}, \tag{1.1}$$

which is often read as H and D representing a hypothesis and observed data respectively, such that one reads: "the probability of the hypothesis given observed data (the Posterior Distribution) is proportional to the probability of the data given the hypothesis (the Likelihood Function) times the probability of the hypothesis (the Prior Distribution). The normalisation factor $P(D|I)$ is referred to as the Evidence Term.

The I term serves as a reminder that the form of the distributions used are always conditional upon the amount of information available to the individual who defines them (Jaynes, 2003a). As the I term is not an actual value in the calculations, it is often omitted. When the analysis is used to address parametric uncertainty the process is referred to as model calibration. In this case the hypotheses that comprise the H terms are sets of model parameters, so that the posterior distribution expresses confidence in the suitability of a given set of model parameters given observations and prior beliefs/information. Hence, rather than obtaining a single optimised model, one obtains a distribution of possible realisations, which can be used to estimate prediction uncertainties. Using the probability calculus also allows the inclusion of relevant uncertainties in the observations and the model in the calculation of this posterior.

The evidence term can be considered the distribution of the observations over the model space, at all possible parameter values; i.e.,

$$P(D|M) = \int P(D, \theta|M)\ d\theta = \int P(D|\theta, M)P(\theta|M)\ d\theta, \tag{1.2}$$

where M is the selected model and θ the model parameters. The choice of model M here replaces the I term as it defines the form of the distribution based on available information; i.e., the possibility of M being a useful description of the data D. As the evidence term is constant when considering the calibration problem and as its determination is typically analytically intractable, it is often ignored. However, as the term describes how well a given model fits the data “on average”; i.e., over the distribution of potential model parameter values (Sivia and Skilling, 2006b), its value can be used to compare the robustness of different models; cf., Burnham (2004), Dose and Menzel (2004), and Bhat and Kumar (2010). Alternately, model comparison through a Bayesian variant of analysis of variance methods has been developed (Kaufman and

Sain, 2010; Sain et al., 2010).

A reason for the current increase in popularity of Bayesian methods is the availability of the computational resources needed for extensive random sampling. Such sampling is used to estimate the density of and to draw samples from analytically intractable distributions. These are often produced by Bayesian inference due to difficulties evaluating the evidence term. These complex distributions can be avoided by selecting initial distributions whose combination results in closed form solutions, but this greatly limits the flexibility of the Bayesian approach. One notable sampling method is the Latin Hypercube. Here, the range of each parameter is broken up into predetermined subsets. The method randomly generates many parameter combinations, but includes samples from each subset only once. This ensures that the full ranges of individual parameters are represented using the least possible combinations (McKay et al., 1979). It has been argued to be more efficient and effective for experiment design and for providing computer inputs in machine learning applications (Urban and Fricker, 2010b), than standard gridded searches; cf., MacKay (2003). There are several variations on this method as well as metrics to compare how well dispersed the values in different hypercubes are; cf., Tang (1998), Grosso et al. (2008a), and Abdellatif et al. (2010), as well as proposed alternatives; e.g., Morris (1991). Another class of sampling techniques are **Markov Chain Monte Carlo (MCMC)** methods. These create random walks biased towards locations of higher probability, so that the distribution of samples (locations visited in the random walk) approximates the theoretical probability distribution. A basic implementation is the Metropolis-Hastings approach (Metropolis et al., 1953). There are many variations which attempt to locate informative regions of the sample space efficiently, without becoming trapped in local maximums. When a limited number of parametric distributions are used, Gibbs sampling is effective (Gelman et al., 1995). Here parameters are varied sequentially in proportion to the conditional probability defined by fixing the other parameter values. Another MCMC approach is Simulated Annealing, which controls the probability of the random walk being allowed to move to a lower probability location through the course of the sampling. This allows for greater probability space exploration at the beginning of the walk (Kirkpatrick et al., 1983). Alternately, Hybrid Monte Carlo sampling incorporates a gradient-based method; i.e., the walk proceeds in the direction which has produced probability increases in past steps, with a stochastic component which increases as the gradient decreases. This attempts to prevent the walk being trapped in local maximum by increasing the randomness of the walk at these points (Neal, 1996a). Importance Sampling is used to focus the sampling, potentially using an iterative search procedure (Neal, 2001), when there is a known region of interest within the sample space (Denny, 2001). These variations typically require additional information in order to appropriately tune the method to the problem under consideration. It has been argued that Slice Sampling, a non-parametric approach to Gibbs Sampling, can outperform other methods when this additional information is not available (Neal, 2003b).

1.3 Model calibration

The first portion of this thesis documents the implementation of a Bayesian calibration methodology using a simplified **General Circulation Model (GCM)**. One objective of model calibration is to addresses parametric uncertainty, which represents a major source of uncertainty for GCMs used for climate studies. This is done by using Bayesian inference to investigate plausible values for simulator parameters using observational targets; i.e., large scale climate features that form the basis for comparison between model output and observational data. By comparing simulator outputs against these observations, prior beliefs about reasonable parameter values are evolved into posterior distributions describing parameter suitability. Repeated sampling of model parameter values from these distributions allows for the creation of probabilistic projections for future events based on simulator output. Calibration, however, addresses more than parameter selection. As the formal routine accounts for all relevant uncertainties, this allows quantified statements about what may be inferred from the simulations. As such, error bars can be created for forecasts. Also, this enables applications which use these forecasts to account for input uncertainties. Without such uncertainty estimates, the practical usefulness of GCM simulations is limited. Bayesian calibration has been presented as an alternative to parameter optimisation for earth systems modelling (Wagener et al., 2001b). The latter approach does not allow for fully[1] probabilistic forecasting since it identifies only a single set of values, which are assumed optimal based on a selected metric. Such optimization approaches are also limited by their sensitivity to the chosen metric, without providing any criterion for comparison between alternative choices (Khu, 2005).

The practise of using Bayesian methods for calibration and probabilistic projection has been developed largely in the field of hydrology and other geosciences; e.g., Sambridge and Mosegaard (2002a). Early work on choosing computer simulation parameters based on observations was referred to as addressing the “inverse problem” (Mosegaard and Rygaard-Hjalsted, 1999; Sambridge, 1999); i.e., inferring a process from the outcomes it produces rather than direct observation. More formal procedures incorporating all the previously discussed uncertainties; cf., Kennedy and O’Hagen (2001), have been applied to simulations in many different areas of study; e.g., glacial modelling (Schaefli et al., 2005; Tarasov et al., 2012), agriculture (Hue et al., 2008),

[1] Ensembles of multiple optimised models using a variety of initial conditions are used to produce uncertainty estimates, although this method is more limited and less flexible than a truly probabilistic forecast.

and forestry (van Oijen et al., 2011). Bayesian calibration has been advocated as a rigorous approach to quantify uncertainties in climate forecasts. Early work in this area focused on applying data assimilation methods to the inverse problem; e.g., Hargreaves and Annan (2002) and Hargreaves et al. (2004). However, the formulation of these methods; e.g., the requirement of Gaussian priors, is typically too specific to allow them to be effectively extended to parameter estimation. Recently, large forecasting centres have worked to implement the hydrological methods, as described for climate modelling by Rougier (2007a), for their simulators; e.g., Sanso et al. (2008), Sanso and Forest (2009), Sexton and Murphy (2011), and Sexton et al. (2011). Outside of the examples given here, climate focused discussions of calibration are limited in the literature and there are many open questions regarding effective execution. The work presented in this thesis provides an example of the procedure within the context of climate modelling, while making preliminary tests of some alternate implementation approaches.

The computational expense of standard climate models is a major hindrance to the implementation of Bayesian calibration. Exploring the parameter space using MCMC methods requires running large numbers of individual simulations, with run times for most GCMs being on the order of days to months (Muller and Storch, 2004). The computational cost can be reduced by allowing certain portions of the calibration routine to be performed using statistical emulation[2] of the climate simulator response to changes in parameter values (Annan and Hargreaves, 2007a). Since it is not possible to know exactly how changes in parameter values will affect model output, a probabilistic representation of the inferred uncertain relationships is necessary. Such a representation allows a computationally efficient exploration of the parameter space

[2] There is a (not universal) convention in calibration literature to refer to the dynamical model being calibrated as a “simulator” and (if used) the empirical emulation of that model’s behaviour as an “emulator”. This leaves the term model to be applied more generally without confusion.

that takes into account the so-called code uncertainty, allowing it to be formally included in the Bayesian approach (Craig et al., 2001a). It has been argued that for certain applications an ensemble of well-calibrated low complexity models; i.e., computationally fast, might produce more useful results than a complex model that can only be emulated to a certain degree of certainty (Higdon et al., 2004). However, most climate forecasting objectives, such as determining tipping points or transient features, cannot be met with simulations simplified enough for direct MCMC sampling to be computationally feasible.

Typically, emulation is done using linear methods such as **Gaussian Process Emulators (GPEs)**, which are essentially an extension of the smoothing method known as kriging (Rougier et al., 2007b). In this thesis, I examine whether nonlinear **Bayesian Artificial Neural Networks (BANNs)** can be a viable alternative for calibration. Such networks have been used in a similar way for other applications; e.g., Khu and Werner (2003) Tarasov and Peltier (2005a). **Artificial Neural Networks (ANNs)** are a regression method where a network of nonlinear functions linked by prescribed weights and biases is used to map given inputs to expected outputs. These nonlinear components, which can be arranged sequentially in multiple layers, make the method more flexible than many empirical regression methods that are based on linear correlations; cf., Kanevski et al. (2009). It has been shown that ANNs are able to produce accurate emulations of GCMs (Knutti et al., 2003b). A single optimised ANN, however, does not describe the emulation uncertainties needed in the formal calibration framework. An ensemble of ANNs, where Bayesian inference is used to determine the network weights and bias, defines a BANN (Neal, 1996a). This enables the required estimation of the emulation uncertainties, and allows the BANN to be considered an example of a non-parametric statistical model; cf., Lee (2006). As well, hyper-parameters, which control characteristics of the distributions the weights and biases are drawn from, provide a means to regulate the complexity of the model. For many ANN applications regulation of model complexity; i.e., the smoothness of the prediction function it describes, is performed manually. This is often done through early-stopping; i.e., halting training when the ANN parameters become so specific to the training data that performance wrt. a separate set of test data is diminished. Alternately, this can be achieved by including terms in the training metric that penalise model complexity. The use of hyper-parameters fit through Bayesian inference allows a data driven means to adjust how specific BANN parameters are allowed to become. It has been argued that this technique, combined with the use of ensemble statistics that marginalise over parametric uncertainties, potentially makes BANNs more resistant to overfitting than ANNs fit to common maximum likelihood measures (MacKay, 2003). This potential has been demonstrated in experiments performed by Neal (1996a). However, BANNs are not immune from overfitting and care must be taken when evaluating predictions.

Under certain conditions, the properties of a single layer BANN can be shown to approach that of a GPE given an arbitrarily large number of components (Neal, 1996a). However, it is argued that, in practise, ANNs have the potential to outperform GPEs at extrapolating complex structures, and so can potentially create more detailed emulations from limited data, especially when they incorporate multiple layers (MacKay, 2003). Within an iterative search procedure, BANNs may provide better guesses about the location of unobserved high probability areas in the parameter phase space than is possible using linear interpolation. To test the feasibility of including BANNs in a climate simulator calibration, BANNs are used as the statistical emulator in all the calibration experiments presented in this thesis. This requires determining BANN structures that can emulate abstract statistical characteristics of the simulator output when limited to the amounts of training data typically available when calibrating a full GCM. Tests are performed on how to best structure iterative search routines to improve results from the BANNs' data learning capabilities. As well, to incorporate the BANNs into the formal approach, estimates of their emulation uncertainties must be made from the ensemble of emulator predictions. The results of these experiments will suggest whether further comparisons of climate model emulators are warranted.

While the main focus of calibration is to address parametric uncertainty, this requires estimates of structural model errors as well (Edwards et al., 2010b). In order to identify preferred model parameters, it is necessary to quantify how accurate the simulation has the potential to be. Attempts to fit models to a degree of precision beyond their abilities force parameter values to attempt to compensate for shortcomings that they were not designed to address. This results in simulations with much reduced or no predictive capability (Annan et al., 2002). Think of fitting a straight line with a fixed y-intercept of zero to data taken from a line with an intercept significantly different from zero. While it is possible to produce a line that passes close to a limited sample of data, this fit is not informative about the true trends. The further from the sample the fit line is used for prediction, the more inaccurate these predictions will be. As such, any uncertainty determined from the mismatch between the fit line and the sample will be inadequate to describe the prediction errors. A fit that reproduces the slope of the data while acknowledging that the observed values will actually be offset an estimated distance from the fit line is much more useful for prediction. In this case the prediction is always within the estimated uncertainties. In this simple example it's easy to say that the problem could be simply solved by a more complex model. However, the problem of structural error is unavoidable in all descriptions of complex and open real world systems, where no one model will be able to capture all relevant attributes. The challenge in addressing structural error is the creation of a rigorous statistical description of this uncertainty (Doherty and Welter, 2010).

Some researchers advocate that rigorous statistical descriptions of structural uncertainty are not possible and various alternate approaches have been suggested. One proposal is quantifying this error by performing incomplete Bayesian calibrations using increasing amounts of observational data until the simulator projections for a set of reserved observations no longer fall within a given tolerance level (Gaganis and Smith, 2001). At this point it's considered that the simulator parameters have been over-tuned to make up for structural deficiencies in the model and the tolerance is considered to be an approximation of the scale of the irreducible error. An alternative method is finding a separate optimal simulator setting for every given observation, and using this information to define the smallest possible volume of parameter space, over which the simulator is optimised (Gaganis and Smith, 2006). The difference between this compromise solution and the optimal solution for a given observation is considered the inherent structural error of the model for describing the given observation. A comparison between the methods shows that the first approach gives more conservative uncertainty estimates, and suggests that the optimisation scheme is only useful when observations are plentiful and it is possible for the simulator to recreate every individual observation (Gaganis and Smith, 2008). Neither approach is capable of identifying model biases as they do not acknowledge the possibility. Avoiding the issue of structural error, and instead calibrating a model within a given tolerance, does not allow simulations to be seen as ways to sample the state of current knowledge. Rather, they can only be seen as hypothesis to be evaluated, and most likely rejected (Beven, 2009). Or in the case of GCMs, certainly rejected. This approach may be pragmatic for highly specific risk assessment applications such bridge weight tolerance or determining the (non)existence of locations for nuclear power plants where safety can be guaranteed. It is, however, impossible to define such rigid performance criteria for exploratory and multi-objective studies of climate systems, whose nature preclude the possibility of fully comprehensive models.

For climate simulations, formal probabilistic structural error estimates are especially challenging to produce (Allen et al., 2006). These estimates are typically prescribed using subjective beliefs about the range of errors produced by unresolved model processes; e.g., Goldstein and Rougier (2009) and Holden et al. (2009), or by error estimates based on comparisons between the simulator and those of other modelling studies; e.g., Murphy et al. (2007a). The first approach is problematic as it may require a great deal of prior information beyond the current state of knowledge regarding the relevant systems[3]. The second depends heavily on ensembles of multiple GCMs, which should not be considered to be an unbiased estimate of current climate knowledge (Knutti et al., 2010) and require significant computational expense to generate. In this thesis an alternative and more formal approach is tested. The nature of the structural error is described using a few tunable parameters and values for these terms are evaluated along with simulator parameters as part of the model calibration. This method has been successfully applied to less computationally demanding geophysical modelling applications; e.g., Keats (2009). The experiments presented here use a very simplified description in order to perform an initial test of the methodology. It is examined whether these terms will evolve through the course of the calibration and that the posterior values do not underestimate mismatches between the calibrated simulator and observations. The limitations created by such a simple approximation are also observed. These investigations will suggest directions for further development of Bayesian estimates of structural errors.

Similar issues to those created by structural errors arise when fitting models to

[3] It is often argued that if these errors were understood well enough to make such estimates then they could be addressed within the simulation design.

imprecisely known observations, and so estimates of the uncertainties of the calibration targets is also required. Selection of calibration targets is dependent on simulator capabilities and application. They must be features which can be reproduced, in spite of structural errors, by the simulator. As well, they must relate to the variables which the simulator is being used to describe or predict, since the reliability of these forecasts will be estimated based on calibration results (Beven, 2009). As discussed, for climate studies targets are often statistical properties rather than particular events[4] or site chronologies. The next portion of this thesis deals with such features.

1.4 Atmospheric Circulation Regimes

Due to the low resolution of the simulator in the experiments in the first part of this thesis, the calibration targets are limited to climatological averages over continental scale regions. The second section of the thesis examines more subtle features such as regional atmospheric circulation regimes. These regimes are not directly observed as meteorological events or physical phenomena, but rather are mean states representative of certain patterns of atmospheric circulation; e.g., preferred storm track locations or typical locations of persistent high pressure features, that emerge at different scales. At decadal time scales, the frequency and residence time of such regimes is informative about the natural variability of the system (Corti et al., 1999). They are also informative about smaller spatial/temporal scale phenomena (Benestad et al., 2008). It is thought that external forcing from climate change drivers can influence the frequency, residency, and transitions between regimes (Palmer, 1999), while enacting limited change in spatial structural (Terray et al., 2004). As such, these regimes are

[4] Exceptions include paleo studies designed to examine climatic shifts or simulations of long-term events such as the North Atlantic great salinity anomaly. However, these events are shifts in large scale trends rather than individual meteorological occurrences.

potentially stationary features whose evolution and associations can offer more predictive and explanatory power regarding modern phenomena than more linear measures of climate evolution (Palmer, 1999).

One of the first documented examples of a chaotic nonlinear dynamical system is the Lorenz system; a simplified set of equations meant to approximate local atmospheric convection (Lorenz, 1963). One component of these systems are attractors; which can include quasi-steady-state solutions that system realisations tend towards without converging. These describe regimes of typical behaviour within the chaotic evolution of the system. An exhaustive study of the now well known Lorenz system is given by Sparrow (1982), and a general tutorial on the properties of chaotic systems can be found in Hale and Kocak (1991). The chaotic dynamics of the global atmosphere affect the predictability of the system on different time scales (Kalnay, 2002) On short time scales (days), sensitivity to initial conditions limits predictability. For longer time scales though, it is argued that observed central tendencies of these chaotic behaviours suggest atmospheric circulation regimes can be considered as elements of chaotic attractors (Palmer, 1999). Recent climate simulations suggest current atmospheric regimes remain stable under projected changes in external (anthropogenic emissions) forcing (Terray et al., 2004). However, theoretical studies have found that with sufficient forcing attractor structure can break down (Lorenz, 2006).

The dynamics of the earth-ocean-atmosphere system do not allow for analytical descriptions of attractors. Rather, statistical methods are used to look for multimodal behaviour within observed variability. Describing standard regional weather types has been practised since the late 19th century; e.g., Hanns (1887). Algorithmic detection of possible regimes is more recent. There are many approaches to describing such regimes and the following common methods represent a departure form earlier linear techniques of mapping teleconnection patterns; cf., Wallace and Gutzler (1981), and

Principal Component Analysis (PCA) (Preisendorfer, 1988b); e.g., Fraedrich et al. (1997), Bergant et al. (2002).

Identifying regions in the data phase space with high densities of occurrence through kernel density estimates has been used in early studies; e.g., Corti et al. (1999). The method estimates the **Probability Density Function (PDF)** of a multivariate distribution by considering each data point to be the centre of a Gaussian distribution with a preassigned width, summing the distributions and examining the resulting topography for peaks. The kernel widths can be estimated using rule of thumb formulations or through experiment (Roberts et al., 2012). Significance of results is determined by comparing outcomes against those obtained using the same parameters to model red-noise simulations. This allows for an estimate of the probability that the peaks observed in the data PDF are not artifacts of the estimation method.

Another approach is hierarchical clustering (Wilks, 2011). The technique initiates with each data point being considered an individual cluster. Then the two most similar, typically by a variant of Euclidean distance, elements are merged to become a new element. This process repeats until only one element remains; which is the mean of all the original elements. The approach produces many possible clusters and cluster combinations through a sequence of iterations. Significant modes from within the total set of produced clusters are considered to be the most “reproducible”; i.e., the routine is repeated with different subsets of data and the clusters that repeatedly appear in the analysis are considered the most informative (Cheng and Wallace, 1991). The algorithm is simple to implement but produces a large amount of data which must be subjectively analysed. Often it is used as means to check results of other methods; i.e., that the clusters created also appear within a hierarchical investigation (Cassou et al., 2004).

The K-means clustering method divides the data set into a predefined number of subsets (clusters) of similar elements. This results in cluster centroids; i.e., the mean of all elements assigned to the same cluster, which describe characteristic patterns common to their elements. The algorithm is designed to subdivide the data to maximise the distance between centres, while also maximising the similarity of the members assigned to individual clusters (Kaufman and Rousseeuw, 1990). The metric for similarity is the squared Euclidean distance, which results in spherical clusters of similar sizes (radii). Outcomes are examined using various methods, including cross-validation, testing random initial centres, and comparison with results produced from clustering red-noise samples with similar characteristics to the data (Cassou et al., 2004).

The hierarchical and K-means methods referred to as crisp clustering. These group data such that each data point is a member of exactly one distinct cluster. This can be described by saying that the membership of a given data point to a certain cluster is binary; i.e., either zero or one. Alternatively, fuzzy clustering allows a continuous range of membership to a cluster on the range [0,1]. The method grew out of the idea of fuzzy sets, originally motivated as a way to quantify imprecise uses of descriptors in casual language (Zadeh, 1965); e.g., often people do not consider all men above 174cm to be in the set of tall men, and those below in the set of short men, but use the term subjectively to imply a continuum. The fuzzy extension of the k-means clustering method is given by Bezdek (1981), and implementation algorithms are described by Kaufman and Rousseeuw (1990). Fuzzy clustering has been previously applied to identifying local circulation patterns (Ghosh and Mujumdar, 2006). Heuristics for evaluating results are similar to those of other clustering methods (Klawonn and Hoppner, 2003; Ghosh and Mujumdar, 2007). Fuzzy clustering is sometimes referred to as probabilistic clustering as the degree of membership can be interpreted as the probability that a data point belongs to a given cluster (Bezdek, 1981). It can be shown that the method is a nonparametric variant of formal probabilistic models; cf., MacKay (2003).

One such parametric probabilistic method are **Gaussian Mixture Models (GMMs)**. This approach describes a set of multivariate data by considering it as being generated from a combination of Gaussian distributions. The task is to estimate how many distributions comprise the sample, the percentage each distribution contributes to the data sample, and the mean and variance terms for each distribution (Fraley and Raftery, 2002). GMMs are conceptually quite similar to the described heuristic methods. The k-means algorithm and fitting a GMM where the covariance matrices are set to be diagonal, equal, and the same for all clusters, both depict drawing spheres within the phase space to define data groupings (MacKay, 2003). However, GMMs are a mathematically formal approach, fit to different metrics, that give a continuous probabilistic measure of membership. This gives some advantages over the other discussed clustering algorithms. It is trivial for GMMs to classify newly observed data points as they define membership probabilities continuously across the observation space. Once fitted, the model can be used for stochastic simulation by sampling from the defined PDFs. As well, they can be fit with Bayesian methods, giving a formal means for comparing models and describing parametric uncertainties. The parametric nature; i.e, the use of defined distributions, of the model can however overly restrict the form of the solution. As with all the methods described here, time correlation is not taken into account, limiting the usefulness of simulation. For the results of this method to be fully interpretable; i.e., to receive the full benefit of the model's formal structure, the modelled data would need to be temporally independent.

Bayesian implementations for GMMs have been developed (Neal, 1991). One variation is known as an Infinite Mixture Model. Rather than pre-selecting an initial number of clusters for the model, the number is set *a priori* to infinity, with an associated prior distribution, which takes the form of a concentration parameter, representing how diffuse the observed data is believed to be. This gives a means to calculate uncertainty for the number of clusters, but makes it difficult to make an ensemble estimate of the other parameters, since their number and meaning are different for each sample. Alternately, the number of clusters can be determined by comparing different potential values using Bayesian model comparison, which compares values for the Bayesian evidence term; see Sivia and Skilling (2006b). As this value is typically analytically intractable, an estimate can be computed using the **Bayesian Information Criterion (BIC)**; cf., Burnham (2004). Essentially this calculation rewards goodness of fit, but penalises the number of parameters; i.e., the principle of Occam's Razor. Use of this method for model selection is discussed by Lee (2006) for ANNs, and for selecting the form of GMMs in the context of climate data by Rust et al. (2010).

Clustering is not the only classification approach to have been applied in climatological contexts. Most of the described classification methods attempt to define only the most distinct groups within the data set. Alternately, **Self Organising Maps (SOMs)** create sets of similar states so as to highlight more subtle differences and transitions in behaviour (Kohonen et al., 1996). As this is not a clustering method but rather a form of discreet nonlinear regression there is the potential to invent features not actually present in the data. However, when effective, SOMs are able to present a more continuous and nuanced description of the features present in the investigated variables. SOMs present full maps of the same dimension of the data, as is the case for other clustering methods, referred to as nodes. These are arranged in a two-dimensional arrays, referred to as grids. The individual nodes are referenced by their x,y coordinate on the grid or by their sequence of occurrence when reading left-to-right top-to-bottom. Training occurs by finding the node most representative of a given data point, then adjusting this map and those of nearby nodes, relative to the grid, to match the data point based on set learning rates and decorrelation lengths. Training is repeated until convergence is reached. Performance is checked by the error between the final nodes and the data points they are considered representative of. As well, it is checked that nodes that are neighbours on the grid are in fact more similar to each other than to any other nodes. Hewitson and Crane (2002) provides an overview of the application of the method to the field of synoptic climatology. The method has been applied to classify weather patterns in the Arctic (Cassano et al., 2005) and Antarctic (Reusch and Alley, 2007) regions, as well as the North Atlantic (Reusch et al., 2007).

SOMs have been referred to as discreet analogues to other nonlinear equivalents to common data analysis techniques (Hsieh, 2004); i.e., PCA and Canonical Correlation Analysis. Here ANNs are used to create nonlinear functions to reduce a data field to a single variable and then map this variable back to the original data (Hsieh, 2004). Such methods have been successfully used to investigate El Nino Southern Oscillation processes (Monahan, 2001) and Northern Hemisphere atmospheric dynamics (Monahan et al., 2001), but it is suggested that their performance would suffer in the higher noise to signal ratios found in regional studies of higher latitudes (Hsieh, 2004). Recently, another nonlinear technique, Network Analysis, has been employed to analyse climate data. Typically used to describe phenomena such as cellular interactions, disease spread, and Internet systems, this approach considers data points as discreet nodes and looks to map the connections between them (Barabasi and Bonabeau, 2003). Systems that consist mostly of closely linked communities with limited external connections are referred to as complex networks (Steinhaeuser and Chawla, 2010). The method is applied to climate data by replacing the usual binary links between nodes with weights derived from lagged cross-correlations (Steinhaeuser et al., 2010). Investigations using this approach suggest that many large scale climatic phenomena can be thought of as occasionally intercommunicating subsystems (Tsonis and Swanson, 2012). While this method offers an alternative way to study climate phenomena, the results are mainly descriptive and, especially when they differ from classical analyses, difficult to interpret.

Weather regime searches using crisp methods have been conducted for different regions, using either **Sea Level Pressure (SLP)** and/or various geo-potential height fields. Studies over the northern hemisphere have found patterns relating to previously documented variability features (Corti et al., 1999; Monahan et al., 2001; Molteni et al., 2006). The significance of results produced over this broad region have however been questioned (Stephenson et al., 2004; Christiansen, 2007). Studies over different sections of western Europe by Corte-Real et al. (1998) and Casty et al. (2005) revealed common high pressure features and local manifestations of the **North Atlantic Oscillation (NAO)**. The latter is indicative of extra-tropical cyclone tracks across the North Atlantic (Vallis et al., 2004). It is classically defined as the anomaly in the difference in SLP for local measurements in Iceland and the Azores, although, it's broader regional structure has been documented by many studies and methods; cf., Hurrell et al. (2003)

The second section of this thesis focuses on describing atmospheric regimes for the North Atlantic region. It is believed that ocean circulation anomalies over this region and the recently observed variability in Labrador Sea temperatures (Yashayaev and Clarke, 2006), including some of the extreme observations from the last few years can be associated with atmospheric regimes (Zhu and Demirov, 2011). The Labrador Sea plays an important role in global ocean dynamics as a location for the deep water formation believed to partially drive the thermohaline circulation (Haine et al., 2008) and has motivated more detailed regional studies; e.g., Zhu et al. (2010). As such, descriptions of atmospheric processes from this region are particularly important to understanding global climate dynamics and represent important metrics for evaluating GCMs.

Studies of daily SLP anomalies over the North Atlantic region by using K-means clustering have identified NAO+/- regimes, as well as blocking regimes known as the **Atlantic Ridge (AR)**, and **Greenland-Scandinavian Dipole (SG)**, (Cassou et al., 2004). These are also observed by studies using regions shifted to focus more on Europe and alternate methods; i.e., GMMs, although these studies report additional; c.f., Terray et al. (2004); Rust et al. (2010), or fewer; c.f., Franzke et al. (2011), features of interest than the four commonly discussed. Discussion of the 'correct' number of patterns is ongoing and often dependent on what researchers are attempting to describe through these patterns. Work in this thesis focuses on describing the four patterns whose associations are most commonly documented in the literature, although Appendix B offers some alternative analyses using more involved classification methods, that tentatively support the observations of (Franzke et al., 2011). The four regimes, NAO+/-, AR, and SG, have been related to various regional climactic features. The NAO+ is linked to above average precipitation for North Europe and Eastern US and cold events for Eastern Canada, and the NAO- with above average precipitation for Southern Europe and the Canadian Arctic (Yiou, 2004). The Madden-Julian Oscillation[5] appears to be associated with shifts between NAO+ and NAO- over timescales of one to two weeks, possibly as a driving influence (Cassou, 2008). It has been argued that given the extreme dominance of the NAO- regime in the winter of 2009/2010 and the trends typically associated with the regime, that European winter was actually far milder than could be expected, despite many

[5] A cyclic pattern of eastward propagating moist convection that appears in the Indian and West Pacific Oceans (Zhang, 2005).

extreme cold events being observed. This has been interpreted as a climate change signal (Cattiaux et al., 2010). The SG is associated with extreme precipitation for East Greenland and the Mediterranean and the AR to high temperatures and precipitation for Newfoundland (Yiou, 2004) and decreased rainfall across the Iberian Peninsula (OrtizBevia et al., 2011). Associations between increased blocking events, which are poorly represented by the traditional NAO index, and warmer sea temperatures as well as the Atlantic Multidecadal Oscillation[6] have also been documented (Hakkinen et al., 2011).

In this thesis, I test whether the weather regimes documented for the North Atlantic can be reproduced using Bayesian GMMs. This approach should theoretically be able to reconstruct patterns of the form described by previous studies, while quantifying uncertainties in their spatial structure and classification. While applying a Bayesian method involves more detailed implementation and analysis than more commonly used algorithms, these features have been identified as potentially desirable calibration targets (Palmer, 2012). Use of such features within the calibration framework requires descriptions of observational uncertainty. Typical distributions of regimes may provide additional information about the system state (Michel et al., 2012). This is assessed in this thesis by classifying interannual atmospheric trends through fuzzy clustering. This non-parametric approach does not overly restrict the form of the results, while the calculated memberships levels allow classification uncertainties to be accounted for. It is examined whether the resulting modes are indicative of shifts in the distributions of weather regimes. These interannual modes are also compared against results from previous studies and ocean data. One of the motivations for considering long-term trends is that they are more directly comparable to the evolution of the ocean states that they are believed to be linked with (Marsh et al.,

[6] An observed oscillation between anomalously warm and cold Atlantic water with a period of around 70 years (Schlesinger and Ramankutty, 1994).

2008; Zhu and Demirov, 2011). Another is the observation that current GCMs do not provide consistent descriptions of regional weather regimes (Rust et al., 2010). This may be because the weather regimes are closely tied to detailed synoptic features that many GCMs struggle to represent in detail (Muller and Storch, 2004). Interannual trends are potentially more within the reach of lower resolution simulations.

1.5 Weather Generator

That the described regimes can be statistically linked to more limited spatial/temporal scale regional phenomena such as storm activity (Cattiaux et al., 2010; OrtizBevia et al., 2011), makes them useful as predictors which can be used to constrain projections for more variable local events referred to as predictands. GCMs typically underestimate variability in general and at mesoscale spatial and/or temporal scales in particular since they oversimplify or do not represent the related processes and interactions (Muller and Storch, 2004). The likely significant role of comparatively small scale variability as a driver of climate variation is still under discussion, cf. Frankignoul and Hasselmann (1977) and Monahan et al. (2010). Even when it is possible to accurately simulate climatic trends, it is often an open question how much this information constrains the potential range of the related smaller scale behaviours (Milliff et al., 2011). Describing the variability of these processes is required to further quantify the residual variability portion of simulation uncertainties. These studies are also needed to provide information for regional applications and risk assessment given that local responses can vary significantly from region to region (Benestad et al., 2008). While it may be optimal to couple as many linked processes as possible into unified simulations, the ability to do this is constrained by theoretical and computational limitations (Sato, 2004), especially when considering the need for ensemble simulations for uncertainty quantification.

Using larger scale features to make projections for smaller scale phenomena is referred to as downscaling. Downscaling can be done dynamically by using output from a global model as boundary conditions for a higher resolution **Regional Climate Model (RCM)**. Alternately, empirical-statistical downscaling uses data to derive statistical relationships between variables. This greatly decreases computational expense compared to RCMs, and avoids the boundary effects inherent to those models (Wilby et al., 2004). However, the method requires adequate data to determine the statistical relationships and the assumption that these relationships will be stable over the time period for which the model is to be used (Wilby et al., 2004). Many approaches have been implemented, including linear regressions, ANNs and non-parametric classifications (Benestad et al., 2008). A common example is the analogue method, which selects from a collection of historical states based on which occurred under large scale predictors most similar to current conditions (von Storch and Zwiers, 1984). There are many general reviews of downscaling methods including Murphy (1999), Campbell (2006), and Benestad et al. (2008).

Weather generators are a downscaling technique which produce stochastic simulations of the evolution of a small scale process conditioned on large scale predictors[7]. The focus on system evolution and the stochastic representation of unresolved processes means these models are not calculated directly from the predictor state, as is the case for many other downscaling approaches (von Storch, 1999; Benestad et al., 2008). Most examples of weather generators are focused on precipitation modelling. Simple examples involve fitting one or more standard distributions, typically Poisson-type, to rainfall amounts and defining rules to choose which distribution to sample from for a given interval; cf., Ferraris et al. (2003) and Maraun et al. (2010).

[7] Some authors; e.g., Wilby and Harris (2006), refer to weather generators which are conditioned on external factors as being hybrids between weather generators and regression models.

For the third portion of this thesis, a weather generator that describes the variability within given long-term atmospheric regimes is developed. The interannual regimes described in Section 1.4 are used as predictors to condition simulations of the daily features within the North Atlantic region. The method combines Hidden Markov and regression models to simulate state shifts and within state variability respectively; see Corte-Real et al. (1999a) and Furrer and Katz (2007) for examples of conceptually similar approaches. In Markov models, the probability of observing a given state is conditional on the previous states of the system. Hidden Markov models are an extension where there are multiple sets of these transition probabilities and a higher order process determines which will be in use for a given time step; cf., Rasmussen and Akintug (2004); Cappe (2005). A typical example, adapted from Rabiner (1989), is to consider the occurrence of a sunny or cloudy days. This can be modelled by a set of transition probabilities defining the odds that the next day will be sunny or cloudy based on the state of the previous day. Consider a region where cloudy days are more common in winter, and sunny days in summer. Thus, different transition probabilities are appropriate for different seasons. The changing seasons, which modifies the Markov properties, and thus the statistics of the observed sequence of events, is considered the hidden process. In this example the hidden process is quite regular, but typically it is also thought of as being stochastic or externally prescribed by an unknown mechanism. For weather generators the higher order, hidden, process is the evolution of predictor values, in this case the state of atmospheric regimes. The Markov process(es) describes the evolving state of the predicted variable.

Using identified atmospheric circulation patterns as downscaling predictors is common practise[8]; e.g., Corte-Real et al. (1999b), Bardossy et al. (2005), and Kannan and

[8] Boe et al. (2006) present an alternate bottom-up approach. They consider typical local events and then classify large scale predictors based on what predictor arrangements are most likely to result in a given behaviour. The method has been successful, but only for very specific regional settings; e.g., what sort of local atmospheric conditions are most likely to result in heavy precipitation on a

Ghosh (2010). Often large scale features are mapped to site specific observations; e.g., Semenov and Stratonovitch (2010). The goal, however, in this work is to reproduce the evolution of full data fields produced by reanalysis; i.e., high resolution models corrected with extensive data-assimilation (Kalnay et al., 1996a).

This approach is also known as empirical model reduction (Kravtsov et al., 2005). A review of different methods is given by Strounine et al. (2010). The preferred method, a lagged second order regression with stochastic noise terms is referred to as a **Linear Inverse Model (LIM)** (Kravtsov et al., 2010), by the developers who considered their approach to be a variation on earlier methods of the same name; c.f., Kravtsov et al. (2005). This method and naming convention is adopted in this thesis, but it should be noted that in the literature a LIM often refers only to models without higher order terms. Further comparison with other mid-range forecasting techniques is given by Hawkins et al. (2011). Conceptually, empirical model reduction is similar to model emulation, but with a different focus. The goal of this method is to produce self-evolving fields that match the observed variability, rather than to model how key features will change due to different simulator settings. Using this method to study unresolved behaviour within larger processes puts the work presented here in the framework of reduced order stochastic models. These have been used to investigate potential interactions in the earth system such as ocean responses to external forcing; e.g., Frankignoul and Hasselmann (1977), Monahan et al. (2010), and Prange et al. (2010). It has been argued, and for some cases demonstrated, that formally incorporating stochastic models into GCMs and other simulations to describe unresolved processes could effectively increase resolution without prohibitive increases in computational expense; see, Palmer et al. (2005), Jung et al. (2005), Wilks (2008), and Palmer (2012).

given day.

Typically weather generators produce values for a limited number of distinct local sites within the region of interest (Benestad et al., 2008), and so downscale both temporally and spatially. Alternately the evolution over a region is depicted by discreet transitions between finite sets of predetermined states through versions of the analogue method; e.g.; Boe et al. (2006). For the weather generator developed here this approach is augmented to produce continuous values over the entire region. The motivation for this is that the dominant sources of variability over the North Atlantic, extratropical cyclones, occur on spatial scales on the order of thousands of kilometres (Hoskins and Hodges, 2010) and are highly variable in their manifestation. An analogue type weather generator cannot sufficiently represent this variability, while a regression model of the full system requires more model parameters than can be determined by the amount of data available. This study experiments with the feasibility of combining the methods by using an analogue type model, developed from a SOM analysis, for basic features, and describing the residual variation with a regression model. Experiments are performed to determine a suitable regression method for the continuous portion of the model. Two methods are tested for creating the continuous portion: a LIM parametric linear approach and a BANN model. Testing different methods in different applications to determine when nonlinear approaches are advantageous is ongoing in weather generator research; e.g., Tang and Hsieh (2002), Hashmi et al. (2011), and Hawkins et al. (2011). Experiments using ANNs for forecasting and as weather generator components have produced mixed results; cf., Tangang et al. (1998), Tang et al. (2001) and Aguilar-Martinez and Hsieh (2009). The presented weather generator provides an estimate of the residual variability between the large scale features described previously and higher frequency processes. Such a model can be used for more detailed local studies within the region it describes. Also, as the weather generator is defined to behave differently for individual interannual regimes, comparing how these differences manifest structurally within the model may be informative about properties of the long-term modes.

1.6 Summary

As outlined, the organisation of this thesis follows a top-down trajectory. First issues relevant to quantifying global modelling uncertainties are addressed. Then, regional scale modes are examined, using different methods to describe observed features and the associated classification uncertainties. Finally, day to day local behaviours are described, looking to match the full range of variability despite limited information of related sub-processes and interactions. In practise however, it is difficult to separate the topics addressed in the various sections. Simplistic or missing descriptions of local sub-processes contribute to global model errors and biases. This makes estimates of unrepresented variability an important tool when testing the range of uncertainty induced by these necessary simplifications. The combined behaviour of these processes defines the global response to so-called external forcing, of which only a portion is produced physically external to the earth system, which creates feedbacks for regional subsystems. Regional subsystems interact across the climate system (Tsonis and Swanson, 2012) driving both global trends and local variation. Hence, reproducing and predicting the behaviour of these modes and interactions are important to both global modellers and local forecasters. The statistical approaches used here provide a common language for addressing various uncertainties across the different scales. It also makes possible stochastic simulations whose computational efficiency potentially allows for a wider range of investigations than more computationally exacting approaches. As well, a probabilistic framework puts an emphasis on constraining possibility rather than creating narratives of indeterminate relevance. That is, the less information/understanding we have the greater the range of possible outcomes we must accept, rather than creating false certainty by ignoring elements that we cannot explain.

As outlined, there are three main areas of study presented in this thesis. The first, described in Chapter 2, demonstrates a formal Bayesian calibration for GCMs, which is necessary to quantify the uncertainties in their projections. This experiment focuses on two areas of current research within calibration; emulation and estimating structural uncertainty, specifically

1. Testing the effectiveness of BANNs as climate simulator emulators given limited training data;

2. Testing the estimation of posterior distributions of parametrised structural error models in the context of a climate simulator.

The next section, described in Chapter 3, examines classifying patterns of atmospheric variability for the North Atlantic Region. This is an important region within the global climate system and so meaningful calibration targets and their observational uncertainties must be determined. The specific goals of this investigation are to:

3. Examine the reproducibility of published results using methods that better describe associated classification uncertainties;

4. Describe long-term shifts in the distribution of these regimes;

5. Relate these shifts to regional processes.

The final study, described in Chapter 4, is the construction of a local scale weather generator conditioned on the regime shifts described in the previous study. Such models are one way to estimate the residual variability between calibration targets and observations. As such, this study experiments with empirical model reduction and stochastic modelling of unresolved processes. The main objectives are to:

6. Determine a computationally efficient approach to creating realistic simulations of local variables for the sub-polar North Atlantic, that capture the range of observed variability;
7. Test if BANNs are needed to describe the residual between the discreet portion of the generator and the data to be simulated, or if this can be accomplished with a less opaque model;
8. Use this model to investigate the daily signals of the trends described in Chapter 3.

1.7 Thesis Overview

This thesis is written in manuscript format. Content is presented as three journal articles that have either been published elsewhere or submitted for publication. Because they are written as standalone articles there is overlap between the articles, and between them and material that has already been presented in the introduction. Note that occurrence of the terms “above” and “below”, in relation to the placement of information, refer to content within the individual chapters. To meet the requirements of Memorial University thesis guidelines each article is presented with its associated bibliography and, in addition, there is a bibliography for the entire thesis. Appendices and supplementary material for the articles is appended after thesis bibliography.

The original research papers appear in Chapters 2, 3 and 4. An overall summary of the body of work, and comments on envisioned future efforts are presented in Chapter 5.

1.8 Co-authorship statement

Authorship for the research paper presented in Chapter 2 is in the following order: Tristan Hauser (thesis author), Dr. Andrew Keats, and Dr. Lev Tarasov (thesis supervisor). Dr. Keats is currently a Senior Associate of Financial Advisory and Analytics with Deloitte. Dr. Tarasov is an Associate Professor with the Department of Physics and Physical Oceanography at Memorial University and holds a Canada Research Chair. Dr. Tarasov developed the initial idea and direction of the project. Implementation and analysis was performed by Mr. Hauser, with assistance from Dr. Keats. The manuscript was prepared by Mr. Hauser and critically reviewed by Dr. Keats and Dr. Tarasov, as well as by two anonymous reviewers.

Authorship for the research paper presented in Chapter 3 is in the following order: Tristan Hauser (thesis author), Dr. Entcho Demirov (thesis supervisor), and Dr. Igor Yashayaev. Dr. Demirov is an Associate Professor with the Department of Physics and Physical Oceanography at Memorial University. Dr. Yashayaev is a Research Scientist at the Bedford Institute of Oceanography. Dr. Demirov and Dr. Yashayaev developed the initial idea and direction of the project. Dr. Yashayaev provided observational data. Implementation and analysis was performed by Mr. Hauser. Interpretation of results was performed by Dr. Demirov and Mr. Hauser. The manuscript was prepared by Dr. Demirov and Mr. Hauser, with the exception of Figures 3.4 and 3.5, which were produced by Dr. Yashayaev. The manuscript was critically reviewed by Dr. Yashayaev.

Authorship for the research paper presented in Chapter 4 is in the following order: Tristan Hauser (thesis author) and Entcho Demirov (thesis supervisor). Dr. Demirov developed the initial idea and direction of the project. Model development, experimental design, implementation, analysis and manuscript preparation was performed by Mr. Hauser. The manuscript was critically reviewed by Dr. Demirov as well as by two anonymous reviewers.

The thesis as a whole as critically reviewed by Dr. Brad deYoung and Dr. Joel Finnis.

Connecting Text

The following article presents a more detailed discussion of a Bayesian formulation and implementation, in the context of a simplified GCM calibration problem. This provides examples of the uncertainties outlined in Chapter 1, with the exception of residual variability, being addressed within the approach. More specifically, this article addresses objectives (1) and (2), described in Section 1.6. This article has appeared as Hauser et al. (2011), in the journal *Climate Dynamics*. Additional discussion can be found in Section A.2.

Chapter 2

Artificial neural network assisted Bayesian calibration of climate models

2.1 Abstract

Earth systems models that attempt to make long-term predictions are sensitive to the approximations they employ. These approximations crucially depend upon model parameters whose values and uncertainties ought to be defined using objective and repeatable methods. In this study we approach this problem by using observational data to generate Bayesian posterior probability distributions for the model parameters. This allows us to determine high-probability parameter values along with their credible intervals, and accounts for the observational uncertainties related to the calibration data. For complex climate models, evaluating these distributions can require a prohibitive degree of computational expense. In the experiments presented here, Bayesian artificial neural networks (BANNs) are trained with output from a general circulation model (GCM) and used as statistical emulators of the full model to allow a computationally efficient Markov Chain Monte Carlo (MCMC) sampling of the Bayesian posterior of the GCM calibrated against seasonal climatologies of temperature, pressure, and humidity. Constraint data is categorized using principal component analyses of the observations. For these initial investigations we vary only five model parameters, which influence radiation, heat and momentum transport. We validate the methodology by calibrating to targets produced by a model run with added noise. A calibration is then performed to an observational data set. This requires us to incorporate a posterior assessment of the model structural error, which in turn allows the model to be used to make probabilistic forecasts for future climate states. All calibration experiments are performed with emulators trained using a maximum of one hundred model runs, in accord with typical resource restrictions imposed by computationally expensive models. We conclude by summarizing remaining issues to address in order to create a complete and validated operational methodology for objective calibration of computationally expensive models.

2.2 Introduction

Earth systems models are unavoidably incomplete descriptions of environmental phenomena. While the mathematical descriptions of the modeled processes are often very sophisticated and consistent with physical theory they inevitably contain approximations. Furthermore, given the complexity and nonlinearity of the Earth system, such models will invariably lack critical processes. 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 or field measurements. As Earth systems models generally have nonlinear dependencies on these parameters, determining appropriate parameter values and estimating the related forecast uncertainties is a challenging task. The models employed in the area of climate study are more complex and computationally expensive than those used in many other applications, with the nonlinear nature of these models further increasing the difficulty of identifying the relationship between parameters and model output. As well, the time and spatial scales considered in climate modeling make it difficult to prescribe with certainty appropriate calibration data. As a result, climate and Earth system models generally have a host of parameter values which have been fixed through subjective “hand tuning” to undocumented metrics. While this practice can result in models that provide arguably “reasonable” descriptions of the current climate system, it is unclear how accurately they will respond to changes in external forcing (Jackson et al., 2008). Furthermore, such hand-tuning precludes the determination of objective uncertainty estimates for model predictions.

One result of these issues is that similar models can often produce very different forecasts with no well-defined estimate of the degree of uncertainty in their predictions. Due to its nonlinear nature and dimensionality, the evolution of the climate system is impossible to describe in an explicit deterministic fashion. Therefore, policy discussions regarding climate issues are inherently about risk management. Unfortunately, the projections of current climate models are, for the most part, not presented in a form that allows for that type of decision making. These concerns have been expressed by many sources, including current reports by the Intergovernmental Panel on Climate Change (Solomon et al., 2007).

Due to the inherent (and difficult to quantify) divide between models and reality, as well as the uncertainties in observational data, calibration has often been viewed not singularly as a question of optimization, but as the probabilistic description of a range of parameter sets (and therefore model forecasts) in which the modeler has confidence (Wagener et al., 2001), which lends well to a Bayesian formulation of model calibration. Such formulations result in complex solution spaces which require numerical integration schemes such as Markov Chain Monte Carlo (MCMC) sampling methods (Mosegaard and Sambridge, 2002). Many variations of these methods have been developed and applied to a variety of geophysical inverse problems (Sambridge and Mosegaard, 2002).

It has been argued that the Bayesian calibration of climate models would provide a framework for addressing the above concerns, provided that (i) structural model errors (*i.e.* that can not be reduced by improved calibration) between the models and "reality" can be quantified (Rougier, 2007) and (ii) that the parameter space can be adequately sampled (Jackson, 2009). However, the computational demands of current climate models make most sampling routines unfeasible. Various sampling routines have been proposed and investigated (Jackson et al., 2004; Villagraon et al., 2008), but most are impractical for all but simplified models when computational resources are limited. One proposed method for coping with computational limitations is the use of statistical emulations of model response (Annan and Hargreaves, 2007; Rougier, 2008) to sample the model parameter space with Markov chain methods. One such type of emulator, Bayesian Artificial Neural Networks (BANNs), has been used in this way for different applications (Khu and Micha, 2003; Knutti et al., 2003; Tarasov and Peltier, 2005). However, there is little guidance available concerning the implementation of this method with regard to computationally expensive General Circulation Models (GCMs). The following work is an exploratory examination of its practical application to such models under the constraint of limited computational resources.

We will show that the BANNs enable Bayesian inference to be applied simultaneously to questions of calibration and model discrepancy despite computational resource limitations. This approach allows us to avoid basing structural discrepancy estimates solely on previously observed model errors and meta data concerning the processes being described (Murphy et al., 2007; Holden et al., 2010). As such, this approach is in contrast to using the Bayesian inference to assign weights to members of a much larger ensemble of model realizations; e.g., Holden et al. (2010), where the ensemble members are initially selected according to very general constraints, allowing the ensemble to span the believed range of the inherent model error (Edwards et al., 2010).

To describe these experiments we proceed first with a comprehensive overview of Bayesian inference as applied to the problem of model calibration using an emulator, with an emphasis on the probabilistic formulation. We also describe the practical steps taken to implement the method. Next we describe the model and data used in the presented experiments, with an outline of the model parameters selected for calibration and the method for calculating particular calibration targets from the available data. We then present the results of calibration experiments using “perfect model” and observational climatology targets. We conclude with a discussion of our findings and the issues raised in the course of the experiments.

2.3 Bayesian model calibration

In this section we describe our approach to the model parameter estimation problem, explaining the need for the model emulator and how it is addressed in the probabilistic formulation. This Bayesian formulation of the problem will allow us, in subsequent sections, to explicitly state the assumptions and information that will be used in the model calibration. The "objectivity" of the Bayesian approach does not imply that there is only one way to construct the terms used in the following equations, but rather that given this information, there is a framework to make reproducible and interpretable inferences from it. For example, there is no stipulation of what amount or type of observational data is to be used in the calibration. The probability density functions (PDFs) utilized represent a quantification of our state of knowledge, rather than claiming to be a complete statistical description of the system. Formally the form of the PDFs are always conditional upon the amount of information available to the individual who defines them[1].

2.3.1 Model parameter inference

Confidence in the values assigned to model parameters is expressed probabilistically based on comparison between selected model outputs $\boldsymbol{f}$, produced by running the model with model parameters set to $\boldsymbol{\theta}$, and corresponding observational data $\boldsymbol{z}$. This is expressed as $P(\boldsymbol{\theta} \mid \boldsymbol{z})$, which reads, "The probability of the parameter choice, given observations and background information." This probability distribution is constructed using Bayes' rule, so that:

$$P(\boldsymbol{\theta} \mid \boldsymbol{z}) \propto L(\boldsymbol{\theta}\,; \boldsymbol{z})\, P(\boldsymbol{\theta}). \tag{2.1}$$

[1] To emphasize this, some writers express such probability distributions in terms of $P(\cdot \mid I)$, where the "I" represents the information available to the individual; e.g., Jaynes (2003).

The distribution $P(\boldsymbol{\theta})$ is referred to as the prior as it is defined using prior information about the model parametrization. For example, in a situation where there is little physical basis for assigning proper parameter values the prior can be defined to give zero probability to all parameter sets known to result in physically unrealistic model output, and equal non-zero probability to all others.

The likelihood function[2] $L(\boldsymbol{\theta}\,;\boldsymbol{z}) = P(\boldsymbol{z} \mid \boldsymbol{\theta}) = P(\boldsymbol{z} \mid \boldsymbol{f}, \boldsymbol{\theta})$ expresses what the probability of observing $\boldsymbol{z}$ would be given that the model predicts $\boldsymbol{f}$, with the model prediction being tied to the choice of parameter values $\boldsymbol{\theta}$. Thus "high likelihood" parameter sets are those whose associated model output have a high probability "being close to" observations. The likelihood is derived by considering the existence of a true but unknown system state $\boldsymbol{y}$, which is then marginalized, so that:

$$\begin{aligned} P(\boldsymbol{z} \mid \boldsymbol{f}, \boldsymbol{\theta}) &= \int P(\boldsymbol{z}, \boldsymbol{y} \mid \boldsymbol{f}, \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{y} \\ &= \int P(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\theta})\,P(\boldsymbol{y} \mid \boldsymbol{f}, \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{y} \\ &= \int P(\boldsymbol{z} \mid \boldsymbol{y}) P(\boldsymbol{y} \mid \boldsymbol{f}, \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{y}. \end{aligned} \tag{2.2}$$

This formulation asserts that the relationship between the observations and $\boldsymbol{y}$ is independent of the model. This makes it possible to relate the model parameters to an unknowable "reality" using observed values, provided that judgements can be made on how to separately represent the relationships between both the observations and model to this "true state".

The posterior distribution $P(\boldsymbol{\theta} \mid \boldsymbol{z})$ represents a result of combining prior information with observational evidence. The resulting expression can not typically be calculated analytically. However, it is possible to evaluate the value of the posterior

[2] The expression $P(\boldsymbol{z} \mid \boldsymbol{\theta}) = P(\boldsymbol{z} \mid \boldsymbol{f}, \boldsymbol{\theta})$ results from $\boldsymbol{f}$ being a deterministic function of $\boldsymbol{\theta}$. However, it does not necessarily follow that $P(\boldsymbol{z} \mid \boldsymbol{f}, \boldsymbol{\theta}) = P(\boldsymbol{z} \mid \boldsymbol{f})$ unless the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{f}$ is one to one.

(up to an unknown constant) at a given value of $\boldsymbol{\theta}$. This makes it possible to determine high probability parameter sets by sampling the distribution using MCMC techniques, which generate parameter sets in proportion to the density specified by the posterior distribution. The resulting sample is then dominated by high-probability parameter sets. The algorithms for these methods often involve a large number of sequential evaluations of the posterior distribution. This is problematic as every evaluation of $P(\boldsymbol{\theta} \mid \boldsymbol{z})$ at a given value of $\boldsymbol{\theta}_i$ requires an evaluation of $\boldsymbol{f}_i$; i.e., the model must be run using a prescribed parameter set, with run times for many GCMs being on the order of days or weeks. When only a limited number of runs are possible, MCMC sampling may be implemented using an emulation of model output response to parameters of interest. The computational feasibility of such an approach will be shown below.

2.3.2 Expected model output

Approximating a PDF for the parameters; i.e., the MCMC sampling of the posterior, produces not only samples of desirable parameter sets, but also a measure of uncertainty. In order to see how this parametric uncertainty translates into variability in the model output, an ensemble of model realizations is created; i.e., multiple model runs are performed with the parameters of each run set to a different sample parameter vector $\boldsymbol{\theta}_i$ taken from the Markov chain. As it is necessary to assess the relationship between the model and the "true state" $\boldsymbol{y}$ in the construction of the Likelihood function (see above), this information can then be used to make statements about $\boldsymbol{y}$. It is also possible to extract statistics such as the ensemble mean and variance[3], which can be interpreted respectively as the expected model output and the model output uncertainty due to the uncertainty with which parameter values can

[3] This is if the resulting model output space makes such calculations appropriate; c.f. Sivia and Skilling (2006).

be assigned. Typically, the expected model output $E[\boldsymbol{f}]$ over the parameter space $P(\boldsymbol{\theta} \mid \boldsymbol{z})$ is approximated through the calculation:

$$\begin{aligned} E[\boldsymbol{f}\,; P(\boldsymbol{\theta} \mid \boldsymbol{z})] &= \int \boldsymbol{f}(\boldsymbol{\theta})\, P(\boldsymbol{\theta} \mid \boldsymbol{z})\, \mathrm{d}\boldsymbol{\theta} \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{\theta}_i),\ \boldsymbol{\theta} \sim P(\boldsymbol{\theta} \mid \boldsymbol{z}). \end{aligned} \tag{2.3}$$

As stated above, the required integral is typically not expressible in closed-form, hence the approximation which reads; "the expected output is approximated by the mean of n model runs, where the parameters for each run i are selected in accordance to (drawn from) the posterior distribution[4]." An MCMC method is used to sample from the posterior (i.e., to generate the actual values of $\boldsymbol{\theta}_i$).

MCMC sampling of the posterior requires a large number of model runs (a much greater number than n), which is unfeasible given the computational expense of Earth Systems models. Instead, the above approach is implemented using Bayesian Artificial Neural Networks (BANNs) as an effective non-linear regression of the model response to input parameter values. The BANNs are used in place of the actual model in the MCMC sampling, to permit algorithmic completion in an acceptable amount of time. Given that the emulators are by necessity a simplified approximation to the full model, we must now consider a posterior distribution for the model output $\boldsymbol{f}$: $P(\boldsymbol{f}, \boldsymbol{\theta} \mid \boldsymbol{z})$. As above we marginalize out the unknown term, so that:

[4] The symbol $\sim$ reads as "distributed as"; i.e., samples are concentrated according to the distribution P.

$$\begin{aligned} P(\boldsymbol{\theta} \mid \boldsymbol{z}) &= \int P(\boldsymbol{f}, \boldsymbol{\theta} \mid \boldsymbol{z}) \, \mathrm{d}\boldsymbol{f} \\ &\propto \int P(\boldsymbol{z} \mid \boldsymbol{f}, \boldsymbol{\theta}) \, P(\boldsymbol{f} \mid \boldsymbol{\theta}) \, P(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{f} \\ &= P(\boldsymbol{\theta}) \int P(\boldsymbol{z}, \boldsymbol{f} \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{f} \\ &= P(\boldsymbol{\theta}) \, P(\boldsymbol{z} \mid \boldsymbol{\theta}). \end{aligned} \tag{2.4}$$

The result is that our likelihood function must now represent the relationship between the model parameters and model output, as well as the considerations given above. For parameter sets at which we have run the model this can be done with complete certainty; i.e., $P(\boldsymbol{f} \mid \boldsymbol{\theta})$ is a delta function centred on $f(\boldsymbol{\theta})$. In other cases our ability to describe model outputs and "good" parameter sets is determined by our ability to make inferences about the model phase space. This is represented by the emulator predictions and their uncertainties.

2.3.3 Emulation using Bayesian artificial neural networks

Artificial neural networks (ANNs) are tools for performing nonlinear regression of complex systems, where a network of nonlinear functions described by prescribed weights and biases is used to map inputs to expected output. These networks can be made much more computationally efficient than the numerical model they describe and they can be implemented with a much weaker understanding of the underlying system dynamics. For the BANNs used here, the weights and biases are not single valued. Instead, the BANNs are actually ensembles of ANNs drawn from probability distributions derived by training the network against available data (in this case sets of Earth System Model parameters and associated Earth System Model output) using Bayesian inference as described above. The parameters defining these distributions are determined by MCMC sampling as part of the training routine. Note that this is a separate, though conceptually identical, occurrence of using Bayesian inference and MCMC methods to determine model parameters than that discussed elsewhere in this work. The implementation details are given by Neal (1996) and are beyond the scope of this paper.

This approach results in the BANNs being much more resistant to over-fitting than ANNs constructed by optimising values for the weights and biases (Neal, 1996). Furthermore, a trained BANN actually represents a posterior distribution for possible weights and biases values, given the training data. As a result the BANN prediction is comprised of an expected value together with an associated uncertainty (Lee, 2004). This property is why we refer to the BANNs as emulators. Other artificial neural network methods do not generally provide an uncertainty estimate and are therefore inappropriate for this application. We construct the BANNs using the freely available Software for Flexible Bayesian Modeling and Markov Chain Sampling suite[5] (Neal, 1996).

For ease of discourse in this work, the “architecture” of a BANN refers to the elements of the network directly selected by the user in order to improve emulation quality. For the present application these elements are: the number of hidden layers, the number of elements (size) of each layer and the degree of connection between layers. Explanations of these components can be found in Neal (1996). The link between the network architecture and the workings of the system or model that is being emulated is often vague at best[6]. This is a departure from some other approaches to emulation; e.g.; Goldstein and Rougier (2010), which focus more on statistically replicating the structure of the model in question. As a result, in practice compu-

[5] Available at `http://www.cs.toronto.edu/∼radford/fbm.software.html`

[6] It is commonly pointed out that ANNs can have very limited applicability when the goal is to describe the mechanisms behind an observed relationship; e.g., Sanderson et al. (2008).

tational capacity 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 (Lee, 2004). The fact that network construction is to a certain extent rather ad hoc is part of what makes the method flexible and tractable, although it makes it difficult to argue that the utilized network is optimal for the problem at hand. The optimal number of weights and biases required by an BANN depends on the size of the training dataset, the number of inputs and outputs addressed, and their actual relationships. The more non-linear the relationship between the input and the outputs, as well as the more uncorrelated the outputs, the larger and more complex the BANNs will need to be. Therefore, when it is desired to predict multiple outputs with limited training data it is often better to use separate BANNs for different outputs.

BANNs are only one of many available non-parametric regression methods. An overview of these and how they relate to BANNs is given by Lee (2004). The above formulation of the calibration problem can be applied using any emulation method; i.e., any method that can predict model response and give an estimate of the uncertainty of this prediction. It has been observed that as the number of elements of a single layer ANN approaches infinity, the ANN behaviour approaches that of a Gaussian process model (Neal, 1996), which have been successfully used for similar applications; e.g., Rougier et al. (2007). Additionally, it has been suggested that for some applications, BANNs with multiple hidden layers may possess additional attributes that allow them to outperform smoothing-based techniques MacKay (2003). The above, coupled with reports that BANNs can perform very well on high-dimensional, non-linear problems, where there is no known functional form relating inputs to responses (Lee, 2004), has motivated our decision to experiment with the method in this context.

2.3.4 Calibration procedure

The calibration follows an iterative implementation. An interim posterior generated by emulators trained on available data is used to select a new set of parameter vectors which is then used to construct additional model runs to add to the training suite. The iterative process allows these posterior distributions to gradually improve as the BANNs obtain more information about the model space. As the routine iterates, the amount of data available about the high probability portion(s) of the parameter space increases. It is sufficient, and much more efficient, to emulate the parameter subspace(s) of interest well, rather than to exhaustively recreate the entire phase space (Craig et al., 2001)., In detail this procedure is as follows:

1. Create initial ensemble of size m_0
   - Use the prior to generate the initial parameter sets: $\boldsymbol{\theta}^{(0)}_{1:m_0} \sim P(\boldsymbol{\theta})$
   - Run an ensemble of models $\boldsymbol{f}^{(0)}_{1:m_0}$ using these parameter sets
2. Augment the current ensemble using parameter sets identified by MCMC sampling[7].
   - for $j = 1 : N$
     - Create and train BANN (or BANNs) using all available input-output sets $\{\boldsymbol{\theta}^{(0:j-1)}, \boldsymbol{f}^{(0:j-1)}\}$ as training data
     - Generate a MCMC chain that samples from the Posterior, using the BANN emulator:

       $\tilde{\boldsymbol{\theta}}_{1:k} \sim P(\boldsymbol{\theta} \mid \boldsymbol{z}) \propto P(\boldsymbol{\theta})L(\boldsymbol{\theta}\,;\boldsymbol{z}),\ k \gg m_j$
     - Take new parameter sets $\boldsymbol{\theta}^{(j)}_{1:m_j}$ from post burn-in[8] $\tilde{\boldsymbol{\theta}}_{1:k}$

[7] Typically N, the number of times this process is reiterated, is determined by time and computational resources. Ideally this loop would repeat until a selected convergence criteria is met.

[8] The "post burn-in" portion of the sample are the samples that occur after the MCMC chain has converged.

   - Run the model using the new parameters sets to create new ensemble members $\boldsymbol{f}^{(j)}_{1:m_j}$
   - Augment previous ensemble with the new parameter sets and model realizations

3. Retrain the BANN(s) using all available data, in order to maximize their skill and further focus the posterior distribution. This distribution can be used to generate an ensemble of calibrated model runs $\boldsymbol{f}^{(j+1)}_{1:m_{N+1}}$, whose expectation (and higher moments) can be calculated following Eq. (2.3).

We apply the method iteratively so as to condition our beliefs on newly available information about the model; i.e., model runs at previously untested parameter sets, as well as from the observational data. In these experiments m is kept constant for each iteration (although this is not necessarily optimal) to allow experimentation with different values for m.

2.4 Model and observational data

As stated, the above described calibration methodology is a framework for making inferences based on a given data set. The selection of appropriate targets; i.e., the $\boldsymbol{z}$ in the above equations, can depend on many factors. These targets should be able to adequately constrain the model, and should be relevant to its spatial and temporal resolution, with the physical properties they represent appropriately resolved by the model (Müller and von Storch, 2004).

For the calibration exercises described here, the climate fields utilized are sea level pressure, surface temperature, and surface specific humidity as they relate to key model processes. As the model is run at a low resolution, we consider regional mean seasonal climatologies for these fields (herein calculated over the years 1958 - 2008).

2.4.1 Planet simulator general circulation model

The Planet Simulator is an Earth systems model of intermediate complexity (EMIC) developed by the Meteorological Institute of the University of Hamburg. This model solves the atmospheric primitive equations with a slab ocean model (Lunkeit et al., 2007a). For this study, it is run at a low resolution of T21 with five vertical levels. The model is forced with observed annual atmospheric CO_2 concentrations from the years 1958 - 2008 as provided by Tans (2009), and with data for present day ice extent, vegetation, and surface roughness. The Planet Simulator is run for the full cycle of fifty model years with this forcing (following a ten year initial spin up cycle, the results of which are discarded). For each model run, seasonal climatologies are calculated for each grid point, averaged from mean values over the fifty simulated years.

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 context that calibration is not being used 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. An overview of the role of these particular constants in the model equations can be found in Lunkeit et al. (2007a), and are described briefly as follows:

θ_1: 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.

θ_2: Used in the calculation of the ocean vertical diffusion coefficient for the three layer slab ocean model.

θ_3: Links time scale for divergence to damping time scale in the parametrization of atmospheric horizontal diffusion.

θ_4: Accounts for back scatter as proportional to the solar zenith angle in the calculation of cloud transmissivity for visible and ultraviolet Short Wave Radiation (SWR).

θ_5: Adjusts mixing length in equations for calculating exchange coefficients for momentum and heat in vertical diffusion relating to wind, temperature, and specific humidity. This calculation of vertical diffusion is used to approximate atmospheric turbulent exchange.

2.4.2 Data suite

The calibration target data fields used for the following experiments are regional mean seasonal climatologies for two meter temperature, sea level pressure, and two meter specific humidity. To validate the methodology, we first calibrate the Planet Simulator model against results from the model run at its default parameter settings with added noise to simulate observational uncertainty. The calibration data used in the second experiment is taken from NCEP/NCAR reanalysis data (Kalnay et al., 1996), and is transformed to match the locations of the model output as the reanalysis field is of higher resolution than the model. Note that if we were to use the reanalysis data to calibrate the same model that is used to conduct the reanalysis then we would break our assumption from Eq. (2.2), *i.e.* independence between observations and model. The observational uncertainties for the reanalysis data set were approximated from calculated regional inter-annual climatological variability.

2.4.3 Calibration targets

Given the low resolution of the model employed, the calibration data must first be reduced to an approximate synoptic spatial scale. Time centered Principal Component Analysis (PCA), as described in Preisendorfer (1988), is used to objectively select regions of the world to average over.

A time centered PCA[9] is used to identify orthogonal patterns of variance from the temporal mean of each global field. Projecting the data onto the leading (i.e., most statistically significant) basis vectors gives a projection coefficient vector which shows the dominance of an identified temporal variance pattern (the basis vector) at a given location. Therefore, these projection coefficient vectors can be used to identify similarly behaving regions, as locations of positive or negative values with respect to a given basis vector share a common trend of temporal variance that is either respectively in agreement with or in opposition to the basis vector. For the present analysis, each projection coefficient vector is used to create two regions of interest, one consisting of locations where the values are positive, the other where it is negative. The calibration targets for a given variable are taken to be the weighted means, for each season, of that variable over each region. The weights for each region are calculated from the respective portion of the projection coefficient vector which defines it by normalising the absolute values. Locations corresponding to the opposite portion of the vector are given a weight of zero, as are any weights bellow 1×10^{-5}. Thus the weights reflect how representative each location within a specified region is of the behaviour described by the relevant variance pattern; i.e., the original eigenvector. These regional means, along with the global mean values, become the 108 calibration targets; i.e., 4 seasons × 9 regions × 3 climate variables. These targets can be compared to model runs by calculating weighted averages of the model output within

[9] PCA is also referred to as Empirical Orthogonal Function Analysis

the above prescribed zones, using the same weights as used for the calibration data. The result for an analysis of NCEP/NCAR surface temperature data is seen in Figure 2.1.

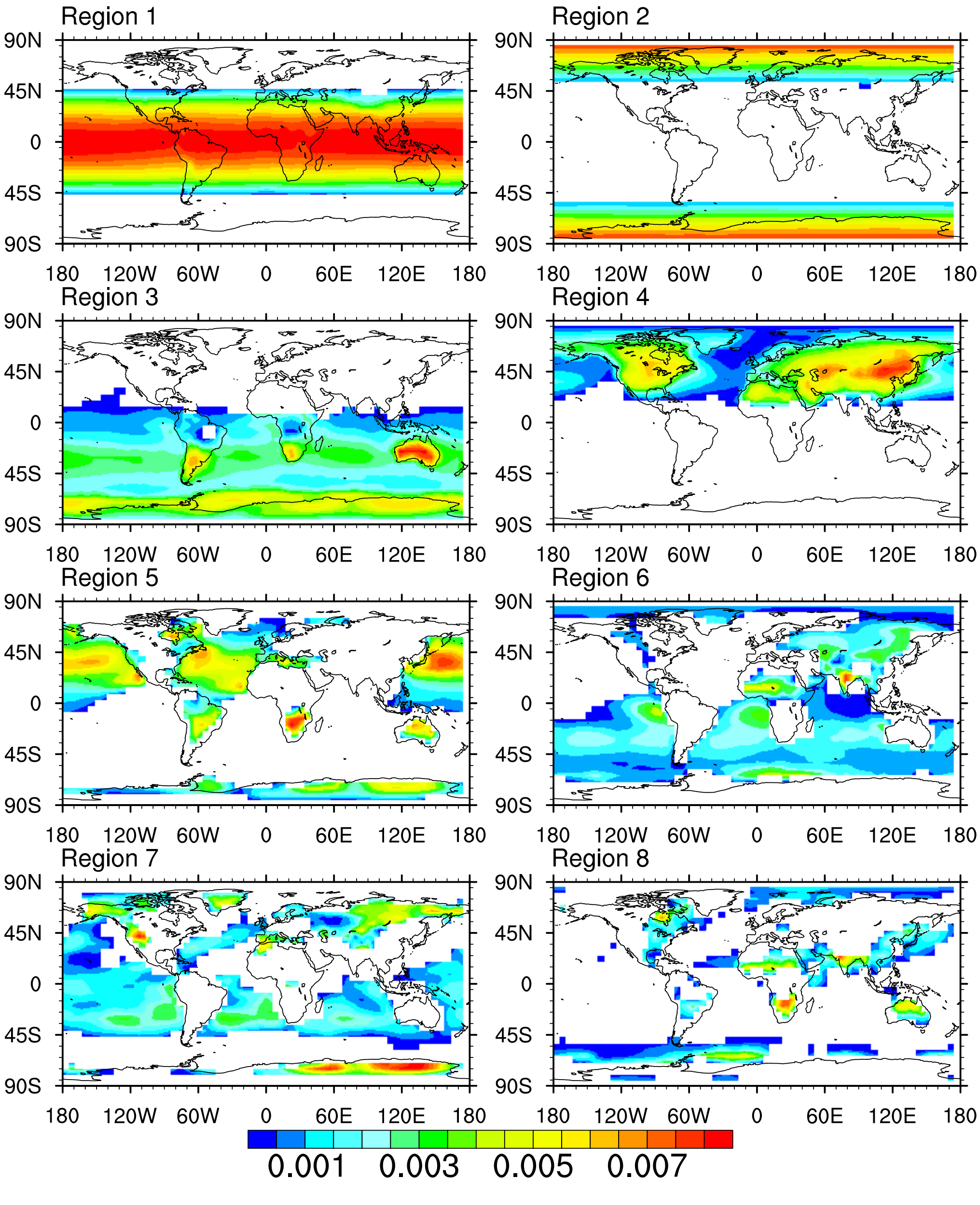


Figure 2.1: Regions averaged over to create surface temperature calibration targets. Colour bar displays weights used in calculating these averages.

A benefit of this method of data classification is that all of the data is utilised yet reduced to limit correlation, and so simplify the calibration. This choice of targets should drive the calibration to favour models that perform well over all climatic regions. However, this does not address correlation between seasons for like regions. Also, as with all PCA methods there is no guarantee that zones describe physically and not just numerically meaningful areas. For example, note in Figure 2.1 how the second region contains both poles. This is not because these regions have a similar seasonal cycle, but rather because their seasonal cycles correlate in being consistently colder than the global mean. Also, while the projection coefficient vectors are orthogonal, the vectors which define the regions are not. Therefore, regions are double counted, although potential complications are mitigated as they are being counted as part of different phenomena. For example, region two acts as a check that the poles are cooler than the mean global temperature, while regions three and four act to check that the seasonal cycles of the Northern and Southern hemispheres are different. Decisions must also be made on how many zones are to be used when defining the calibration targets. It is suggested (Preisendorfer, 1988) that it is required to use a large number of the projection coefficient vectors to account for all of the signal provided by a data set. In Figure 2.1 the vector that resulted in the first two zones was produced from the eigenvector that accounted for 99% of the variance, but the second two vectors also seem to have produced clear signals while the last set of zones appears more noisy.

2.5 Demonstration of calibration methodology

Two experiments are performed to demonstrate the calibration method. For the first experiment, the calibration targets are calculated from synthetic test data produced from the Planet Simulator GCM run at its default parameter settings[10]. This represents a simplified "perfect model" situation where the model is known to be able to reproduce the data, and at least one optimal parameter set is known to exist within the allowed parameter ranges. For the second experiment the calibration targets are calculated from the reanalysis data set. Therefore, there is no reason to believe a priori that there is any acceptable parameter combination where the GCM can entirely reproduce the calibration targets. This is addressed through the construction of a simple error model. We also give an example of how this calibration result can then be used to create a probabilistic forecast for a future event.

2.5.1 Prior distributions, likelihood functions and probabilistic forecasting

For this experiment prior ranges are set by multiplicatively expanding the range of each parameter around its default value, representing a "worst case" scenario of limited intuition as to what would be physically realistic values for the investigated parameters. Resulting ranges are given in Table 2.1 and designated by the labels given in the Planet Simulator code and documentation provided by Lunkeit et al. (2007b). To quantify this decision the prior distributions are set to be null outside the predetermined acceptable ranges and uniform[11] over a logarithmic scale within. As such, they are invariant under power law transforms, and so represent a uniform probability over all orders of magnitude. This formulation for the prior distribution, $P(\boldsymbol{\theta})$, is used in both calibration experiments.

As stated above, defining the likelihood function requires a representation of our

[10] Here we use the term "default" to refer to the values assigned in the original source code for the Planet Simulator model.

[11] While conceptually simple, it has been argued that uniform priors would rarely truly represent ones prior beliefs; c.f., Rougier (2007).

Table 2.1: Investigated parameters and their priors

	Parameter Name	Prior Range
θ_1	*acllwr*	$[0.05, 0.2]$
θ_2	*vdiffk*	$[1 \times 10^{-5}, 1 \times 10^{-3}]$
θ_3	*tdissd*	$[0.04, 0.8]$
θ_4	*tswr1*	$[0.02, 0.08]$
θ_5	$vdiff_{lam}$	$[80, 320]$

ability to relate the observational data and model outputs to the true state of the system. For the perfect model "observations" are created by adding Gaussian noise to the calibration targets, such that $\boldsymbol{z} = \boldsymbol{y} + \boldsymbol{\epsilon}$, where the elements ϵ_i are normally distributed with a mean of 0 and variance of $\sigma^2_{z_i}$, and are independent of each other. Therefore, we define the relationship between the observed and true state by a multivariate normal distribution[12]:

$$P(\boldsymbol{z} \mid \boldsymbol{y}) \triangleq N(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\sigma}^2_z). \tag{2.5}$$

As the "true state" of the calibration data is a model realization, there is no inherent discrepancy between model output and this state.

Emulation of model output is represented in a Gaussian formulation where the expected model output $\overline{\boldsymbol{f}}$ and its associated uncertainties $\boldsymbol{\sigma}_f$ for any given untested parameter set are taken to be the mean and one-sigma range respectively of the predictions sampled from the BANN for a given target (as discussed above), so that:

$$P(\boldsymbol{f} \mid \boldsymbol{\theta}) \triangleq N(\boldsymbol{f} \mid \overline{\boldsymbol{f}}(\boldsymbol{\theta}), \boldsymbol{\sigma}^2_z). \tag{2.6}$$

The formulations in Eqs. (2.5) and (2.6) allow us to solve the integral in Eq. (2.4) so

[12] The symbol $\triangleq$ reads as "equal by definition" and $N(\boldsymbol{z} \mid \cdot)$ represents a Gaussian distribution for $\boldsymbol{z}$.

that the likelihood function is represented as,

$$L(\boldsymbol{\theta}\,;\boldsymbol{z}) \triangleq N(\boldsymbol{z} \mid \overline{\boldsymbol{f}}(\boldsymbol{\theta}), \boldsymbol{\sigma}_z^2 + \boldsymbol{\sigma}_f^2), \tag{2.7}$$

where the $\boldsymbol{\sigma}_z$ are known explicitly and the $\boldsymbol{\sigma}_f$ are given by the emulator for each parameter set of interest. This representation of the emulator is in many cases a simplification of the non-parametric distribution of predictions generated by the BANN. As such, it represents a conservative (*i.e.* imposing as little structure as possible) formalisation of the information available to us.

For the second experiment we use the same model for the relationship between $\boldsymbol{y}$ and $\boldsymbol{z}$, except the $\boldsymbol{\sigma}_z^2$ terms are our approximation of the uncertainty of each calibration target derived from the reanalysis data (described above). Therefore, as for the emulator terms above, the expression $N(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\sigma}_z^2)$ does not necessarily describe the statistical structure of the data, but rather represents the amount of information that is available to us. Similarly, as we do not have evidence for a more complex error relationship between the Planet Simulator GCM and the true state of the Earth system that extends over all possible parameter sets, we use the same "truth plus noise" model to describe the expected misfit between model output and reality. In detail, $\boldsymbol{f} = \boldsymbol{y} + \boldsymbol{\rho}$, where each ρ_i has mean of 0 and variance of $\sigma_{M_i}^2$, so that:

$$P(\boldsymbol{y} \mid \boldsymbol{f}, \boldsymbol{\theta}) \triangleq N(\boldsymbol{y} \mid \boldsymbol{f}(\boldsymbol{\theta}), \boldsymbol{\sigma}_M^2). \tag{2.8}$$

Solving the integral in Eq. (2.4) using Eqs. (2.5), (2.6), and (2.8) results in:

$$L(\boldsymbol{\theta}\,;\boldsymbol{z}) \triangleq N(\boldsymbol{z} \mid \overline{\boldsymbol{f}}(\boldsymbol{\theta}), \boldsymbol{\sigma}_z^2 + \boldsymbol{\sigma}_f^2 + \boldsymbol{\sigma}_M^2). \tag{2.9}$$

As in Eq. (2.7), the terms $\boldsymbol{\sigma}_f$ are provided by the emulator, and the terms $\boldsymbol{\sigma}_z$

have explicit values for each individual data point and parameter set respectively. The model error terms, $\boldsymbol{\sigma}_M$, however, represent inherent model discrepancies over the entire potential parameter space, and so are very difficult (and computationally demanding) to assess. In keeping with the Bayesian framework, our procedure is to define these values as unknown parameters and consider them part of the solution space of the posterior distribution. Therefore, when performing the MCMC evaluation of the posterior we sample for both these terms and the model parameters; i.e., $\boldsymbol{\theta}, \boldsymbol{\sigma}_M \sim P(\boldsymbol{\theta}, \boldsymbol{\sigma}_M \mid \boldsymbol{z}) \propto N(\boldsymbol{z} \mid \overline{\boldsymbol{f}}(\boldsymbol{\theta}), \boldsymbol{\sigma}_y^2 + \boldsymbol{\sigma}_f^2 + \boldsymbol{\sigma}_M^2)\, P(\boldsymbol{\theta}, \boldsymbol{\sigma}_M)$. This approach can dramatically increase the dimension of the calibration problem. As described above, we investigate five model parameters, using 108 calibration targets, making $\boldsymbol{\sigma}_M$ a vector of length 108. Additionally, if we consider the possibility of describing correlations between model errors, (e.g. the possibility that locations in the model that are prone towards erroneous spring temperatures are consistently the same locations that display a similar degree of error in summer) then we face the prospect of estimating a 108×108 member covariance matrix. There are trade-offs to consider concerning the merits of such an extensive investigation. The more complex our posterior distribution becomes, the more computational limitations hamper our ability to approximate it accurately; e.g., we require greater accuracy from the BANNs, enough MCMC samples to be sure of convergence, etc. As well, our limited ability to explore the model space also reduces our ability to acquire prior information regarding the nature of the model error. These issues emphasize the importance of decorrelating the constraint data as much as possible. For the experiment described here, we adopt a much cruder description of model error, that more accurately reflects the current limited state of our understanding of the model and the scope of the investigation we are able to conduct. We define $\boldsymbol{\sigma}_M^2$ as $[\sigma_H^2, \sigma_P^2, \sigma_T^2]$, where $\sigma_H^2 \triangleq$ expected squared model error for specific humidity, $\sigma_P^2 \triangleq$ expected squared model error for sea level pressure, and

$\sigma_T^2 \triangleq$ expected squared model error for surface temperature.

Priors for these parameters were specified by assigning $\log(\sigma_H)$, $\log(\sigma_P)$, and $\log(\sigma_T)$ Gaussian distributions with means and variances estimated from error statistics for the initial spread of model realizations[13] Employing such a general description of model error will likely produce assessments of $\boldsymbol{\sigma}_M^2$ that are more conservative than those that might result from a more accurate description that captures the covariant structure of the errors. However, an estimation of large uncertainty is appropriate given the limited understanding of the system under consideration. This is preferable to ignoring model discrepancy and so obtaining a false estimation of high confidence in the model result, or using an inappropriately detailed description that would be overly sensitive to our (admittedly ill-informed) prior judgements.

A primary motivation for earth systems modeling is the desire to define inferences about some element(s) of the earth system, $\tilde{\boldsymbol{y}}$, that are conditioned on our observations of the physical system $\boldsymbol{z}$; i.e., to be able to estimate $P(\tilde{\boldsymbol{y}} \mid \boldsymbol{z})$. The calibrated model and estimates of its relationship to reality are the means by which these inferences are made. For the example presented here the probability for any potential value of $\tilde{\boldsymbol{y}}$ can be found by considering the joint probability between $\tilde{\boldsymbol{y}}$ and model parameters $\boldsymbol{\theta}$ conditional on $\boldsymbol{z}$ and marginalizing $\boldsymbol{\theta}$, so that:

[13] By working with the logarithm of the standard deviation, we essentially consider the probability of $\sigma/2$ to be equal to the probability of 2σ. This type of treatment is appropriate when describing 'scale' parameters whose uncertainties are relative rather than absolute Sivia and Skilling (2006)

$$\begin{aligned}
P(\tilde{\boldsymbol{y}} \mid \boldsymbol{z}) &= \int P(\tilde{\boldsymbol{y}}, \boldsymbol{\theta} \mid \boldsymbol{z})\,\mathrm{d}\boldsymbol{\theta} \\
&= \int P(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{z})\, P(\boldsymbol{\theta} \mid \boldsymbol{z})\,\mathrm{d}\boldsymbol{\theta} \\
&= \int P(\tilde{\boldsymbol{y}} \mid \boldsymbol{f}, \boldsymbol{\theta}, \boldsymbol{z})\, P(\boldsymbol{\theta} \mid \boldsymbol{z})\,\mathrm{d}\boldsymbol{\theta} \\
&= \int N(\tilde{\boldsymbol{y}} \mid \boldsymbol{f}(\boldsymbol{\theta};\boldsymbol{z}), \boldsymbol{\sigma}_M^2(\boldsymbol{\theta};\boldsymbol{z}))\, P(\boldsymbol{\theta}, \boldsymbol{\sigma}_M \mid \boldsymbol{z})\,\mathrm{d}\boldsymbol{\theta} \\
&\approx \frac{1}{n}\sum_{i=1}^{n} N(\tilde{\boldsymbol{y}} \mid \boldsymbol{f}(\boldsymbol{\theta}_i;\boldsymbol{z}), \boldsymbol{\sigma}_M^2(\boldsymbol{\theta}_i;\boldsymbol{z})), \\
&\quad \boldsymbol{\theta}_i,\ \boldsymbol{\sigma}_M^2(\boldsymbol{\theta}_i;\boldsymbol{z}) \sim P(\boldsymbol{\theta}, \boldsymbol{\sigma}_M \mid \boldsymbol{z}), \qquad (2.10)
\end{aligned}$$

where $n =$ the number of members of the forecast ensemble and $\boldsymbol{f}(\boldsymbol{\theta}_i;\boldsymbol{z})$ is the model forecast for the quantity $\tilde{\boldsymbol{y}}$ with the dependence on $\boldsymbol{z}$ maintained by using model parameters and associated discrepancy terms[14] prescribed by the posterior $P(\boldsymbol{\theta} \mid \boldsymbol{z})$. As the posterior incorporates parametric and observational uncertainties any final estimate of $\tilde{\boldsymbol{y}}$ is conditioned on these as well as $\boldsymbol{\sigma}_M$. The use of the Gaussian distribution in the fourth line follows from Eq. (2.8) and so is particular to this exercise. Note that for this example the model forecast is assumed to be an unbiased estimator of $\tilde{\boldsymbol{y}}$. It is also assumed that (a component of) the estimated discrepancy relationship between the model and the calibration data can be directly applied to $\tilde{\boldsymbol{y}}$. In practice, depending on the model used and the nature of $\tilde{\boldsymbol{y}}$, these assumptions (as well as those made concerning the structure of the observational and emulator errors) will potentially be quite tenuous, especially when predicting future events. As the experiments presented here are meant to form a baseline demonstration of the method, we adopt the simplest possible form for all the assumptions made in this section, even when we may have access to additional information (as seen below). In many cases

[14] Here the discrepancy terms are written $\boldsymbol{\sigma}_M^2(\boldsymbol{\theta}_i;\boldsymbol{z})$ as a reminder that they are sampled jointly with $\boldsymbol{\theta}_i$ from the posterior, and so are distinct for each parameter set.

a more complex, and potentially subjective (Holden et al., 2010), error model will be required, or if this is not possible, the functionality of the model as a descriptor of $\tilde{\boldsymbol{y}}$ will have to be reassessed.

2.5.2 Perfect model experiment

Here we present the implementation details and results of calibrating the selected model parameters to the model run at its default parameter values. In order to assess the functionality of the methodology under realistic computational limitations we restrict ourselves to running the model no more than one hundred times for any given experiment. We break up our ensemble into batches of size m following the routine outlined above, and experiment by creating three different ensembles using values of $m = 20$, $m = 30$, and $m = 50$. We refer to these throughout the text as Ensembles A, B, and C respectively. This gives us an idea of how the quality of the calibration degrades with ensemble size, and allows us to investigate whether benefit is gained by increasing the number of times the routine is iterated, even if this involves training the emulators with a reduced amount of data at each iteration.

2.5.2.1 Implementation

As outlined above, the first step in the calibration routine is to create an initial ensemble of model runs. Parameters for these initial model runs were selected from the prior through Latin hypercube sampling. This form of sampling has been shown to be an effective method for selecting emulator training data (Urban and Fricker, 2010). We generate 100 initial hypercubes and utilize the one with the maximum minimum distance between its members, although more developed algorithms are available for this task; c.f., Grosso et al. (2008).

For this experiment, multiple BANNs were needed to successfully approximate model response. Nine BANNs are used for each climate variable, one for each calibration data region, giving a total of twenty seven individual networks. Initially, for each location, multiple BANNs, with architectures of varying complexity, are created and compared, with the architecture that results in the best emulation being selected to represent the location in the calibration procedure. Ideally the quality of this emulation would be assessed with independent test data. However, since our imposed limitations leave us with very little training data to begin with, rather than reserve some of this information for testing, we take advantage of the nature of the BANNs and their resistance to over fitting and judge their quality based on their mean ability to recreate the training data. While this is not a good measure of the BANNs' predictive ability, it does allow us to identify network architectures that are able to reproduce the system we wish to describe[15]. Where similar results were obtained with different architectures, selection was motivated by the desire to use emulators with varying degrees of complexity.

All the BANNs have as their inputs the five parameters discussed above. As each BANN is trained to express all the data of a specific region, each produces a four element output vector; i.e., the regional average value for each of the four seasons of the associated climate variable. We consider it appropriate, and perhaps even beneficial (MacKay, 2003), to use a single network for the entire temporal output of a region, as these regions are selected on the basis of their having distinct seasonal cycles for the variable in question (see above). Thus, these values will potentially be related to each other and so simplify the non-linear relationship the BANN must approximate. However, the potential for a resulting correlation between the outputs

[15] A more prudent approach is to perform a cross-validation, where a different element of training data is reserved each time and used to test the predictive ability of the resulting BANN. While this does require additional computing resources, it requires less than running a full GCM only to find out that it was calibrated based on the predictions of a poorly performing emulator. Alternatively, the Bayesian structure of the emulator does allow for more sophisticated methods of model selection, including comparison of Bayes factors or use of the Bayesian information criterion (Lee, 2004).

is not reflected in the description of emulator error above.

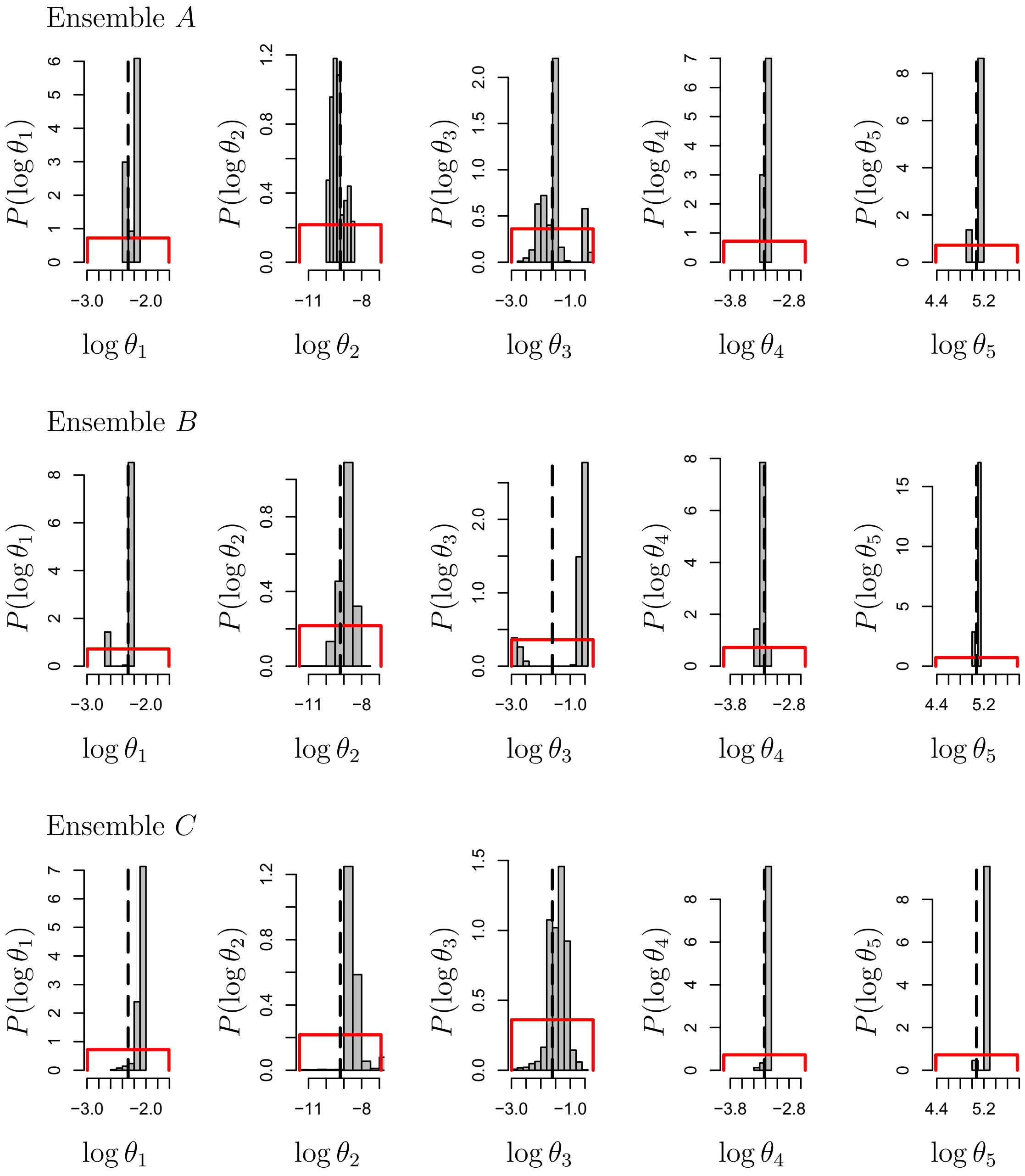
The trained BANNs are incorporated into the likelihood function (described above) and MCMC is used to sample from the resulting posterior distribution. Slice sampling was selected for the MCMC routine since it can be easily implemented and adjusted for efficiency despite limited knowledge of the form of the distribution to be sampled (Neal, 2003).

In keeping with the outlined calibration procedure, parameter sets from the resulting MCMC chains are used for subsequent model runs, which are then used to extend the model ensemble. Data from the resulting (and previous) ensemble(s) are used to retrain the BANNs. These are then used in the generation of new MCMC samples. Samples of parameter sets extracted from these are used to further extend the ensemble, and so on until we reach our one hundred run limit. When the model is rerun as the calibration routine reiterates, its output is used to check the performance of the BANNs. The final posterior distribution is obtained as a result of training the emulators against the entire, iteratively generated ensemble.

2.5.2.2 Results

In the context of the marginal probability distributions for individual parameters resulting from the final posterior distributions (Figure 2.2), the parameters used to construct the synthetic data are all within one standard deviation of the means of these distributions for Ensemble A (and generally very close to the mode of their marginal distribution). Marginal probability densities are not necessarily representative of the true shape of the full N-dimensional (and thus very difficult to visualise) posterior distribution. Investigations of two and three dimensional marginal posteriors (not shown) offer little further information, aside from a strong linear connection between the values for θ_1 and θ_4. This is not unexpected, as both are related to the effects of clouds on radiative fluxes. Still, these results suggest that the method performs well, given that a set of parameters that we know deserves high confidence is so represented within the posterior. This is also the case for Ensemble C, although the result for Ensemble B is problematic for θ_3. For most model calibration scenarios there will be no "true" parameter set, and further analysis (not shown) suggests that even for this example the relationship between fit to targets and distance from the "perfect model" parameter set is highly non-linear, giving reason to believe that there are various distinct parameter sets that will produce similar output for the model, and/or values for the calibration targets.

Figure A.3 shows the evolution of the emulators' ability to predict the model response to the selected parameter sets. As all fields showed similar behaviour only the temperature field is shown, both to simplify the presentation and as its results are the most pronounced. It appears that initially the emulators were unable to consistently emulate the model. Initially the BANN also fails to provide accurate assessments of the prediction uncertainties for any of the ensembles; e.g., for Ensemble A only 27% of the emulator errors were below the 3σ level of the corresponding predicted uncertainty. Further iteration results in improved overall estimates of uncertainties, and in the case of Ensemble A, improved accuracy as well. This suggests that when the total number of runs is limited, there is more benefit to be had in allowing the emulator to "learn from its mistakes" through an iterative process than there is in providing it with large amounts of initial data. We find that by the final iteration of Ensemble A, 93% of the emulator errors are below the 3σ level of the associated predicted uncertainty. However, the remaining outliers can be very far from the mean prediction, up to a distance of 35 times the respective predicted standard deviation. Note that this ratio represents an extreme underestimation of uncertainty, but not necessarily an extreme error. This suggests that the Gaussian approximation of the



Ensemble A

Ensemble B

Ensemble C

Figure 2.2: Histograms show final marginal posterior densities for model parameters (in log scale) as estimated in the perfect model experiment. Red lines show the prior distributions for the parameters, black dashed lines show the parameters used to create the synthetic data. Top row is the result of performing five iterations of the calibration routine, using twenty model runs apiece, middle is result of performing three iterations using thirty model runs apiece, and bottom is result of performing two iterations using fifty runs apiece.

BANN distribution used in Equation 2.6 is too restrictive, and that it would be more appropriate to use a “heavier tailed” distribution in its place. While not ideal, after repeated iterations the representation does capture the bulk of the uncertainty (refer to Appendix A for further discussion of the relationship between the emulator error and its estimation), and does provide quite (when compared to the scale of the targets) accurate predictions. So in practice (for this exercise), this does not preclude reasonable results. As a technical aside, we find that the parameter to target relationship is sufficiently nonlinear that regardless of initial fit to training data, multiple hidden layers are required to ensure the possibility of predictive ability.

To track the progress of the calibration we calculate the natural log[16] of the likelihood function, Eq. (2.7), for each model run produced by the calibration, except here the model output is produced from the GCM itself (rather than from the emulator) and so the $\boldsymbol{\sigma_f}$ term is ignored. Figure 2.4 shows the evolution of this measure of misfit between the ensembles and the calibration data as the calibration routine iterates. The comparatively good fit for Ensemble B at iteration two suggests that the emulators’ ability to predict their own errors was sufficient to prevent the targets for which they have low skill from overly affecting the calibration (82% of the errors are constrained by the 3σ level of the associated predicted uncertainty), although it also suggests limitations in the targets’ ability to constrain the calibration. The overall result shows that given reasonable performance by the emulators, the calibration routine can identify parameter sets that produce better fits to targets than can be expected to be discovered through Latin hypercube sampling alone.

[16] The logarithm is used to avoid computational round off errors.

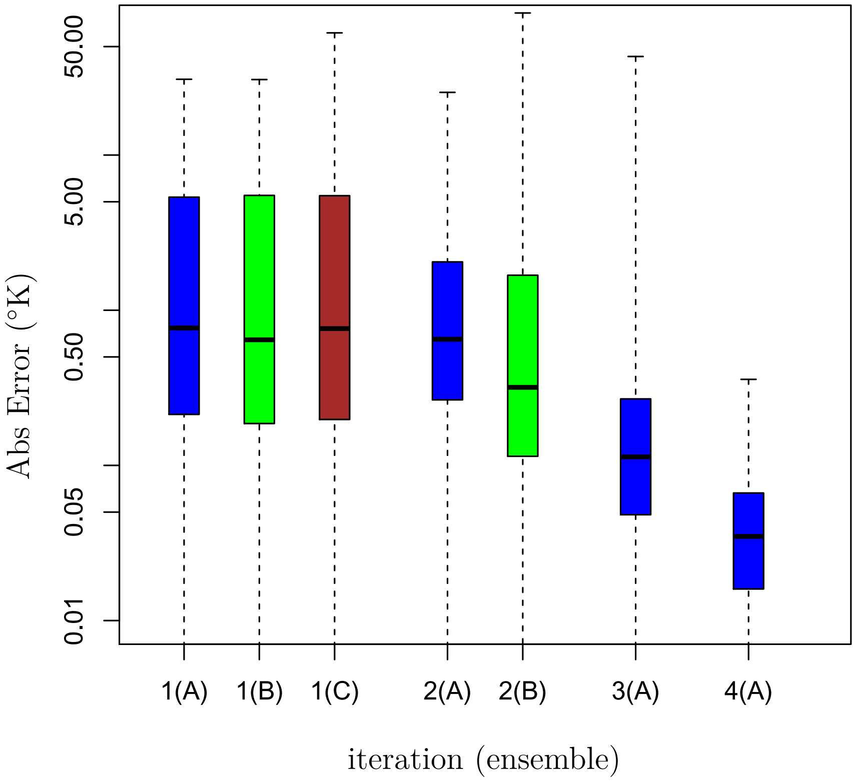


Figure 2.3: Spread of absolute errors (y-axis, log-scale) between actual model output and emulator predictions thereof for each iteration (x-axis, subdivided by ensemble) of the perfect model experiment, are displayed using box-plots depicting quartile values. Ensembles *A*, *B*, and *C* are represented by the colours blue, green and brown.

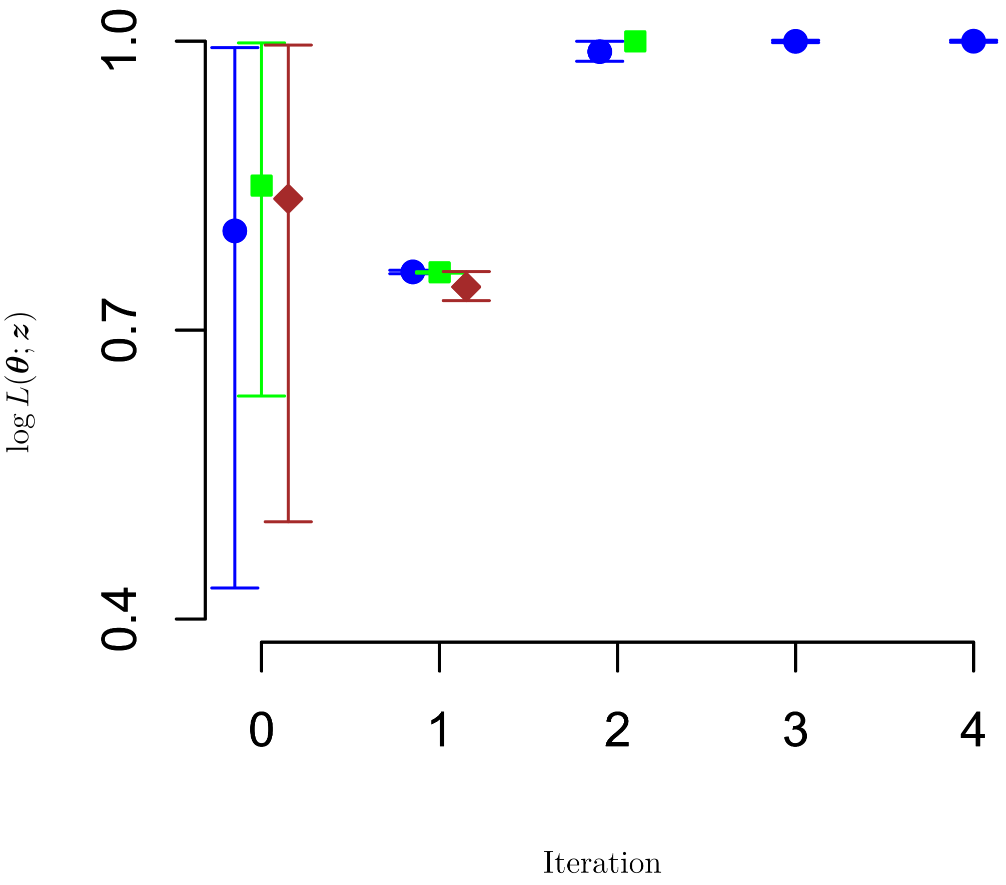


Figure 2.4: The mean of the calculated log-likelihood values of the model runs (y-axis, scaled to a range of [0 : 1]), bracketed by their 10% and 90% quantiles, produced at each iteration of the calibration routine (specified on the x-axis, points are offset for clarity), for the perfect model experiment. Ensembles A, B, and C are represented by the colours blue (circle), green (square), and brown (diamond), respectively

2.5.3 Calibration to reanalysis data

Here we present the implementation details and results of calibrating the selected model parameters to targets calculated from the NCEP/NCAR reanalysis data.

2.5.3.1 Implementation

The same procedure outlined above is followed when calibrating the model to the reanalysis data. We refine our initial search for suitable BANN architectures based on results from the previous experiment.

2.5.3.2 Results

The emulation quality for the three ensembles is more consistent than in the above experiment as shown in Figure A.4, although here again we obtain better performance from repeated iterations, even if this requires smaller sample sizes to be used. By the final iteration of Ensemble A, 92% of the emulator errors are below the 3σ levels of the predicted uncertainty. However, the presence of outliers far beyond the estimated uncertainty is again observed. This, as well as investigation of the distribution of ANN responses that comprise the individual BANN predictions, further support our suspicion that the Gaussian approximation of Equation 2.6 is too restrictive to fully describe the BANN behaviour (refer to the supplement for a QQ-plot based consideration of this).

There is a faster and more uniform convergence to comparatively high likelihood model output (Figure 2.6) for all ensembles. As such it is not surprising that the marginal distributions (Figure 2.7) show similar behaviour between all ensembles. While it is in general impossible to assess the true nature of the distribution from the marginal distributions, we see that the latter are most focused for Ensemble A, and express the lowest expected model discrepancy values (Figure 2.8). This is sensible as

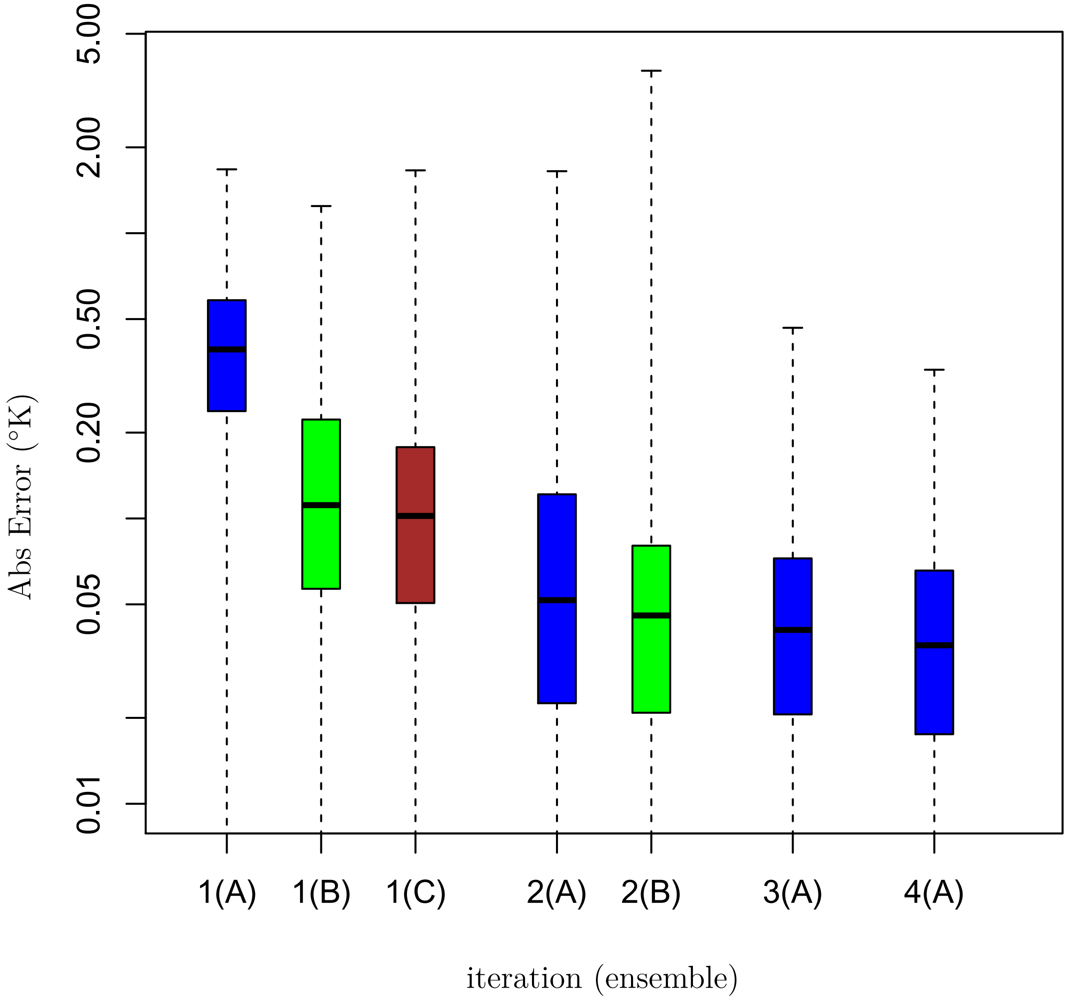


Figure 2.5: Spread of absolute errors (y-axis, log-scale) between actual model output and emulator predictions thereof for each iteration (x-axis, subdivided by ensemble) of the calibration to NCEP/NCAR data, are displayed using box-plots depicting quartile values. Ensembles A, B, and C are represented by the colours blue, green and brown.

a higher proportion of the training data in Ensemble A is focused on the high probability regions of the parameter space, and so the emulators are better able to provide information about these regions. However, the much larger estimated observational uncertainties for the reanalysis data set than for those constructed for use with the perfect model experiment, as well as the presence of model discrepancy, do result in a wider range of potential parameter sets in general. This behaviour was observed in tests (not shown) where additional noise was added to the calibration data for the perfect model experiment (described above), and then modelled using the additional model discrepancy terms. The larger the model discrepancy becomes, the wider the range of "reasonable" parameter sets becomes. Still, it is notable that many of the marginal posterior distributions do not appear (Figure 2.7) to have evolved significantly from the prior distribution. Investigations of the two and three dimensional marginal distributions (not shown) give little additional information beyond the association between θ_1 and θ_4 discussed above. To investigate how well the marginal parameter densities describe the full posterior, an additional ensemble was run using parameter values sampled independently from the individual marginal distributions. These model runs were frequently lower in likelihood than runs with parameter sets sampled from the multivariate posterior, with less than a quarter of the runs produced using the marginal posteriors having likelihood values within the range of the top fifty percent of values produced from the final iteration of Ensemble A. This suggests that the marginalization masks a more focused multivariate distribution.

Comparing the global expected model output from the last sub ensemble of Ensemble A[17] to the entirety of the reanalysis data (Table 2.2) shows that the calibration has reduced global biases and errors for all fields. For the sea level pressure field, the

[17] Note that this sub ensemble is not actually produced by the posterior distribution addressed in Figure 2.7 which would require generating a new set of runs with the GCM. Given the limited evolution between the final iterations of the calibration routine for Ensemble A, we assume that this sample is an adequate approximation to that which would be generated by the final posterior.

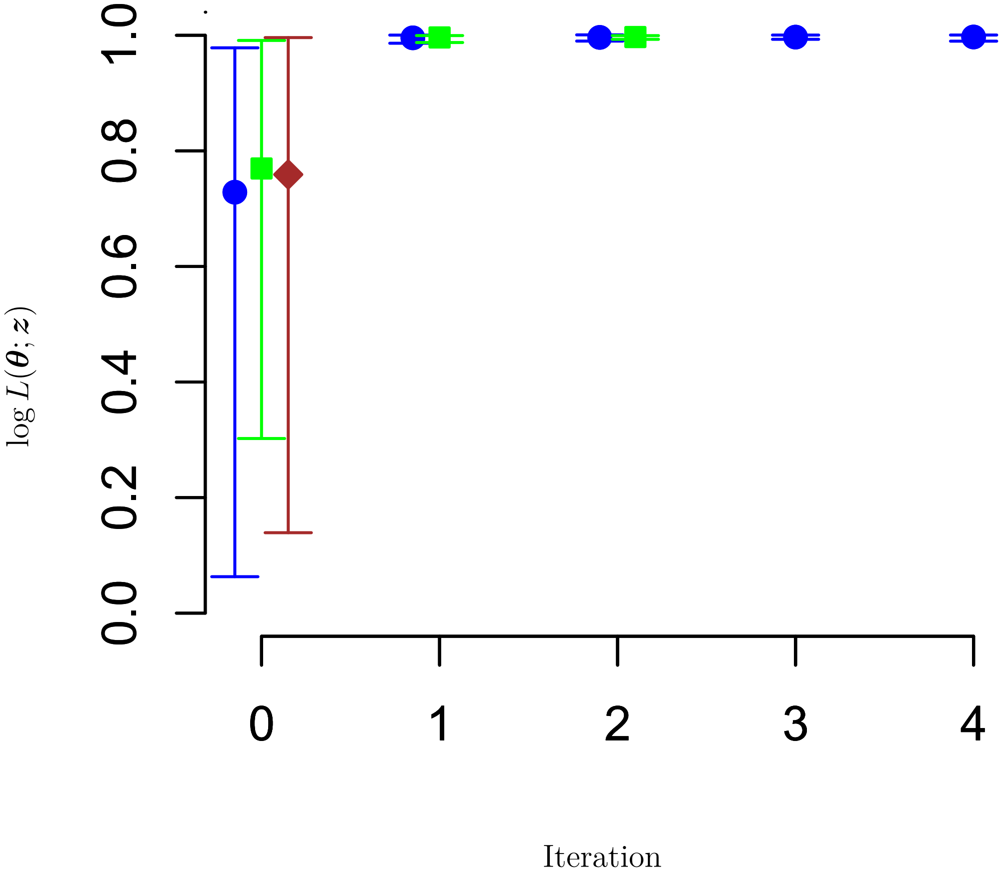


Figure 2.6: The mean of the calculated log-likelihood values (y-axis, scaled to a range of [0 : 1]) of the model runs, bracketed by their 10% and 90% quantiles, produced at each iteration of the calibration routine (specified on the x-axis, points are offset for clarity) for calibration to the NCAR/NCEP data set. Ensembles A, B, and C are represented by the colours blue (circle), green (square), and brown (diamond), respectively

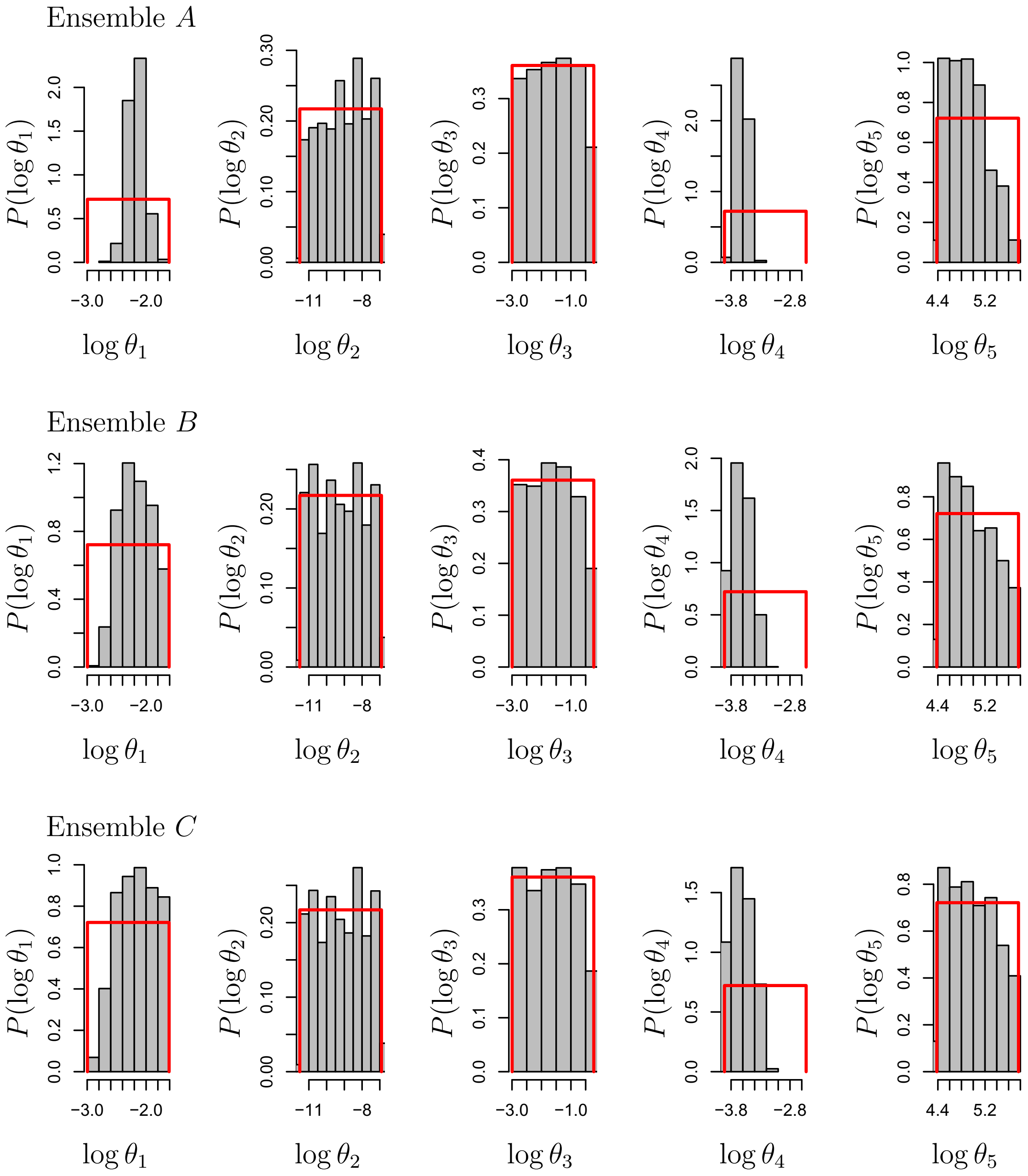
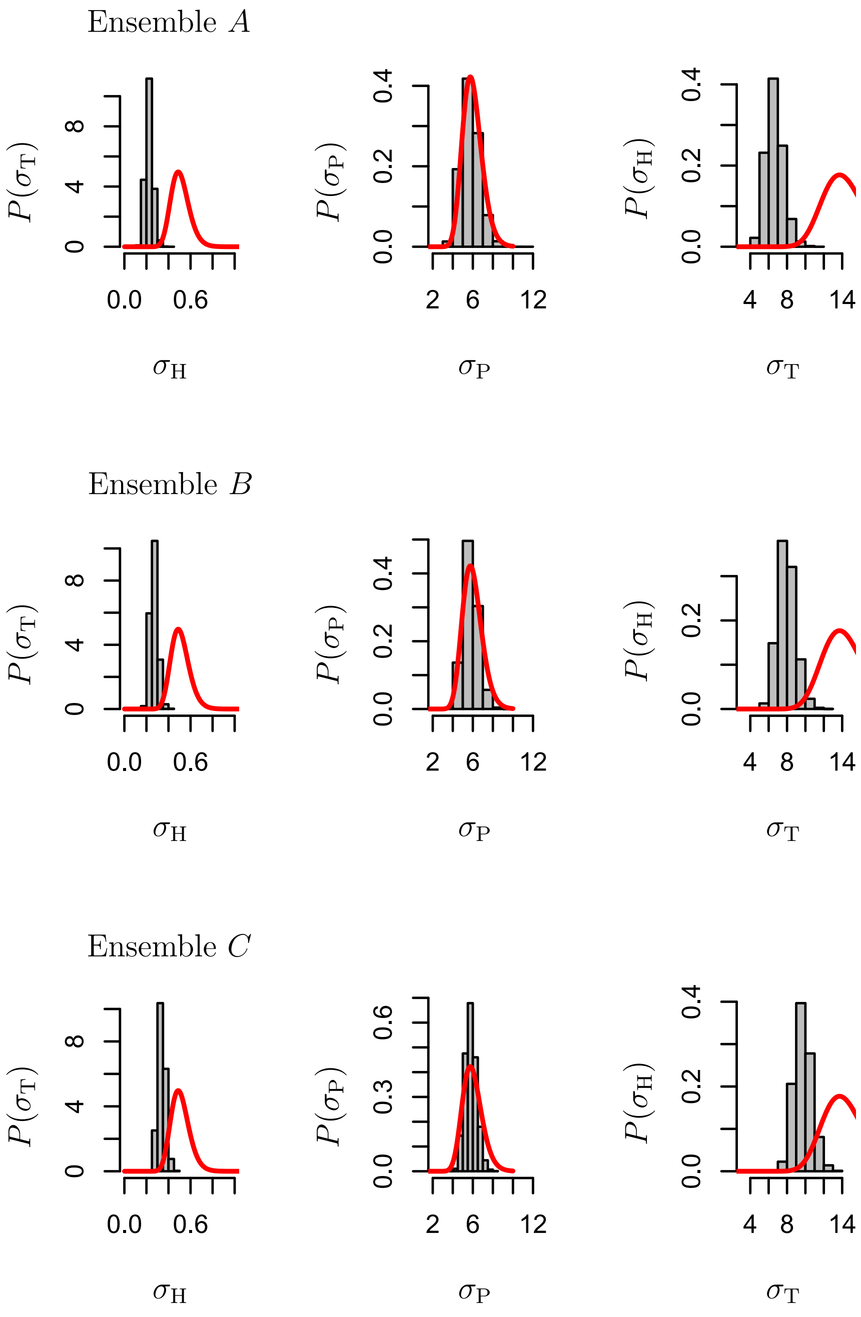


Figure 2.7: Histograms show marginal posterior densities for model parameters as estimated by the BANNs when calibrating to the NCEP data. Red lines show the prior distributions for the parameters. Top row is the result of performing five iterations of the calibration routine, using twenty model runs apiece, middle is result of performing three iterations using thirty model runs apiece, and bottom is result of performing two iterations using fifty runs apiece.


Ensemble A

Ensemble B

Ensemble C

Figure 2.8: Histograms show marginal posterior densities for error model parameters as estimated by the BANNs when calibrating to the NCEP data. Red lines show the prior distributions for the parameters. Top row is for Ensemble A, middle Ensemble B, and bottom for Ensemble C

model run with the maximum calculated (over all targets) likelihood performs worse than the model run with the lowest calculated likelihood, suggesting that these targets are not dominant contributors to the total likelihood calculation. This is appropriate considering that the relative observational uncertainties for the sea level pressure targets are an order of magnitude higher than the targets for the other fields, and so the calibration results are more focused on matching these latter targets. The ensemble produced from the marginal distributions (discussed above) has a similar Root Mean Square Error (RMSE) with respect to NCEP/NCAR fields though with a bit more bias and wider standard deviation (Table 2.2). The latter again suggests that the final posterior has a structure that is not fully captured by Figure 2.7. For larger parameter sets, more complex models, and non-diagonal error models, the marginal distributions will likely be more unreliable.

As an example of the spatial distribution of the fields used to calculate the statistics of Table 2.2, the difference maps for mean annual surface temperature between model output and the NCEP/NCAR field are shown for the original model settings and the ensemble mean (Figure 2.9). Similar patterns are seen for all other fields, although misfit levels over the poles have a strong seasonal component not shown here. While certain areas see increases in misfit, the overall error is reduced and is more evenly spread across the globe. This suggests that the regions averaged over to produce the calibration targets were well selected to address model performance in a variety of regions. This may however not be the most desirable result in practice. Considering that the model in question is an EMIC without realistic ocean circulation, producing reasonable approximations of equatorial and mid latitude phenomena and allowing a polar cold bias may be a more “physically realistic” calibration goal. Here we observe one of the dangers of calibration and parametrization in general; compensating for model shortcomings through potentially unrelated parameters. Due to the lack of

Table 2.2: Comparison of the total model output for each field against the corresponding NCEP/NCAR reanalysis map, for a default model run at the original parameter settings, and for the mean field calculated from the samples composing the final iteration of Ensemble *A*. Runs from this subset with the highest and lowest calculated likelihood are also included, as well as temperature results from a model ensemble created using the marginal densities for individual parameters as estimated from Ensemble *A*. Results are summarized by the mean difference (model output - reanalysis), and by the root mean square difference between the fields. The standard deviation of the model ensembles about their mean fields is also presented.

	mean difference	RMSE	ensemble-1σ
specific humidity [kg/kg] DJF			
default model settings	-0.0028	0.0029	
ensemble mean	0.000035	0.0016	0.00038
max-likelihood run	-0.00018	0.0016	
min-likelihood run	0.00061	0.0017	
specific humidity [kg/kg] JJA			
default model settings	-0.0031	0.0033	
ensemble mean	-0.00013	0.0019	0.00041
max-likelihood run	-0.00034	0.0019	
min-likelihood run	0.00046	0.0020	
sea level pressure [hPa] DJF			
default model settings	1.24	4.33	
ensemble mean	0.82	4.00	0.64
max-likelihood run	0.85	4.21	
min-likelihood run	0.74	3.86	
sea level pressure [hPa] JJA			
default model settings	0.27	3.98	
ensemble mean	-0.091	3.70	0.43
max-likelihood run	-0.063	3.80	
min-likelihood run	-0.16	3.64	
surface temperature [°K] DJF			
default model settings	-5.10	6.08	
ensemble mean	0.83	4.35	0.73
marginal ensemble mean	1.08	4.37	1.10
max-likelihood run	0.41	4.301	
min-likelihood run	1.95	4.68	
surface temperature [°K] JJA			
default model settings	-4.80	6.14	
ensemble mean	1.77	4.04	0.77
marginal ensemble mean	2.00	4.09	1.10
max-likelihood run	1.34	3.95	
min-likelihood run	2.91	4.48	

heat transport, the polar cold bias is corrected by increasing the global available energy, which results in overheating in equatorial and other regions. This issue could be addressed by changing the calibration targets; e.g., reducing the weighting for data from polar or other regions. However, as this is an issue of model discrepancy a preferable solution would be to create a more sophisticated error model. Allowing different error (and potentially bias) terms for targets relating to different latitudes would provide the opportunity to use the Bayesian inference to determine where (spatially) the model can perform best. Including such terms in the approximation of the likelihood covariance matrix would focus the calibration to physically realistic solutions where possible, and quantify the degree of error where they are not. Provided that these terms can be successfully estimated in the posterior this may produce a more desirable result than our current “compromise solution” which results from fitting the model to our simplistic assumptions about its information content.

Table 2.2 shows that the model never describes the observational data within the ensemble standard deviation. However, the ranges of the estimated discrepancy terms (Figure 2.8) completely capture this error. These estimates are, as predicted, overly conservative. This results from the oversimplification of the likelihood covariance matrix as discussed above. Testing the effect of adding additional noise to the calibration data for the perfect model experiment (described above), suggests that when the error and its model are of the same statistical form, the resulting estimates typically lie within the one-sigma range of the “true” synthetic error. However, the quality of this description decreases as the synthetic error is located further into the tails of the prior distribution, showing that the error model can be sensitive to the utilised priors. Figure 2.8 suggests that use of a log-normal prior for this experiment is perhaps restrictive for the case of σ_H and σ_T. While it describes our initial information, the relatively narrow tails of the log-normal distribution represent an assumption that it

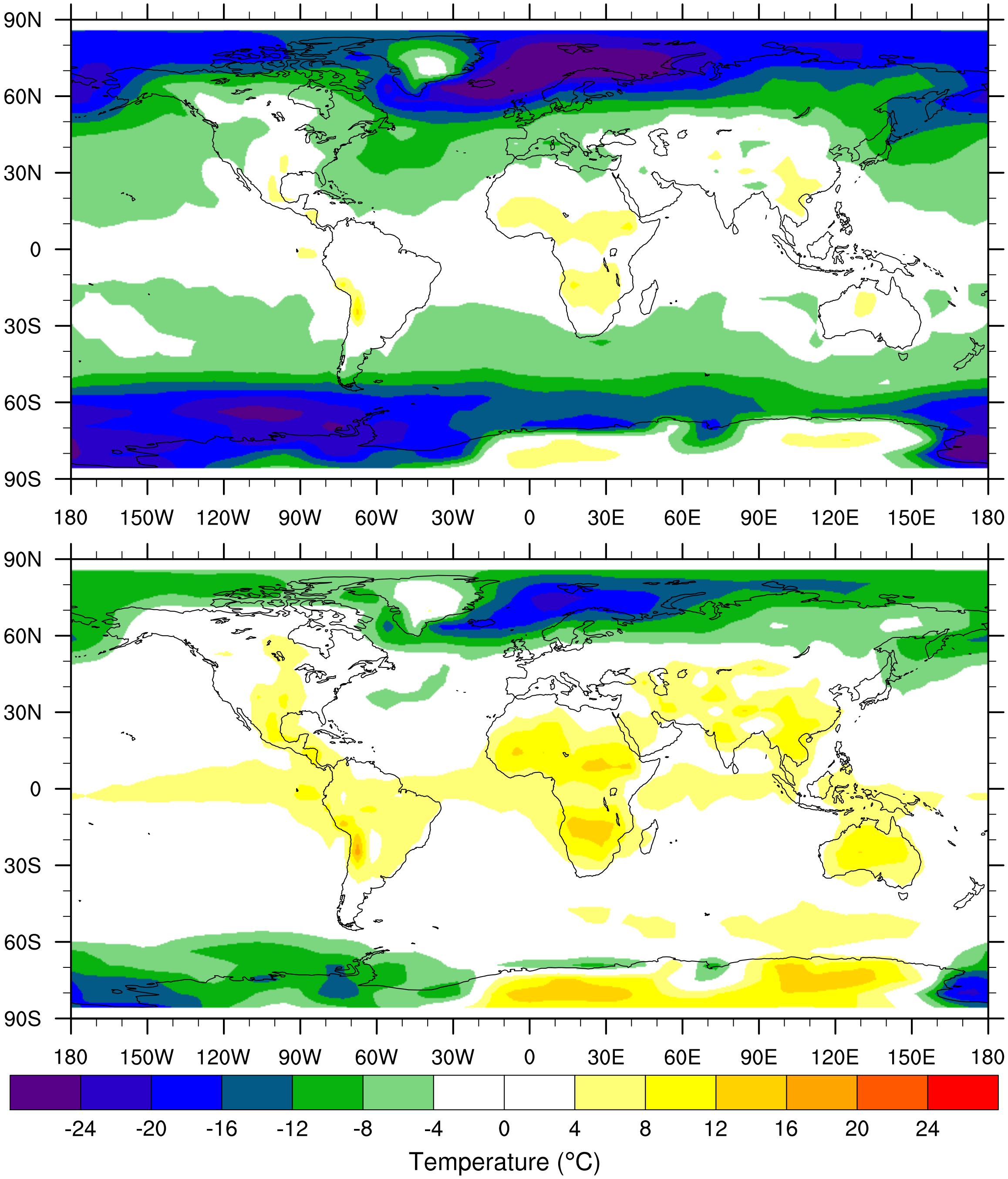


Figure 2.9: Difference in surface temperature between the model run at its original settings and the reanalysis data (top), and between the the ensemble mean produced by the final iteration of Ensemble *A* for the NCEP calibration and reanalysis data (bottom)

is very unlikely to find within the parameter space a model realization that performs significantly better or worse than what we have seen in our limited initial samples. In general, whether such an assumption is justified will be very specific to the amount (or lack thereof) of initial information available concerning the model in question. In contrast the marginal density of the discrepancy for sea level pressure (σ_P) is only slightly shifted from its prior. This is in agreement with the small change in this field produced by the calibration discussed above.

As an example of the use of this result for probabilistic forecasting we create two ensembles of twenty model runs each, using the posterior produced by Ensemble A. These ensembles are run into the future, one using an approximation of the A2 climate forcing scenario as described by Nakicenovic et al. (2000), the other having CO_2 stabilised at 2008 levels. These ensembles are used to compute Eq. (2.10) for a range of potential mean global temperatures for the decade of 2048-2059, the results of which are shown in Figure 2.10. It is important to note that given the simplistic nature of the current experiment this figure is not meant to serve as an actual prediction, but simply as an illustration of the potential of the method. Given the degree of convergence of Ensemble A we could alternately have minimized computational time by running the final sub-ensemble forward to the desired future date.

2.6 Conclusions

As part of the construction of a Bayesian posterior distribution, we have documented the information, assumptions and inference used to constrain model parameter selection. The BANNs, despite a limited supply of training data, emulate the model behavior to a sufficiently high degree to allow identification of high-probability parameter sets which improve model fit to observational data. The resultant computational

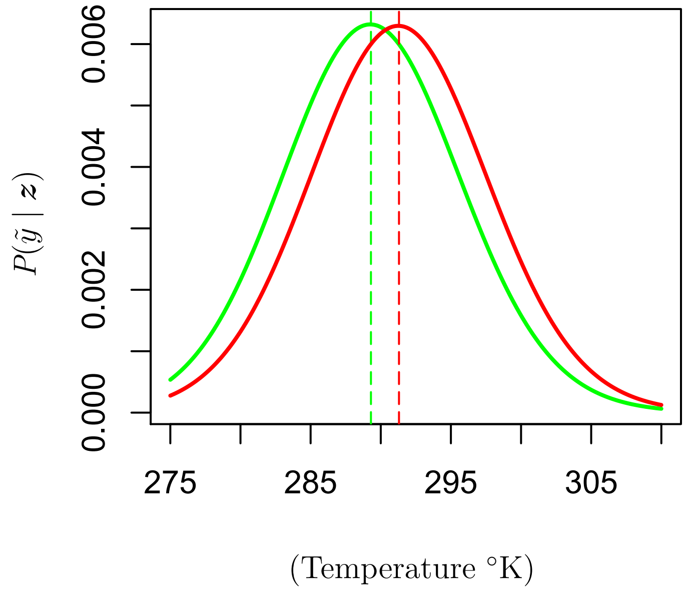


Figure 2.10: Probabilities for different mean global temperatures for the decade of 2048 - 2059, for increasing (red, right-most) and stabilized (green, left-most) CO_2 forcing, taking into account estimates for observational, parametric, and model uncertainties. Dashed lines show temperatures that correspond to peak probability

feasibility of MCMC methods enables the sampling of terms describing model error while exploring the posterior space. This in turn permits estimation of the inherent uncertainties over the resulting space of calibrated model realizations, and avoids the false assumption that parametric uncertainty can capture the entirety of the model-reality discrepancy. Therefore model ensembles can be used to construct rigorous probabilistic forecasts.

Our results show that the error model will need to be more complex in practice than the one we have experimented with here. In particular it will need to be more responsive to differences in the model's ability to resolve distinct targets. Also, it will be necessary to estimate some portion of the targets' covariance structure over the model space, particularly temporally, so as to properly address the information content of individual targets as regards overall model performance. Further work must be done to assess what practical limitations exist in how complex these error models can become. Also, implementing more detailed descriptions of model error will require more exact representations of emulator error.

The "smoothing" of errors over the model domain (Figure 2.9) suggests that the calibration targets are well distributed. However, it is not clear whether the degree of reduction in information was appropriate for our calibration goals. As there exist a wide variety of data classification methods, criteria for the selection and pre-processing of calibration targets for particular models and applications need further development. As the model discrepancy is estimated through the MCMC sampling of the posterior, our assessment and interpretation of these terms is inherently linked to the choice of calibration target. The more we can elucidate the potential covariance structure between modelled calibration targets, the more we simplify the task of accurately assessing the model discrepancy.

In general, problems of emulator design, approximation of the likelihood covariance matrix, selection of calibration targets and specifying prior information, as well as discerning the appropriate use and interpretation of ensembles produced in accordance to the resulting posterior, will all have to be tailored to the available model and data, with solutions that will vary from situation to situation. However, in these initial tests the methodology has shown potential for the objective and tractable Bayesian calibration of computationally expensive Earth system models. Furthermore, this is to a degree of completeness such that the generated information and related uncertainties can be directly used to make statistically rigorous inferences about the physical system(s) being investigated.

2.7 Bibliography

Annan, J. D., Hargreaves, J. C., 2007. Efficient estimation and ensemble generation in climate modelling. Philos. Trans. Roy. Soc. London 365, 2077–2088.

Craig, P., Goldstein, M., Rougier, J., Seheult, A., 2001. Bayesian forecasting for complex systems using computer simulators. J. Amer. Stat. Assn. 96 (454), 717–729.

Edwards, N., Cameron, D., Rougier, J., 2010. Precalibrating an intermediate complexity climate model. Clim. Dyn., 1–14.

Goldstein, M., Rougier, J., 2010. Reified Bayesian modelling and inference for physical systems. J. Stat. Plan. Infer. 139.

Grosso, A., Jamali, A., Locatelli, M., 2008. Iterated local search approaches to maximin Latin hypercube designs. Innov. Advan. Tech. Sys., Comp. Sci. Softw. Eng.

Holden, P., Edwards, N., Oliver, K., Lenton, T., R., W., 2010. A probabilistic calibration of climate sensitivity and terrestrial carbon change in genie-1. Cli Dyn 35 (5).

Jackson, C., 2009. Use of Bayesian inference and data to improve simulations of multi-physics climate phenomena. J. Phys. 180.

Jackson, C., Sen, M., Huerta, G. Deng, Y., Bowman, K., 2008. Error reduction and convergence in climate perdiction. J. Climate 21.

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

Jaynes, E. T., 2003. Probability Theory; The Logic of Science, 1st Edition. University Press, Cambridge.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., M., C., Ebisuzaki, W., Higgins, W., Janowiak, J. Mo, K., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R. Joseph, D., 1996. The NCEP/NCAR 40-year reanalysis project. Bull. Amer. Meteor. Soc. 77.

Khu, S. T., Micha, G., 2003. Reduction of Monte-Carlo simulation runs for uncertainty estimation in hydrological modelling. Hydro. Earth Sys. Sci. 7 (5).

Knutti, R., Stocker, T., Plattner, G., 2003. Probabilistic climate change projections using neural networks. Climate Dyn. 21.

Lee, H., 2004. Bayesian nonparametrics via Neural Networks. ASA and SIAM.

Lunkeit, F., Böttinger, M., Fredrich, K., Jansen, H., Kirk, E., Kleidon, A., Luksch, U., 2007a. Planet Simulator Reference Manual. University of Hamburg, 15th Edition.

Lunkeit, F. Blessing, S., Fraerich, K., Jansen, H., Kirk, E., Luksch, U., Sielmann, F., 2007b. Planet Simulator User's Guide. University of Hamburg, 15th Edition.

MacKay, D., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Mosegaard, K., Sambridge, M., 2002. Monte Carlo analysis of inverse problems. Inv. Prob. 18.

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

Murphy, J., Booth, B., Collins, M., Harris, G., Sexton, D., Webb, M., 2007. A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. Philos. Trans. Roy. Soc. London 365, 1993–2028.

Nakicenovic, N., Alcamo, J., Davis, G., de Vries, B., Fenhann, J., Gaffin, S., Gregory, K., Grubler, A., Jung, T., Kram, T., La Rovere, E., Michaelis, L., Mori, S., Morita, T., Pepper, W., Pitcher, H., Price, L., Riahi, K., Roehrl, A., Rogner, H., Sankovski, A., Schlesinger, M., Shukla, P., Smith, S., Swart, R., van Rooijen, S., Victor, N., Z., D., 2000. IPCC Special Report on Emissions Scenarios. Cambridge University Press, Cambridge, United Kingdom and New York.

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

Neal, R., 2003. Slice sampling. Ann. Stat. 31 (3), 705–767.

Preisendorfer, R., 1988. Principal Component Analysis in Meteorology and Oceanography. Elsevier.

Rougier, J., 2007. Probabilistic inference for future climate using an ensemble of climate model evaluations. Climate Change 81, 247–264.

Rougier, J., 2008. Efficient emulators for multivariate deterministic functions. J. Comp. Graph. Stat. 17 (4), 827–843.

Rougier, J., Guillas, S., Maute, A., Richmund, A., 2007. Emulating the thermosphere-ionosphere electrodynamics general circulation model. Tech. rep., Statistical and Applied Mathematical Sciences Instituted, Research Triangle Park, NC, USA.

Sambridge, M., Mosegaard, K., 2002. Monte carlo methods in geophysical inverse problems. Rev. Geophys. 40 (3), 1–29.

Sanderson, B., Piani, C., Ingram, W., Stone, D., Allen, M., 2008. Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics GCM simulations. Climate Dyn. 30, 175–190.

Sivia, D., Skilling, J., 2006. Data Analysis A Bayesian Tutorial, Second Edition. Oxford University Press Inc.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., Tignor, M., Miller, H., 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., Peltier, W., 2005. Arctic freshwater forcing of the Younger Dryas cold reversal. Nature 435, 662–665.

Urban, N., Fricker, T., 2010. A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of a climate model. Comp. Geos. 36, 746.

Villagraon, A., Huerta, G., Jackson, C., Sen, M., 2008. Computational methods for parameter estimation in climate models. Bayes. Analy. 3 (4), 823–850.

Wagener, T., Boyle, D., Less, M., Wheater, H., Gupta, H., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydro. Earth-Syst. Sci. 5 (1), 13–26.

Connecting Text

The following article was written as a review of atmosphere and ocean interannual variability, over the past fifty years, for the North Atlantic region. The evolving understanding of interactions between atmospheric and ocean processes on multiple scales is outlined. These are related to dominant modes of atmospheric variability, which are described in different degrees of regional and temporal detail; addressing objectives (3)-(5). As such, this article offers a perspective on issues inherent in defining statistical climate features for calibration targets,; e.g., identifying features, determining relationships between processes, and formulating targets such that their uncertainties can be quantified. This article has been submitted to the journal *Progress in Oceanography.*

Chapter 3

North Atlantic atmospheric and ocean interannual variability over the past fifty years - spatial patterns and decadal shifts

3.1 Abstract

This article presents results from a study of the patterns of interannual variability of the North Atlantic atmospheric circulation over the past fifty years, and their links with the observed subpolar ocean variability. A fuzzy clustering analysis is used to identify the patterns of atmospheric variability in the interranual spectral interval. Four dominant patterns of North Atlantic interannual variability are found, which describe phases of two asymmetrical alternating modes. The first two patterns have the spatial structures of positive and negative phases of the North Atlantic Oscillation. The third and fourth patterns define the opposite phases, here referred to as G+ and G-, of an alternating mode that closely resembles the regional manifestation of the Pacific-North American Pattern (PNA).

Alternatively, the patterns of interranual variability are characterised through the associated distributions of subseasonal weather regimes. The latter are defined from Sea Level Pressure (SLP) anomalies using Bayesian Gaussian mixture models. In the 1960s the distribution of weather regimes favoured blocking patterns over the North Atlantic and warmer than normal upper ocean temperatures. In the late 1980s and early 1990s the dominant weather regimes favoured intensification of the Icelandic low and cold winters over the Subpolar North Atlantic. The change of the distribution of the weather regimes between 1960s and 1990s is associated with a decadal shift in the dominant interannual patterns from NAO- and G+ in the 1960s towards NAO+ and G- in the late 1980s and early 1990s. While there are strong indications that the recent warming of the Subpolar North Atlantic since the mid-1990s was triggered by internal ocean dynamics and Atlantic Multidecadal Oscillation, it is suggested that the atmospheric variability related to the domination of the G+ pattern in the past 20 years was a factor that additionally contributed to this process.

3.2 Introduction

The global oceans have warmed since the mid 20th century (Levitus et al., 2009). This trend is well correlated with contemporaneous greenhouse gas driven variations in radiative forcing (Solomon et al., 2007). On regional scales interannual and decadal variations are often superimposed on the long term climate trends. One example is the observed decadal cooling of the Subpolar North Atlantic Ocean and warming of the continents in the 1980s and early 1990s (Wallace et al., 1996; Corti et al., 1999).

Studies of the impact of atmospheric interannual variability on the North Atlantic climate date back to the works of Sir Gilbert Walker (Walker, 1922, 1924). Walker and Bliss (1932) demonstrated that the intensity of the Icelandic surface pressure minimum is well correlated with the cold winters in Eastern North America, Greenland and the Middle East, and with warm temperatures in and Northwest Europe. The reverse pressure anomaly over Iceland is related to opposite tendencies in these regions. The anomalous low pressure over Iceland is dynamically consistent with stronger than normal cold advection over the Labrador and Greenland, and with intense southwestern flow of mild ocean air over the Northwestern Europe. More recent studies (for a review see Hurrell and Deser (2009), Greatbatch (2000), and Hurrell et al. (2003)) have demonstrated that this alternating mode, identified by Walker (1922, 1924) as the North Atlantic Oscillation (NAO), dominates the variability of the North Atlantic.

The origin of the NAO has been related to processes of baroclinic instability, and with eddy generation and decay (Thompson et al., 2003), with typical time scales of about ten days (Simmons and Hoskins, 1978). The influence of these processes on the ocean is usually presented as a background noise providing energy to internal ocean dynamics, which have higher inertia and longer time scales of variability (Frankignoul and Hasselmann, 1977). The ocean, which has a higher heat capacity, damps the atmospheric variability leading to a red spectrum in climate variability (Vallis, 2009). This paradigm of climate is often called null hypothesis of climate variability (Greatbatch, 2000). Alternate descriptions have been presented and are reviewed in this text. These put more emphasis on isolated ocean processes for decadal scale variations. Shorter term variation is driven by responses to specific regimes of atmospheric forcing. This approach requires more detailed descriptions of atmospheric variability. It has been suggested that, while informative, typical measures of NAO activity obscure more subtle, yet significant features (Monahan et al., 2001; Cassou et al., 2004; Reusch et al., 2007).

The mechanism of the NAO on time scales longer than the typical life time of atmospheric weather events and disturbances is less understood (Vallis, 2009). Peaks in the observed atmospheric power spectrum on interannual and decadal time scales suggest the presence of regime like behavior. While atmospheric regimes with robust statistical and dynamical foundations have been identified (Molteni et al., 2006), their predictability and the predictability of the atmospheric interannual and decadal variability on periods longer than a year is not significant (Vallis, 2009).

This article presents results from a study of the patterns of interannual North Atlantic atmospheric variability. It is based on the paradigm of the North Atlantic atmosphere as a dynamical system, which exhibits chaotic variability on many time scales with complex feed-backs between its components (Molteni et al., 2006; Lorenz, 2006; Monahan et al., 2010). More specifically, the following discussion addresses possible mechanisms of coupled atmosphere-ocean variability of the Subpolar North Atlantic (see Figure 3.1).

In the following we first give a brief overview of atmospheric and oceanic interannual variability, as observed over the North Atlantic for the past fifty years, focused on the NAO. Then we discuss the paradigm of the atmosphere as a dynamical system, and related methods for describing atmospheric variability. We present results from

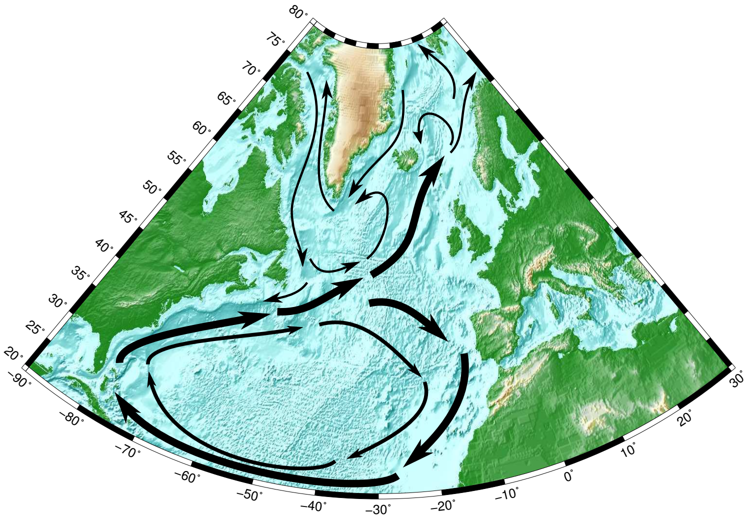


Figure 3.1: North Atlantic Region, with major ocean circulation patterns outlined.

the analysis of interannual variations in the North Atlantic atmospheric circulation. Finally, associations with interannual and decadal ocean variability are presented followed by general discussion.

3.3 North Atlantic atmospheric and oceanic variability from years to decades

The long-term mean surface atmospheric pressure distribution over the North Atlantic region is dominated by the Azores high and Icelandic low pressure centres (Hurrell and Deser, 2009). The Azores high pressure system is stronger over the summer season when it covers a large area of the North Atlantic. In winter the Icelandic minimum dominates and the Azores high weakens and moves equatorward. The zonal westerly flow driven by the mean pressure gradients dominates the regional circulation at mid-latitudes throughout the year. The westerlies extend through the troposphere and have a maximum at a height of about 12km (Hurrell and Deser, 2009). This mid-latitude jet stream coincides approximately with the storm tracks between North

America and Europe, and its variability influences the climate of the North Atlantic region (Willet and Sanders, 1952; Hoskins and Hodges, 2010).

The NAO is typically defined by variations in the strength of the Azores and Icelandic pressure centres (Hurrell et al., 2003). The NAO index; cf., see Hurrell (1995), is often calculated from the difference between normalised Sea Level Pressure (SLP) anomalies in Iceland and the Azores. Alternatively, the NAO index can be defined as the dominant Principal Component (PC) of the SLP (Hurrell et al., 2003).

The NAO index is indicative of the position of the jet stream and extra-tropical cyclone tracks across the North Atlantic; i.e., whether these features tend northward or southward of their climatological positions (Vallis et al., 2004). A positive NAO index is associated with warmer wetter weather in northern Europe, and cold dry weather in northern North America, while a negative index is associated with an opposite variability (Hurrell et al., 2003). Current investigations; e.g., Luo et al. (2011), suggest that there is a relationship between long term NAO phase and storm intensity. Links to local-scale phenomena beyond the North Atlantic have also been made; e.g., Feliks et al. (2010). While the NAO signature is typically strongest in the Northern Hemisphere winter months the relevant processes exist in weaker forms throughout the year (Feldstein, 2007).

The mechanism of the NAO has been a centre of debate over the role of global atmospheric process for the North Hemisphere (Thompson et al., 2003), North Atlantic regional dynamics (Deser, 2000), and Sea Surface Temperature (SST) (Rodwell, 2003; Kushnir et al., 2002). Numerical simulations have demonstrated that the NAO does not owe its existence to the ocean-atmosphere interaction but is a result of intrinsic atmospheric dynamics (Hurrell et al., 2003), and on monthly to yearly time scales the SST has weak impact on NAO (Kushnir et al., 2002). Thompson and Wallace (2000) and Thompson et al. (2003) found that the NAO is a component of Northern

Hemisphere large scale atmospheric circulation variability. More specifically, Thompson and Wallace (2000) demonstrated that the NAO is a regional manifestation of the dominant mode of variability of the zonal mid-latitude flow, known as the North Hemisphere Annular Mode (NAM), or also as the Arctic Oscillation. The NAM variations are related to pressure anomalies with opposite signs along 55°N and 35°N. The NAM spatial pattern, which is defined as the leading PC of SLP anomalies over the Northern Hemisphere, is zonally symmetric over most of the hemisphere, but intensifies locally over the North Atlantic, where it resembles the NAO spatial pattern. The NAM has a strong signature in the pressure fields of the middle and upper troposphere. In the stratosphere it has a zonally symmetric structure typical for the atmospheric annular modes of the Northern and Southern hemispheres (see Thompson et al. (2003)).

Figure 3.2 shows the NAM pattern calculated as dominant PC of SLP for the Northern Hemisphere (Figure 3.2a) and the correlation pattern for the time series of the first PC of the North Atlantic region only (Figure 3.2b). The two patterns are very close over the North Atlantic. The time series for the NAO index calculated from Lisbon, Portugal and Stykkisholmur/Reykjavik, Iceland station data (Figure 3.3a) and as the first PC of the North Atlantic SLP (Figure 3.3b) show very close temporal variability with the NAM index calculated as the dominant PC of the Northern Hemisphere SLP (Figure 3.3c). The close similarities in the spatial patterns and temporal variability of the NAO and NAM reflect the fact the NAO is the regional projection of the NAM (Thompson and Wallace, 2000; Thompson et al., 2003). In the following we will refer the dominant mode of North Atlantic atmospheric variability as the NAM/NAO.

The time variations of the NAM/NAO reflect the variability in the position and strength of the mid-latitude zonal jet stream over the North Atlantic, with respect to its climatological position (Thompson et al., 2003). This is an eddy-driven jet, which

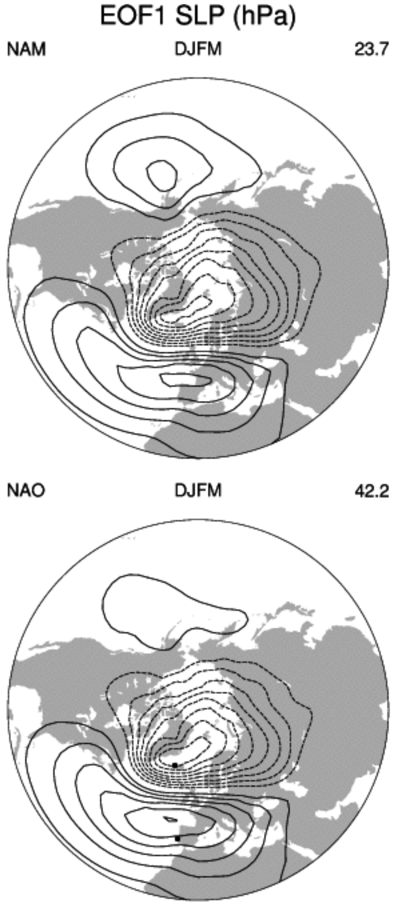


Figure 3.2: Hurrell's North Atlantic Oscillation and North Annular Mode patterns: **(a) Top panel:** The first PC of Northern Hemisphere (20^o-90^oN, 90^oW-40^oE) winter SLP data. It explains 23% of the extended winter mean (December-March) variance. **(b) Bottom panel:** The correlation pattern for the time series of the leading PC of SLP anomalies over the North Atlantic. Contour interval is 0.5 hPa for both images. These figures are from the UCAR web site: climatedataguide.ucar.edu/climate-data/hurrell-wintertime-slp-based-northern-annular-mode-nam-index

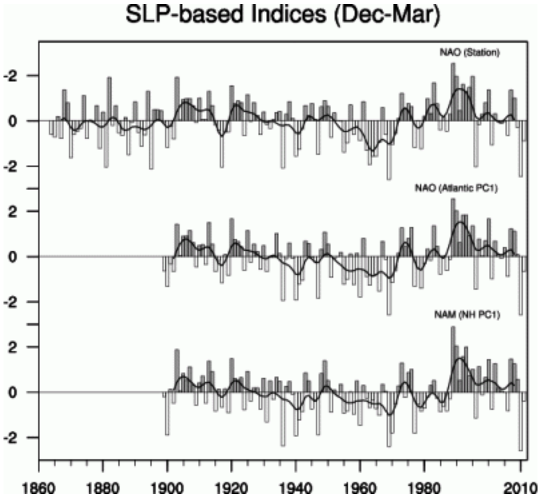


Figure 3.3: Hurrell's North Atlantic Oscillation and Northern Annular Mode indices. **(a) Top panel**: The winter (December through March) NAO station-based index based on the difference of normalized sea level pressure (SLP) between Lisbon, Portugal and Stykkisholmur/Reykjavik, Iceland since 1864. **(b) Middle panel**: The NAO PC-based index of SLP anomalies over the Atlantic sector, $20^o - 80^o$N, 90^oW-40^oE. **(c) Bottom panel**: The NAM PC-based index of SLP anomalies over the Northern Hemisphere $20^o - 90^o$N. The figures are from the UCAR web site: climatedataguide.ucar.edu/category/data-set-variables/climate-indices/

is triggered by eddy fluxes of momentum between the mean flow and transient eddies. The latter are usually defined as departures from the time mean zonal flow. Panetta and Held (1988) showed that the eddy-mean flow interaction can maintain jets in the absence of thermal forcing. When a perturbation grows and decays it drives back the mean jet through the convergence of westerly eddy momentum flux. The eddy driven jets are usually weaker, and at the same time exhibit much stronger spatial and temporal variability than their thermally driven sub-tropical counterparts (Thompson et al., 2003). The variability related with their dynamics is particularly strong when the meridional extension of the baroclinic zone exceeds the size of the eddies, which permits meanders to develop in the jet (Lee and Feldstein, 1996). The presence of warm ocean surfaces in the subpolar North Atlantic permit atmospheric eddy activity over a larger latitude sector, which supports particularly strong baroclinic instability and eddy-mean flow interaction in the region. This is one possible explanation of the particularly strong NAM variability in the North Atlantic sector (Thompson et al., 2003).

NAM/NAO exhibits temporal variability on all time scales from sub-seasonal, to interannual and decadal (Hurrell and Deser, 2009). The physical mechanism of NAM/NAO variability is related to the processes of atmospheric baroclinic instability, Rossby wave propagation and breaking, and eddy generation and decay. The typical time scales of these processes varies from seven to sixty days (Thompson et al., 2003). Longer NAM/NAO variability may be triggered through interactions of the zonal jet with the quasi-biennial oscillation in the equatorial troposphere which has an estimated period of 27 months (M. P. Baldwin et al., 2001). Less is known about the mechanism of NAO/NAM related atmospheric variability on longer time scales. The observed atmospheric spectrum exhibits peaks on interannual and decadal time scales, suggesting the existence of atmospheric regimes on these scales (Vallis, 2009).

Trends for the NAM/NAO to favor certain phases can persist for time periods longer than years and decades (Hurrell and Van Loon, 1997). Historical data indicates that the NAO was positive for much of the first three decades of the 20th century (see Figure 3.3). From the 1930s to early 1970s the NAO index exhibited a downward trend (see Figure 3.3). Since the 1980s the NAO has been in a predominantly positive phase, with the highest values of NAO index ever observed occurring in the late 1980s and early 1990s (Hurrell and Van Loon, 1997). This NAO/NAM variability at interannual and decadal time scales is linked to variations in the surface atmospheric conditions over the North Atlantic Ocean in the past fifty years, which had a significant impact on the subpolar water mass characteristics and ocean circulation; e.g., Dickson et al. (1988); Dickson et al. (2000); Curry and McCartney (2001); Yashayaev and Clarke (2006);Yashayaev (2007); Lohmann et al. (2009); Lozier et al. (2010); Zhu and Demirov (2011).

During the 1960s when the NAO was predominately negative in phase, the winters over the Subpolar North Atlantic were mild and surface cooling weaker than average. The deep convection and ocean circulation were less intense than normal (Yashayaev, 2007; Zhu and Demirov, 2011). In the late 1960s a low salinity anomaly in the surface layer was observed to propagate around the Labrador Sea. This phenomena is known as the Great Salinity Anomaly (GSA) (Dickson et al., 1988). The GSA was related to an approximately 79×10^9 ton salt deficit, which was advected through the Labrador Sea, and in mid 1970s returned to the Nordic Seas (Dickson et al., 1988). It was equivalent to a 2×10^9 ton increase in the freshwater and sea-ice transport into the Subpolar North Atlantic (Aagaard and Carmack, 1989). Dickson et al. (1988) found that the excessive freshwater transport was triggered by intensified sea-ice and freshwater export from Arctic through the Fram Strait and Nordic Seas, which in the 1960s entered the Subpolar North Atlantic.

It is well established that the export of Arctic sea-ice through the Fram Strait is positively correlated with the NAO index (Kwok and Rothrock, 1999). In the 1960s the low frequency NAO index was predominantly negative in phase, while at the same time the observations suggest that this was the period when excessive freshwater and sea-ice transport through the Fram Strait triggered the GSA. This so-called the "GSA-paradox"; cf., Dickson et al. (2000), suggests that enhanced sea-ice transport through the Fram Strait is possible during both states of the NAO. In particular, Dickson et al. (2000) demonstrated that there were time periods during the 1960s when the atmospheric circulation patterns which favour stronger than normal sea-ice export dominated the atmosphere over the Northern Seas. It is also important to note, that while NAO trends can persist on decadal time scales, individual years need not conform to these patterns. In the 1950s and 1960s such short term changes may have impacted ocean transport through the Fram Strait, considering that the GSA was triggered by 25% higher than normal sea-ice and fresh water export through the Fram Strait, for a period of time of only about two years (Aagaard and Carmack, 1989).

Another mechanism that could have potentially contributed to the salt deficit in the Nordic Seas, during the development of the GSA in the 1960s, is related to variations in the export of salty and warm Atlantic Water to the Nordic Seas. The model simulations of Lundrigan and Demirov (2012) suggest that the export of Atlantic Waters into the Nordic Seas in the 1960s was anomalously low following the decay in the intensity of sub-polar gyre circulation. The latter was weaker than normal due to the weak surface forcing during the 1950s and 1960s, when the NAO index was predominately negative (Zhu and Demirov, 2011). Lundrigan and Demirov (2012) suggested that the weakening of the salty Atlantic Water inflow to the Nordic Sea additionally intensified the GSA, which had been triggered initially by the excessive

Arctic freshwater and sea-ice inflow.

Figures 3.4 and Figures 3.5 show the major trends in properties of the surface and intermediate water masses in the Labrador Sea in the past fifty years. During the GSA the surface layer salinity dropped by about 0.38psu. This is the largest variation in the surface layer salinity for the whole period which occurred within a time period of five years from 1967 to 1971. Following the decadal variability in the atmospheric forcing, the surface layer temperature was high in the 1950s and 1960s when the NMA/NAO was mostly negative in phase, and low in the 1980s and 1990s when NAO was predominately positive. These decadal trends are well pronounced also in the intermediate water mass properties. The surface layer, which has a smaller inertia than the intermediate and deep layers, shows an intense variation on time scales of two to three years, that are superimposed on the decadal trends.

The NAO index was in predominantly positive phase in the past three decades since the 1980s (see Figure 3.3), with exceptionally high values in the late 1980s and early 1990s when the NAO index reached its highest values since 1860 (Hurrell and Van Loon, 1997). The severe winter surface winds and cooling triggered intense deep convection (Yashayaev, 2007) and intensified the subpolar gyre circulation (Zhu and Demirov, 2011). Starting from the mid 1980s, the deep convection progressively developed to record depths. The estimated annual production of the Labrador Sea Water (LSW) between 1987 and 1994 was about 4.5 Sv with peaks in some years close to 7Sv (Yashayaev, 2007). The fresh and cold LSW which formed in this period (see Figure 3.5) was the deepest, densest and largest Labrador Sea Water (LSW) ever observed (Dickson et al., 2002).

The Subpolar North Atlantic has warmed substantially (see Figure 3.4) since the mid-1990s (Yashayaev and Clarke, 2006). This warming was initially stronger in the eastern North Atlantic (Marsh et al., 2008). In the western part of the basin, the warming was contemporaneous with a decrease in the intensity of the deep convection. Once winter convection had lost its strength after the winter of 1994/1995, the deep LSW 1987-1994 layer lost ’communication’ with the mixed layer above (see Figure 3.5), consequently losing its volume, while gaining heat and salt from the intermediate waters outside the Labrador Sea (Yashayaev, 2007). The surface 1000 m layer has been steadily becoming warmer and saltier since 1994/1995 (see Figure 3.4), although, there were two periods when cooling caused by an abrupt increase in the deep convection in the 1999/2000 and 2007/2008 (Yashayaev, 2007; Yashayaev and Loder, 2009; Vage et al., 2008).

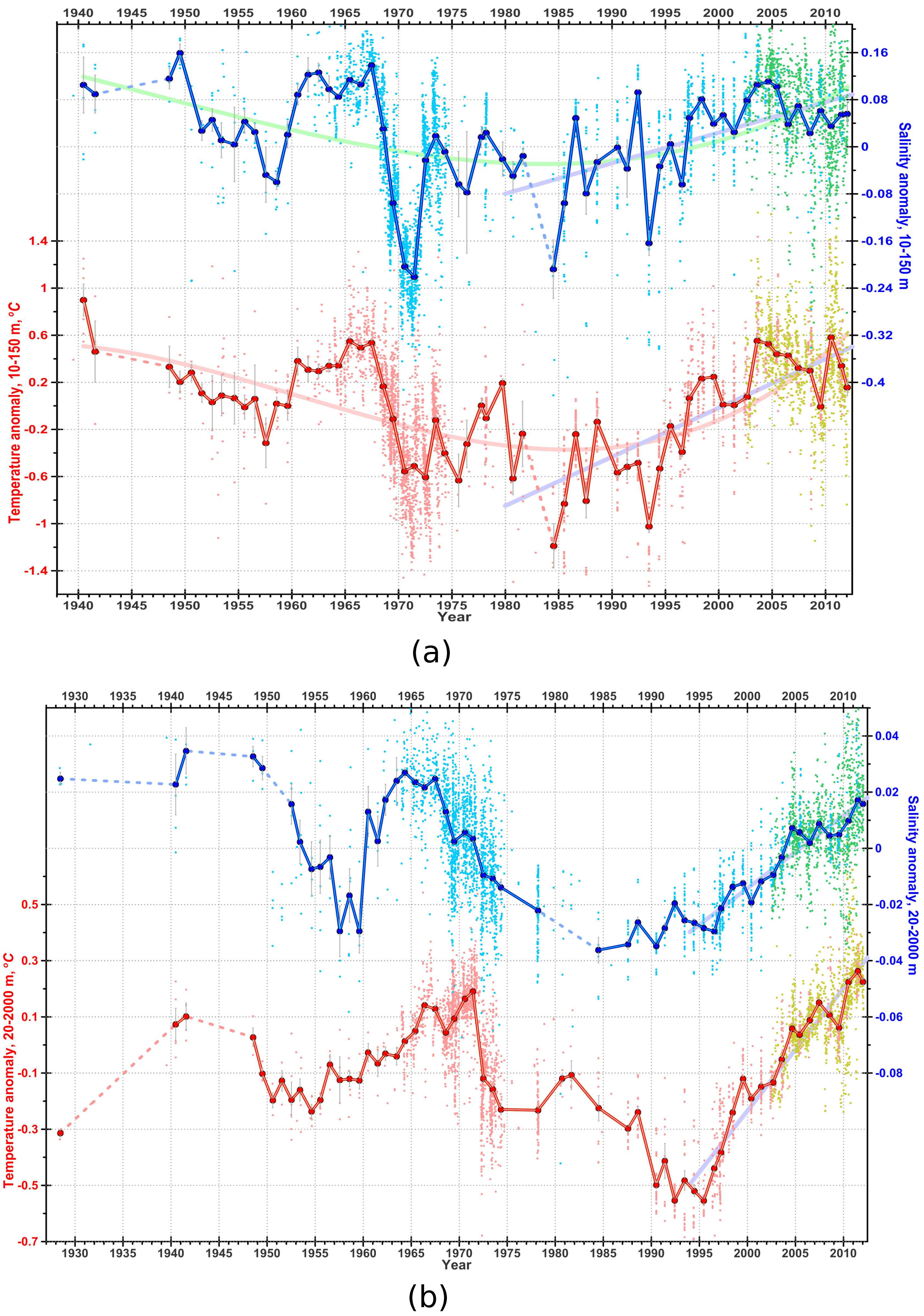
The mechanism of the recent warming in the North Atlantic was a focus of the debate over the role of interannual variations in atmospheric forcing, and in that of the Atlantic Meridional Overturning Circulation (AMOC). The North Atlantic gyre circulation intensified during the period of strong positive NAM/NAO in the late 1980s and early 1990s (Curry and McCartney, 2001). In the second half of the 1990s the intensity of the subpolar gyre circulation declined (Hakkinen and Rhines, 2004), which was contemporaneous with the onset of the warming trend (see Figure 3.4) in the Subpolar North Atlantic (Hatun et al., 2005). A number of studies have indicated that the NAO related atmospheric changes since the mid-1980s played a role in the dramatic changes of the heat storage observed in the Subpolar North Atlantic Ocean; e.g., Lozier et al. (2010); Robson et al. (2012); Zhu and Demirov (2011).

The NAO index magnitude has declined since 1995, while it has remained mostly in the positive phase. There were only isolated one and two year periods when the NAO index increased in magnitude. In particular, when the NAO was negative in the winter 1995/1996. Robson et al. (2012) suggested that the changes in the atmospheric characteristics related to this negative NAO phase after a prolonged period of positive NAO triggered the onset of warming period for the North Atlantic. Hakkinen et al.

(2011) found that the long term North Atlantic atmospheric variability was influenced by blocking events, which between 1996 and 2010 were more frequent than normal. For the North Atlantic region, "blocking" typically refers to atmospheric phenomena where large anticyclonic patterns persist for more than several days (Berrisford et al., 2007), although in general the term has broader application; cf., Tibaldi and Molteni (1990). These features block westerlies and divert the jet stream and storm tracks from their climatological positions. Years with more frequent blocking correspond to warmer and saltier Subpolar North Atlantic Ocean (Hakkinen et al., 2011). These are important features in their own right, with unique associations documented in Section 3.5. However, they are often grouped with different phases of the NAO in basic atmospheric analysis. Woollings et al. (2008) found that blocking can be triggered by upper level breaking of Rossby waves in the atmosphere over the North Atlantic. They found that Rossby waves breaking and episodes of blocking occur frequently when the NAO is negative in phase. The blocking effects are rare when NAO is positive. The study of Croci-Maspoli et al. (2007) confirms that the blocking in the North Atlantic anticorrelates with the NAO index. These authors made a comparison of the blocking effects over the North Atlantic and Pacific Ocean. Similarly to the North Atlantic case, they found that the blocking over the North Pacific anticorrelates with the Pacific North American pattern (PNA).

3.4 Patterns of the North Atlantic atmospheric variability

Historically, two conceptual views have been used by meteorologists in studies of atmospheric variability. These concepts differ in their interpretation of the distribution of instantaneous states of the atmosphere (in the space of all possible states). The



(a)

(b)

Figure 3.4: (a) Upper layer (10-150 m) temperature (red) and salinity (blue) anomalies in the Labrador Sea based in a combination of ship and Argo drifter measurements. (b) Temperature (red) and salinity (blue) anomalies in the Labrador Sea for the 20-2000 m layer based in a combination of ship and Argo drifter measurements. (Figure produced by Igor Yashayaev)

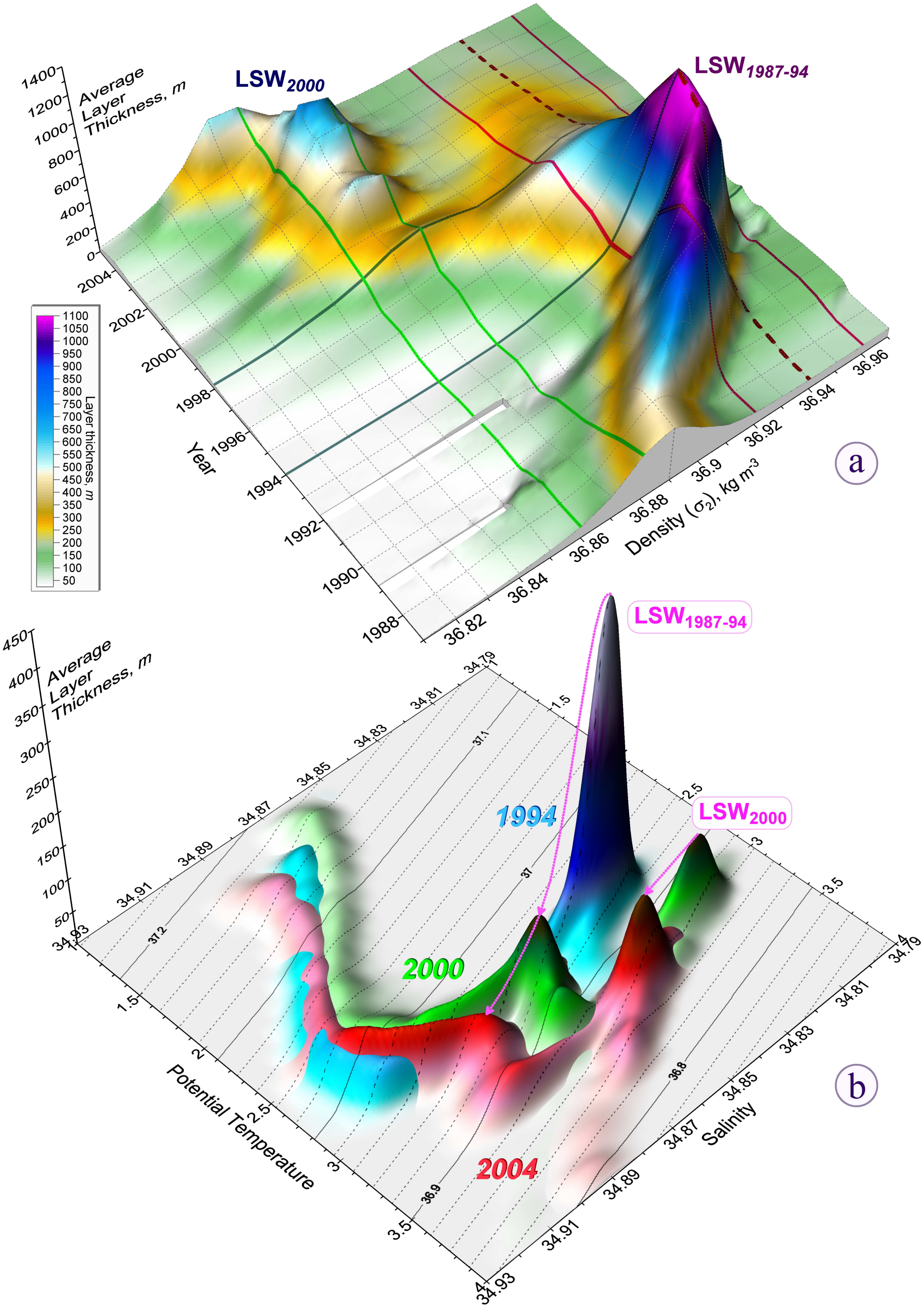
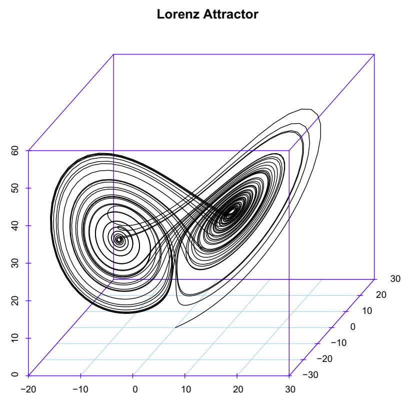


Figure 3.5: (a) A σ_2-time plot showing average thickness (m) of $\Delta\sigma_2 = 0.01$ kgm^{-3} layers in the Labrador Sea (σ_2 is potential density anomaly referenced to 2000 dbar]. (b] Potential temperature (θ)salinity (S) 'volumetric' projections of the 1994, 2000 and 2004 AR7W hydrographic section (Figure produced by Igor Yashayaev)

linear paradigm assumes that the atmospheric states are normally distributed around the climatological mean. Hence, the density of observed states decreases with the distance from the origin, which is the climatological state. Within the linear paradigm, the spatial structure of the patterns of atmospheric interannual variability as well as other teleconnection patterns were originally estimated using site to site correlation measures by Wallace and Gutzler (1981). More generally, they can be described as the leading modes of variability of pressure anomalies (at various atmospheric levels) for the region of interest defined by using Principal Component Analysis (PCA) or rotated variants of the same (von Storch and Zwiers, 1984).

The non-linear paradigm assumes that the climatological mean is not necessarily the most frequently observed state of the atmosphere. Rather, the distribution in the state space is skewed and multi-modal. An example of a similar system is provided by the Lorentz model (Lorenz, 1963), shown in Figure 3.6, where the distributions are indentified by the modes of the chaotic attractor. The most frequently observed states are not close to the mean of the solution, but instead in the neighbourhood of the stationary points of the system.

Chaotic non-linear dynamical systems, such as the Lorenz system, are sensitive to initial conditions; i.e., small differences in state can lead to a completely different system evolution over time. The resulting divergence of trajectories and non-repetitive behavior limits description and predictability, as information about "similar" trajectories or previous states does not necessarily prescribe current behavior (Lorenz, 1963). The attractors of such systems are quasi-steady-state solutions that the trajectories tend towards. As such, these tendencies give some predictability/structure to the system (Kalnay, 2002). These attractors can be quite complex; e.g., the Lorenz "Strange Attractor" where the quasi-equilibrium states involves chaotic oscillation between two distinct orbits (Sparrow, 1982), as shown in Figure 3.6.



Lorenz Attractor

Figure 3.6: A trajectory of the Lorenz 1963 dynamic system demonstrating the multimodal nature of the attractor.

The nature of the attractors and the location of bifurcation points (critical thresholds which determine which attractor a trajectory will be drawn towards) are dependent on the system dynamics which evolve with changes in external forcing (Hale and Kocak, 1991). Changes in the system dynamics, whether through variation in the coupled systems or larger climatic changes, will manifest themselves through shifts in frequency, residency, and transition statistics between attractors/attractor-modes (Corti et al., 1999), provided that the external changes are moderate. If alternatively the external forcing changes are large (as defined by the thresholds of the individual systems) shifts in external forcing can change the structure (or existence) of the attractors/modes themselves (Lorenz, 2006). Currently it is believed that the near-future evolution of the climate state will be within the first category (Terray et al., 2004).

The atmosphere is a multidimensional nonlinear dynamical system, whose evolution in time is typically described by a set of differential equations. Within this paradigm, so called atmospheric regimes are interpreted as the existence of quasi-equilibrium states or fixed points of attractors in the atmospheric phase space. When a state of the atmosphere is close to one of these attractors in the phase space, it remains there for a period of time much longer than the lifetime of weather disturbances. The persistence of such state is a result of the balance of dynamical tendencies with the eddy-mean flow interaction which exists when the state is close to one of the attractors (Molteni et al., 2006).

The dynamics of the atmosphere do not always allow for tractable analytic descriptions of attractors and bifurcations. Furthermore, multi-modal behaviour in the atmosphere need not be produced by non-linear interactions in the resolved dynamics, but rather by state dependent variations in the influence of unresolved sub-scale processes (Monahan, 2002; Sura et al., 2005). As such, statistical methods are used to look for modes of variability. These search for multi-modal behavior; e.g., Molteni et al. (2006), by identifying regions with high densities of occurrence within observational and simulated data (Casty et al., 2005). A related method is identifying self similar subgroups within the data by cluster analysis; e.g., Cheng and Wallace (1991). These methods have been shown to locate known modes of intensively studied chaotic systems such as the Lorenz attractor (Stephenson et al., 2004). However, such methods have been criticized for not meeting certain frequentest measures of significance especially when applied over hemispheric regions e.g., Stephenson et al. (2004). This is potentially due to the limited duration of observational data. As well, these methods do not necessarily determine the number of individual attractors and modes within a system, and attempts to do so can be sensitive to variations in time period and sampling (Christiansen, 2007). Inherently, the results of cluster analysis can always be further subdivided down to the level of individual data points, and to some degree the number of relevant modes can be more a function of the level of detail needed for a particular study than as an approximation of the system dynamics (Dennett, 1991). For some applications the classification methods presented here can be seen as an overlapping between the theoretical concerns of dynamical systems and the weather typing approaches of descriptive meteorology (von Storch and Zwiers, 1984). Irregardless of the sometimes exploratory nature of the analysis, the results and methods presented in this article have been successfully applied in the fields of predictive meteorology and downscaling; e.g., Corte-Real et al. (1999), Boe et al. (2006), Kannan and Ghosh (2010).

3.4.1 Fuzzy Clustering

Cluster analysis is a classification method which divides a data set into a predefined number of subsets of similar elements. Typically these subsets, referred to as clusters, are thought of having centres, which describe the characteristic pattern common to their elements. Algorithms are designed to subdivide the data so to maximise the distance/difference between centres, while also maximizing the similarity of the members assigned to individual clusters (Kaufman and Rousseeuw, 1990). The method has been used previously to describe large scale atmospheric circulation regimes over hemispheric domains; e.g., Cheng and Wallace (1991), as well as the North Atlantic region; e.g., Cassou et al. (2004), and Yiou (2004). Once defined these climate regimes can be statistically linked to local mesoscale phenomena; e.g., Cattiaux et al. (2010), OrtizBevia et al. (2011).

Most clustering methods classify data according to “crisp” or “hard” clusters where each data point is a member of exactly one distinct cluster. This can be described by saying that the membership of a given data point to a certain cluster is binary; i.e., either zero or one. Alternatively, fuzzy clustering allows a continuous range of membership to a cluster on the range [0,1]; e.g., a data point may have a 0.35 degree of membership to Cluster A and a 0.65 degree of membership to Cluster B, with the conditions that a data point’s memberships must be greater than or equal to zero and sum to one. These requirements result in this method sometimes being referred to as “probabilistic clustering”, where the interpretation is that the membership degree represents the probability that one would assign a given data point to a given cluster (Bezdek, 1981). However, as is the case with many clustering algorithms, this is a heuristic approach of which the formal mathematical properties of the results is not well studied. The more common interpretation is that fuzzy membership allows one to think of an object as being able to belong to two sets simultaneously. Either interpretation highlights that the degree of membership describes an uncertainty regarding classification, rather than reflecting a probability of occurrence (Kosko, 1990). It has been suggested that fuzzy clustering may be a preferred clustering approach to climate data, as notable variability in classification often occurs during cross-validation (Cheng and Wallace, 1991).

For this investigation, membership degrees are determined using the “FANNY” algorithm (Kaufman and Rousseeuw, 1990) as implemented by Maechler et al. (2005) for the software package R (R Development Core Team, 2011). For this application this amounts to a variant of the “fuzzy c-means” algorithm (Bezdek, 1981) using Euclidean distance, rather than the traditional squared Euclidean distance, since the former method has less outlier sensitivity, and better represents non-spherical clusters (Kaufman and Rousseeuw, 1990). To prevent the utilized algorithm from converging to “crisp clusters”; i.e., outputting only membership values of zero and one, it is necessary to transform the membership coefficients (m) within the algorithm so that there is an uneven response between changing “high” ($m \to 1$) and low ($m \to 0$) membership values. Typically the membership value is raised to a power (Klawonn and Hoppner, 2003); i.e., for membership m, $m \mapsto m^k$, where $k = 1$ results in crisp clusters and $k \to \infty$ will produce completely fuzzy; i.e., equal membership, delimitation. This value must be set by hand. Values are tested by comparing distributions of membership coefficients produced by data and by red noise simulations, so as to check that distributions of membership values resulting from the data are bimodal, while distributions from red noise are typically ($>$ 95 % occurrence) uni-modal; i.e., we check that we do not use so low a value of k that would "force" the appearance of distinct clusters onto a uni-modal data set. Seeing that it is possible to create fuzzy yet significantly distinct clusters also serves as a means to check the choice of number of centres (Horenko, 2010).

3.4.2 Gaussian Mixture Models

A more formally probabilistic alternative to the fuzzy clustering described above are Gaussian Mixture Models (GMMs). GMMs model a set of (multivariate) data by describing it as being generated from a combination of Gaussian distributions. The task is to estimate how many distributions; i.e., clusters, comprise the sample, the percentage each distribution "contributes" to the data sample, and the mean and (co)variance terms for each distribution. GMMs are conceptually quite similar to the described heuristic methods. The k-means algorithm and fitting a GMM where the covariance matrices are set to be diagonal, equal, and the same for all clusters, both depict drawing spheres within the phase space to define data groupings (MacKay, 2003). However, GMMs are a mathematically formal approach, fit to different metrics, that give a continuous probabilistic measure of membership across the phase space. As such, the mathematical properties of the resulting models are rigorously defined. This gives some advantages over the clustering algorithms discussed above. Namely, they can be fit with Bayesian methods, giving a format for comparing models and describing parametric uncertainties. However, the parametric nature; i.e, the use of defined distributions, of the model can overly restrict the form of the solution.

One variation on the method is the "Infinite Mixture Model". Here, rather than pre-selecting an initial number of clusters for the model, this number is set a priori to infinity, has an associated prior distribution, which takes the form of a concentration parameter dictating how diffuse the observed data is believed to be (Neal, 1991). This gives a means of calculating uncertainty for the number of clusters, but makes it difficult of make an ensemble estimate of the other parameters, since their number and meaning are different for each sample.

The results presented below are produced using the software provided at `http://www.cs.toronto.edu/~radford/fbm.software.html`.

3.5 The North Atlantic weather regimes

In the early development of atmospheric forecasting, meteorologists invented classifications for typical weather regimes or so-called Grosswetterlagen (Baur et al., 1944). At the time this approach was instrumental in the development of methods for short and medium range weather forecasting. The major assumption behind this approach is that certain atmospheric patterns can persist on time scales larger than the typical life time of atmospheric weather events and disturbances. Three types of weather regimes were identified over Europe by Hess and Brezowsky (1952), e.g. zonal, blocking and mixed. The transition probabilities between these patterns were used as input information for weather forecasts; cf., Spekat et al. (1983).

The weather regimes are defined as points in the phase space where the atmosphere is in statistical quasi-equilibrium. In these points the dynamical tendencies of large scale flow are balanced by mean-eddy interaction (Molteni et al., 2006). If the state of the atmosphere is in a close vicinity of one of these quasi-equilibrium states on the phase space, then the atmosphere will remain in this area over a time frame longer than the typical life time of atmospheric disturbances.

Four dominant weather regimes have been identified, using crisp clustering methods, for North Atlantic SLP anomalies, by Yiou (2004), Cassou (2008), OrtizBevia et al. (2011), Cattiaux et al. (2010), and Hakkinen et al. (2011). These regimes correspond to asymmetrical descriptions of the positive and negative phase of the NAO. This asymmetrical description is able to better classify NAO+/- events than classifications based on linear statistics (Cassou et al., 2004). The analysis also reveals two additional features. One such feature is the Scandinavian-Greenland dipole (SG), with low and high pressure features over Greenland and Scandinavia respectively. The second is defined by a region of high pressure south of Greenland, referred to as the Atlantic Ridge (AR) (Cassou, 2008). The spatial structure of these patterns is robust and shows little sensitivity to the difference in the clustering methods and the period averaging of the analyzed data set (Cassou, 2008). These North Atlantic weather regimes have been successfully used in a number of recent studies of atmospheric variability over Europe and North America. Yiou (2004) found that the regional extreme precipitations and temperature over the North America and Europe are connected to the type of dominating atmospheric weather regimes. The relation with the weather regimes was used by OrtizBevia et al. (2011) to explain the extremes of precipitations over the Iberian Peninsula (Cattiaux et al., 2010).

Here we test using a Bayesian approach to classification so as to better capture uncertainties which will be relevant in following sections. We use winter (DJF) SLP fields from atmospheric reanalysis data, provided by the National Center for Environmental Prediction (NCEP) (Kalnay et al., 1996), for the North Atlantic region, specifically 20°N : 80°N and -90°E : 30°E, so as to match previous studies. The classification methods described above are multivariate methods, but tend to lose effectiveness for high dimensional data sets. In the following analysis the dimensionality of the data set is first reduced by performing PCA and retaining the leading ten PCs, which explain

81% of the variance. The clustering algorithms are then applied to the time-series of expansion coefficients for these PCs rather than to those of individual field elements. Parameters for the GMMs are estimated using Bayesian inference[1]. The GMM is set *a priori* as having four spherical components so as to be in agreement with the k-means approach of Cassou et al. (2004). Further analysis is performed using one hundred samples drawn from the resulting posterior distribution of possible models.

Figure 3.7 shows the posterior mean centers of the four clusters. These match the patterns reported in previous studies. The robustness of the estimated centres is studied here in terms of the standard deviation of centres calculated from the posterior samples. These are shown in Figure 3.7. The range of difference between sample estimates is small compared to the distance between the centres themselves. Most of the posterior variability occurs at the edges of the features described by the centres. For the NAO+ and NAO- features the variability in the models is primarily related to shifts in the North-South orientation of the extents and centres of the dipoles. Variability for the SG Dipole is mostly in the western extent of the Scandinavian high pressure feature. Different models shift the AR feature north or south, and vary to a lesser amount concerning its east-west extent.

Figures 3.8 and 3.9 show the variations of the frequency of the winter weather regimes on two different time scales - daily and annual. The probability of membership to each cluster for each day of the winter of 2012 is shown in Figure 3.8. The mean membership probabilities are displayed, as well as those for individual samples from the posterior. The NAO+ and SG weather regimes dominate the North Atmospheric circulation for the most of the winter of 2012. The AR pattern is present only during two weekly periods in January and February and NAO- probability membership is

[1] Gaussian priors are used to ease implementation but set to be wide enough so as to be essentially uninformative. Further testing shows that the results are insensitive to specifications regarding the width of the priors or variations in sampling.

very low during the whole period. The winter mean probability of membership, again shown as posterior mean as well as sample values for each winter is shown in Figure 3.9. In NAO+ and GS frequency peaks in the winter of 2012. The mean probability membership for AR in 2012 is significantly smaller and for NAO- it is close to zero. The NAO index reported for 2012 was overwhelmingly positive, corroborating our results, which have the additional advantage of showing the presence of the AR events. Typically AR features are grouped with NAO- events in the classical indexes. Both time series as well as the cluster centres match estimates created using other clustering methods as well as optimized GMMs (not shown), but have the extra feature of being able to approximate error bars for the presented descriptions. The interannual variability in regime distribution, for the past fifty years, is discussed in the following sections.

3.6 Patterns of North Atlantic atmospheric interannual and decadal variability

Some weather regimes may dominate in the North Atlantic atmospheric circulation for years or decades before being replaced by other regimes. In the past fifty years, the NAO- and AR regimes dominated in the 1960, while the NAO+ and GS regimes were dominant in the 1980s and 1990s (see Figure 3.9). Such long term variability is often regarded as regime change or sometimes as a climate change signal (Lorenz, 2006). An important question in the context of understanding the long term North Atlantic atmospheric variability is if and how the weather regimes changes are linked to external forcing triggered by global climate change, or large scale decadal atmospheric variability of the zonal circulation in the Northern Hemisphere.

Two types of response by the atmospheric circulation are possible to anomalies

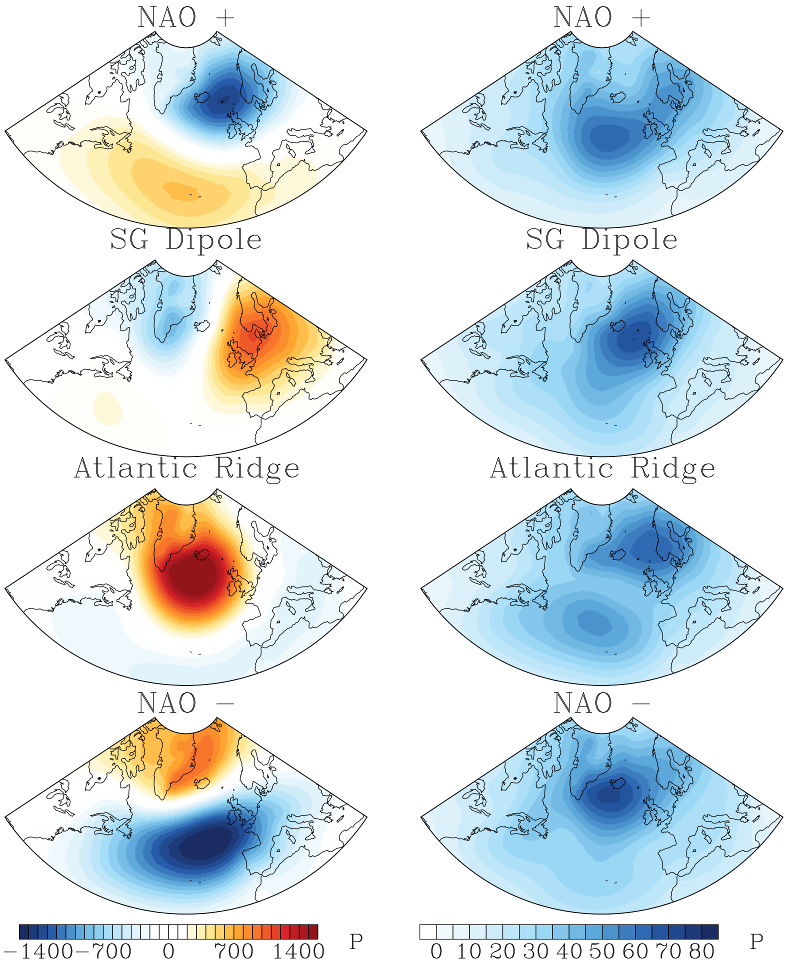


Figure 3.7: The mean posterior estimate of centers calculated using Bayesian Gaussian Mixture Models for winter (DJF) daily SLP anomalies and the standard deviation between samples.

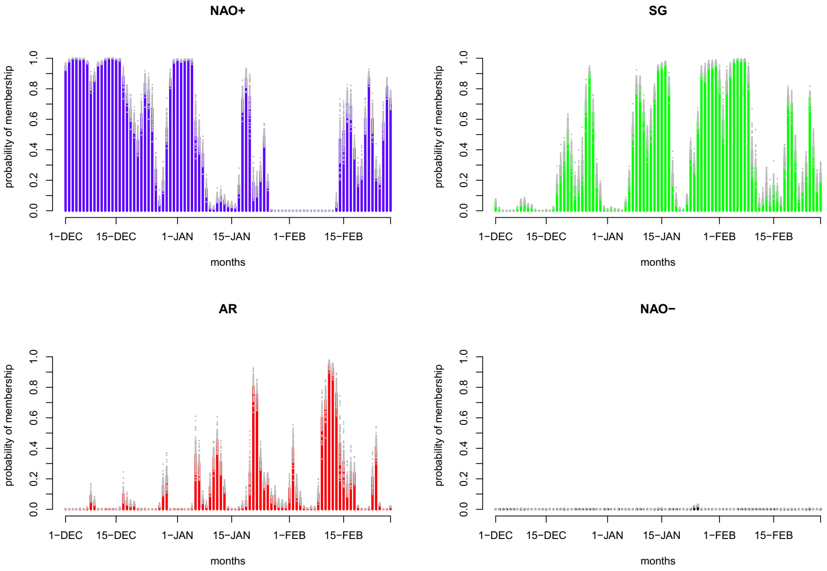


Figure 3.8: Daily probability of membership for the 2012 winter (DJF) season for the Bayesian GNNs, bars give the mean posterior estimate, gray dots give the results for the centres generated by individual samples.

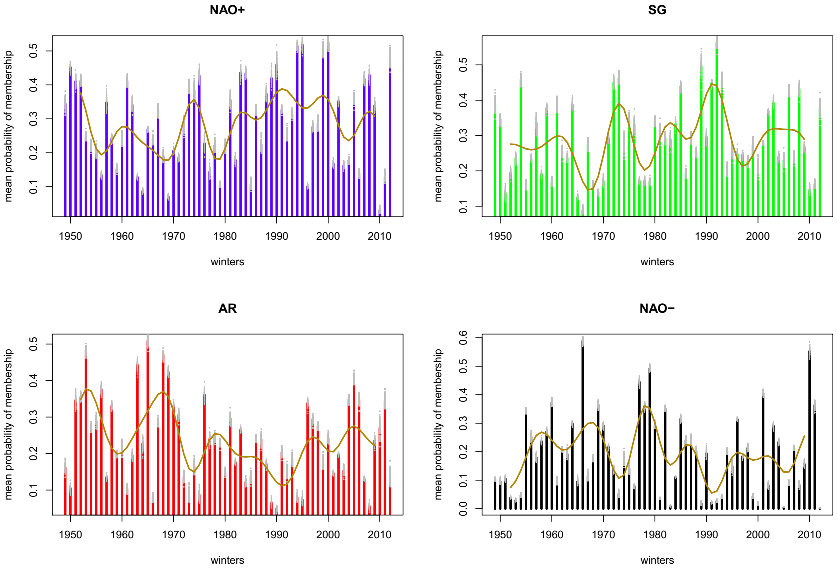


Figure 3.9: Mean winter (DJF) probabilities of membership for the Bayesian Gaussian Mixture Model SLP centres, bars give the mean posterior estimate, gray dots give results for individual posterior samples of centres. Brown curve gives a smoothed time series of the mean probabilities.

in the forcing or boundary conditions (Molteni et al., 2006). If the atmosphere is subject to weak and persistent forcing, then the number of the regimes and their spatial structure remain constant and only small variations in the position of the centres and changes in the frequencies of regimes will occur (Palmer, 1999; Corti et al., 1999). If the external forcing is strong enough, the number and centres of the regimes may also change (Molteni et al., 2006). Existing studies suggest that the change in the North Atlantic atmospheric regimes most probably falls into the first group (Lorenz, 2006). In this case, the weather regimes do not change their spatial patterns, but the external forcing “nudges” the nonlinear dynamical system causing domination of some weather regimes (Corti et al., 1999). Due to its intrinsically chaotic nature, the atmosphere will still occupy states in the phase space that are in vicinity of all quasi-equilibrium points, although the weather regimes that are favored by the forcing will occur more frequently than the others, e.g. the external forcing will cause the atmosphere to stay close to these weather regimes longer than for the remaining regimes. Here this notion is demonstrated using the chaotic Lorenz system subjected to intermittent external forcing, in an experiment similar to that of Corti et al. (1999).

3.6.1 The low frequency patterns: example of the Lorenz System

The solution of the Lorenz equations is given in the top panel of Figure 3.10. The Lorenz system has two dominant regimes and the solution exhibits irregular fluctuations between these two modes. In this experiment external forcing is applied, so that the system favours one mode of the attractor over another, as previously described by Palmer (1999) and Corti et al. (1999). The use of external forcing in these experiments serve only to create statistical shifts such as observed in the atmosphere, it is not meant to imply the physical origin of such shifts in the atmosphere occur at certain scales or that they have well defined predictability.

The characteristics of long term transitions between the states favoured by the external forcing can be identified through analysis of the low pass filtered solution, which is shown in the middle panel of Figure 3.10. The filter removes the most energetic high frequency transitions leaving only the local mean. It is calculated as a moving average, with a smoothing interval comparable to that of the external forcing transitions, so that each point depends on the local mean residence time for the two modes and their amplitudes. As such, the filtered curve indicates the phase of low frequency variations in the system. The extrema in the middle panel of Figure 3.10 correspond to periods when one of the modes dominates in the solution, e.g. the external forcing favors one of the attractors. In the transition periods, the filtered solution has low magnitudes reflecting the fact that the modes are almost uniformly distributed under weak external forcing.

This approach of using low pass filtered solution can be extended towards the study of low frequency variability in the high dimensional atmospheric state at least in the following two ways:

(i) Using fuzzy cluster memberships calculated from the filtered data. The memberships for the Lorenz system is shown on the bottom panel on Figure 3.10. These membership series relate to the entire (multi-dimensional) data set, and so are representative when, unlike for the example problem here, projection onto a single component is insufficient to describe the system. The memberships to the centres of the Lorenz system (bottom panel Figure 3.10) clearly indicate the long term shifts in the solution driven by external forcing.

(ii) Examining the distribution the modes within the clusters identified from the filtered data. This probability distribution for the Lorenz system is shown on Figure

3.11. It indicates the change in the frequency of occurrence of the two modes relating to different external forcing.

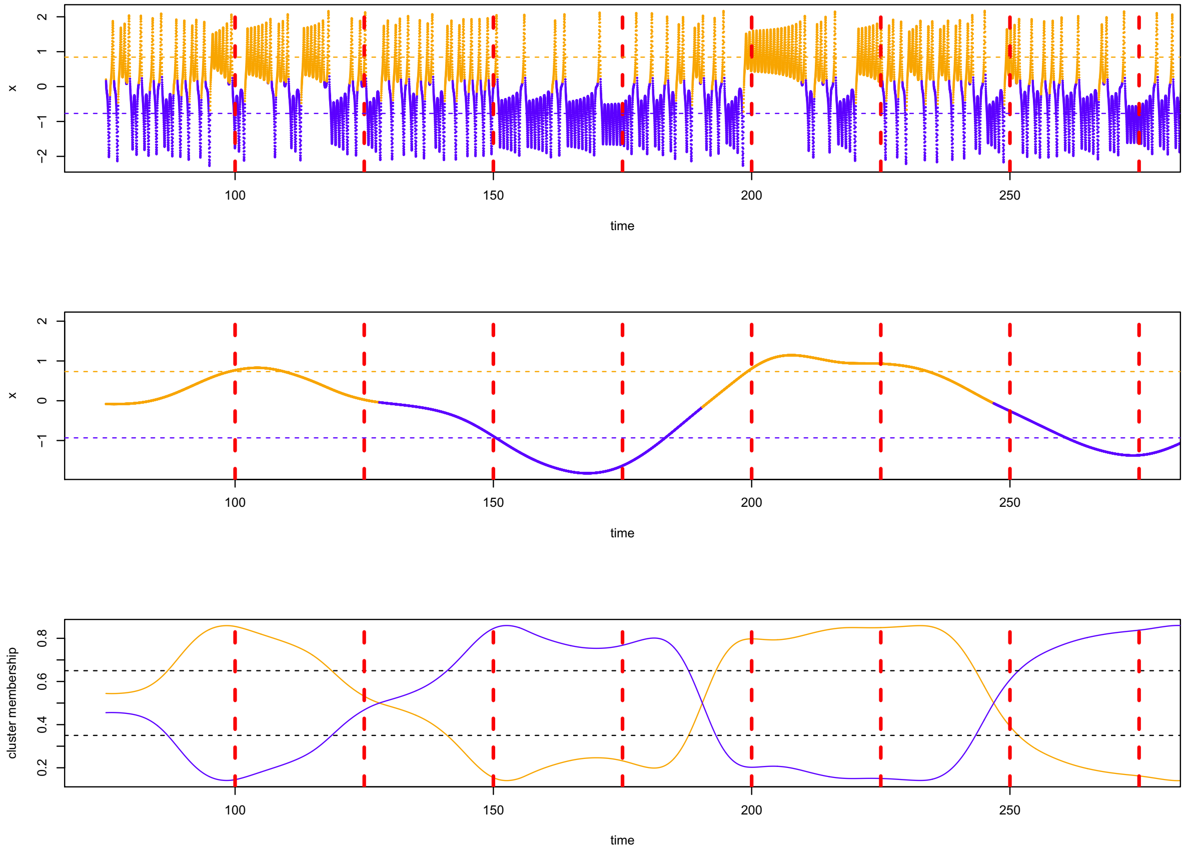


Figure 3.10: Top panel shows the x-component of a simulation of the Lorenz model, with varying external forcing being applied intermittently through out the run. Colouring displays the results of using clustering to identify the two modes of the system. The dashed horizontal lines show the x-coordinate of the calculated mode centres. The vertical lines show where the nature of the external forcing is changed, in a repeating sequence of no forcing, forcing towards positive mode, no forcing, forcing towards negative mode, etc. The middle panel is the same as the top but calculated using the running mean of the original data. The bottom panel shows the fuzzy memberships as calculated for the filtered data presented in the middle panel. Dashed horizontal lines divide the bottom panel into thirds to aid visualisation.

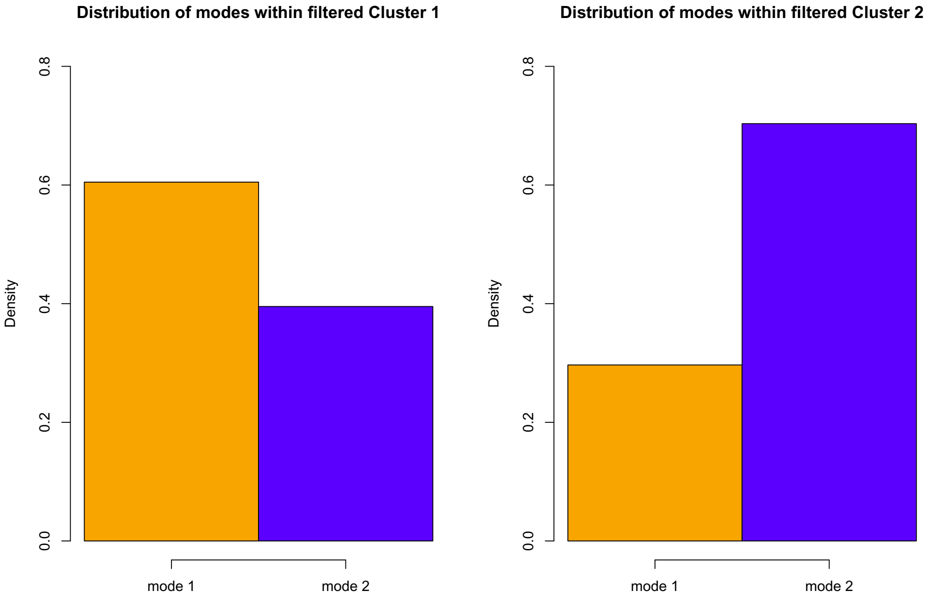


Figure 3.11: The distribution of occurrence for the system modes within the clusters identified using the filtered data for the Lorenz system example.

3.6.2 Patterns of low-frequency North Atlantic atmospheric variability

Esbensen (1984) identified four dominant patterns in Northern Hemisphere 700 mb geopotential data low-pass filtered in the interannual band. They resemble the structure of the Pacific-North American (PNA), North Pacific (NP) patterns; cf., Wallace and Gutzler (1981), the Northern Hemisphere Annular Mode (NAM/NAO) and the Eurasian pattern. Esbensen (1984) also found that the three interannual modes PNA, NP and NAM/NAO are correlated, suggesting that they may not be independent modes of atmospheric variability. The NAM/NAO and PNA are the two interannual atmospheric patterns described that have strong impact on the North Atlantic variability (Esbensen, 1984).

The two dominant PCs for the Northern Hemisphere monthly mean 500 $mathrmmb$ geopotential height (H5) fields are shown on Figure 3.12. The first PC has a spatial structure that resembles the PNA; cf., Wallace and Gutzler (1981), but with the centres over the Gulf of Alaska and Florida shifted northeastward of their canonical positions. This mode describes a variability which is out of phase in the Gulf of Alaska and near the southern tip of Greenland. The second PC descries a pattern that contains elements of NAM/NAO and North Pacific Oscillation. The anomalies related to this PC at the centres in the Gulf of Alaska and near the southern tip of Greenland are in phase.

Here we study the interannual patterns of North Atlantic variability by using the low-pass filtered solution. As discussed in the previous section we focus more specifically on the spatial patterns of the clusters centres and probability distribution of the weather regimes related to each of the clusters. We analyze winter (DJF) H5 anomaly fields from the NCEP reanalysis (Kalnay et al., 1996), over the domain of 25°N : 75°N (Bahamas and Canary Islands to Baffin Bay and Barents Sea) and -105°E : 45°E (Gulf of Mexico and Hudson Bay to edge of Scandinavia and Mediterranean). This is a higher elevation field and wider region than those used for the weather regime classification, which were selected to allow comparison between our method and previous studies. We are looking to compare surface events with variations in long term processes, which include variations in upper troposphere processes such as the jet stream. The H5 is a linking field that offers a compromise between the two levels, and is used this way in previous studies and in meteorological applications. As these processes incorporate continental effects, we expand the east-west range of the region, although the presented results are largely insensitive to small changes in the study area. Low-pass filtering is performed with a Lanczos filter (Duchon, 1979). The clustering is performed on the leading ten PCs of the data, which account for 90% of the variance.

Fuzzy clustering, rather than GMMs, is used to analyze the interannual patterns of

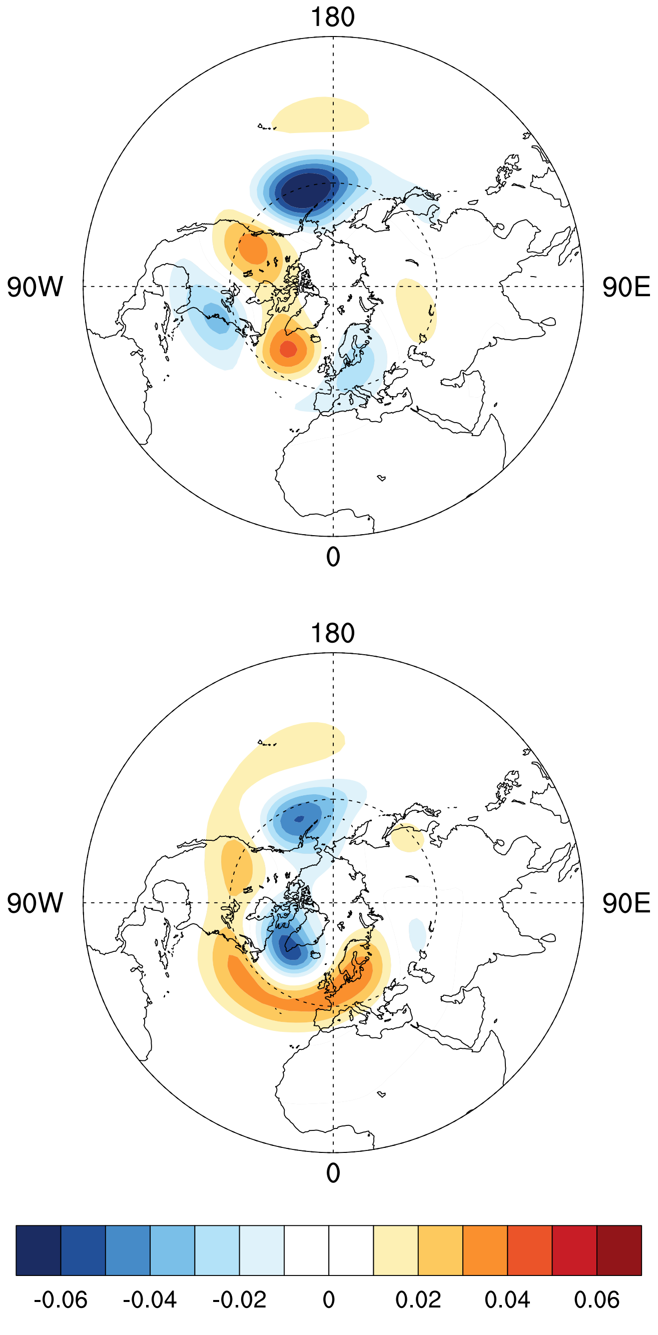


Figure 3.12: Dominant PCs of montly mean 500 mb geopotential height field

variability. The method is non-parametric, yet still allows for continuous memberships to be determined. GMMs assume a structure where the data is clustered around the centres in a Gaussian fashion. This gives a first approximation of the form of the data set, but does not well match the appearance of the data. This assumption also results in very crisp cluster memberships for the classified data points. This is not supported by visual inspection, and given the time scales considered the PDF is not expected to identify unique physical processes or dynamic effects. Rather, multiple centres, skewness and other deviations from linearity are potentially indicative of shifting trends and tendencies at shorter time scales (Teng et al., 2004). The number of clusters to use is investigated by two-dimensional; i.e., using the first two PCs, kernel density estimates of the PDF (Figure 3.13), which suggests four notable modes when compared against red noise simulations. This number is confirmed by fitting an infinite mixture model to all ten PCs and finding four clusters to be the mode of the posterior estimate[2]. The cluster centres are given in Figure 3.14.

The first(counting left to right, top to bottom) and fourth clusters match (see Figure 3.14) depict the positive and negative phase of the NAO, with the maximums and minimums of the meridional dipoles of these two clusters zonally elongated. The second and and third clusters have spatial structures that resembles the regional representation of the two opposite phases of PNA pattern with the Florida centres shifted eastward and the Scandinavian centre shifted northwestern in the second cluster. The dominant element of the second and third centres are pressure anomalies southern of Greenland. In the following, the second centre is refereed as the G- pattern. Its opposite pattern defined by the third cluster centre is labeled as G+. A centre similar to that of G- in Figure 3.14 was identified by Cheng and Wallace (1991) (cluster A,

[2] This result is obtained using an *a priori* concentration parameter set to favor a limited number of clusters. This is deemed appropriate based on the previous investigations and our initial beliefs about the system dynamics.

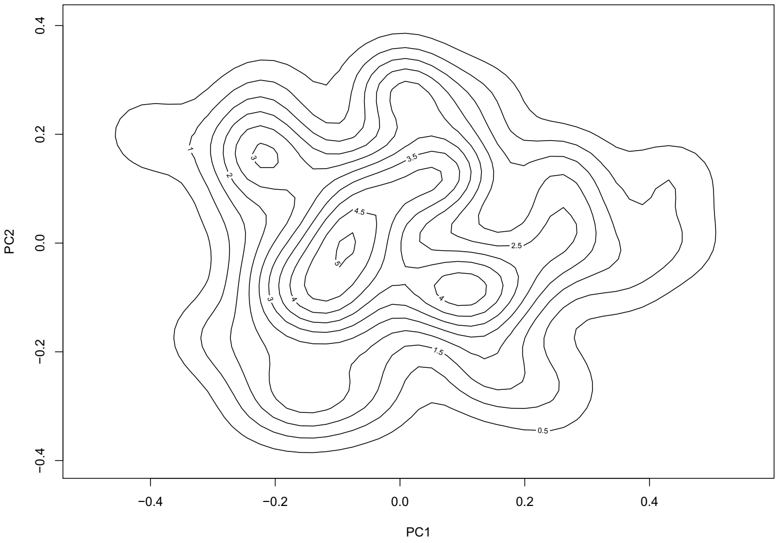


Figure 3.13: Kernel density plot showing the distribution of the leading two PCs of the interannual H5 field.

Page 2681 in that paper) in an analysis of the Northern Hemisphere H5 data, low-pass filtered to remove variations with period less than 10 days. The fourth (NAO-) centre in Figure 3.14 resembles the regional structure of the centre labeled as cluster G by Cheng and Wallace (1991). While the spectral interval represented by the data and the spatial domain for the cluster analysis in Cheng and Wallace (1991) differ from the ones used in our study, the similarities in the regional structure of the centres over the North Atlantic suggest that they are robust. This also suggests that there may be spatial correlations between the regional patterns on Figure 3.14 and some elements of larger scale atmospheric variability over the Northern Hemisphere.

A time series comparing the dominant cluster for each winter against the NAO index[3] is shown in Figure 3.15. The NAO+ and NAO- interannual patterns are dominate in years of high and low NAO index respectively. The two other clusters, G+ and G-, dominate mostly in years of low magnitude of NAO index. More specifically,

[3] As provided by `http://www.cpc.noaa.gov/`.

the G+/G- dominate in some the years in the 1950s when NAO index was negative and since 1995 when NAO index was positive. In both cases the NAO memberships in the low-pass filtered data had a low magnitude.

Another way of representing the time series for the cluster centres is shown on Figure 3.16. The two panels show the time variability of the fuzzy clustering for NAO+/NAO- (upper panel) and G+/G- (bottom panel). The patterns on the two figures clearly exhibit oscillating behavior. This connections between the oscillating pattern can be expressed in terms of the coefficient of correlation which is very similar for the two modes. For both G+/G- and NAO+/NAO- the correlation coefficient between the membership of the opposite patterns is very close and about 0.46. This coefficient is higher when calculated only for years strong NAO+/NAO- or G+/G-. If we remove the years of strong NAO+ or NAO- the G+/G- membership correlation coefficient increases to -0.58. The NAO+/NAO- membership correlation is again very close -0.58, when calculated for years of low G+/G- membership.

To examine the relationship between the structure of the patterns of variability, on interannual and sub-seasonal time scales, we look at the frequency of occurrence for days mapped to the above described sub-seasonal regimes within days mapped to the interannual regimes (when defined as crisp clusters). This result is shown in Figure 3.17. The analysis is performed using posterior samples from the Bayesian GMM discussed above, and the distribution of the results is presented. From this we see that in years of dominant interannual NAO+ pattern, the most frequent sub-seasonal weather regimes are NAO+ and Greenland-Scandinavian (G-S) Dipole features, with a high probability that the G-S Dipole is more dominant. The interannual NAO- pattern shows a mirror effect with a dominance of the sub-seasonal NAO- and Atlantic Ridge features. The other two interannual features are more evenly mixed although they do show a significant redistribution of the sub-seasonal components. We suggest that the interannual clusters presented here represent shifting in distributions of the sub-seasonal patterns, which have tendencies to be grouped in distinct combinations.

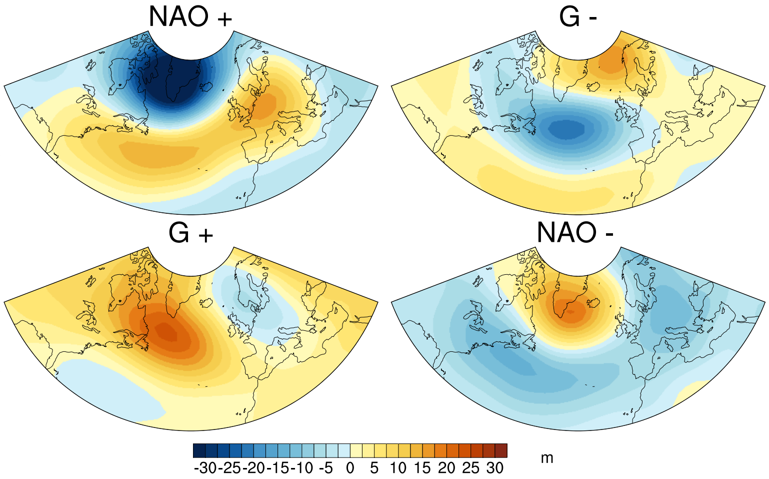


Figure 3.14: Cluster centres of the interannual H5 data, found through fuzzy clustering.

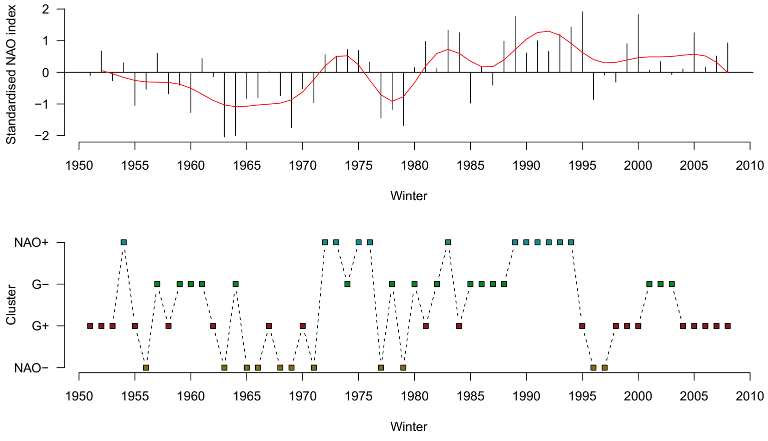


Figure 3.15: (a) The NAO index as provided by NOAA (upper panel) (b) time series of dominant clusters of the interannual H5 data.

3.7 SST patterns of interannual variability

The life time of weather disturbances is significantly smaller than the typical time scales of ocean surface and deep layer variability. Hence, the null hypothesis of climate

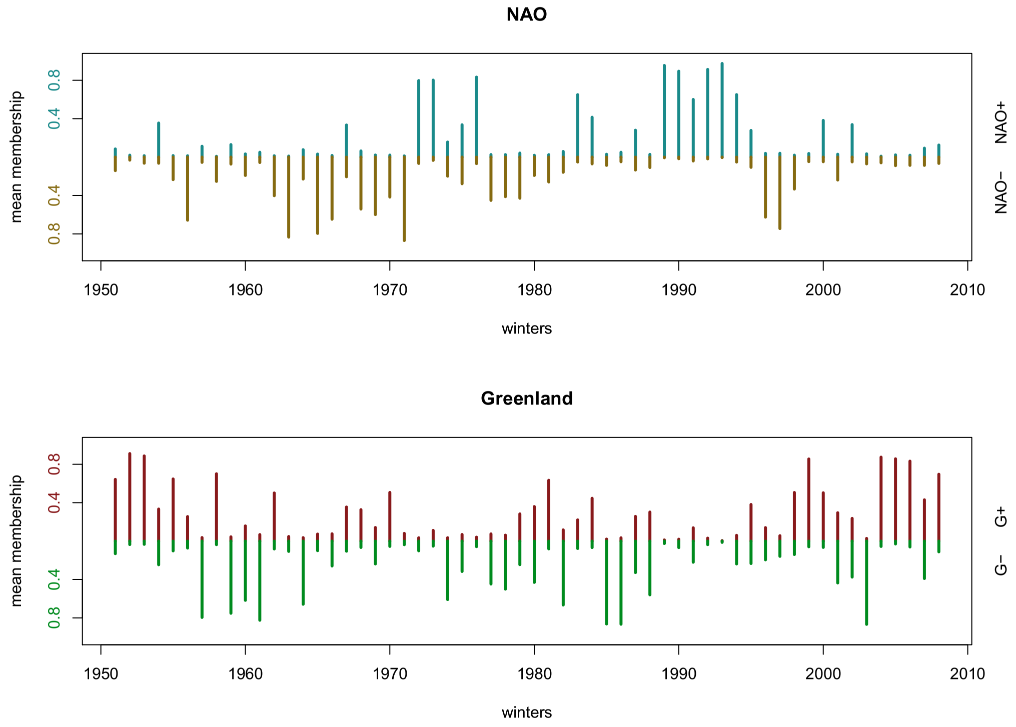


Figure 3.16: Mean winter (DJF) memberships for the Fuzzy Clustering interannual H5 centres.

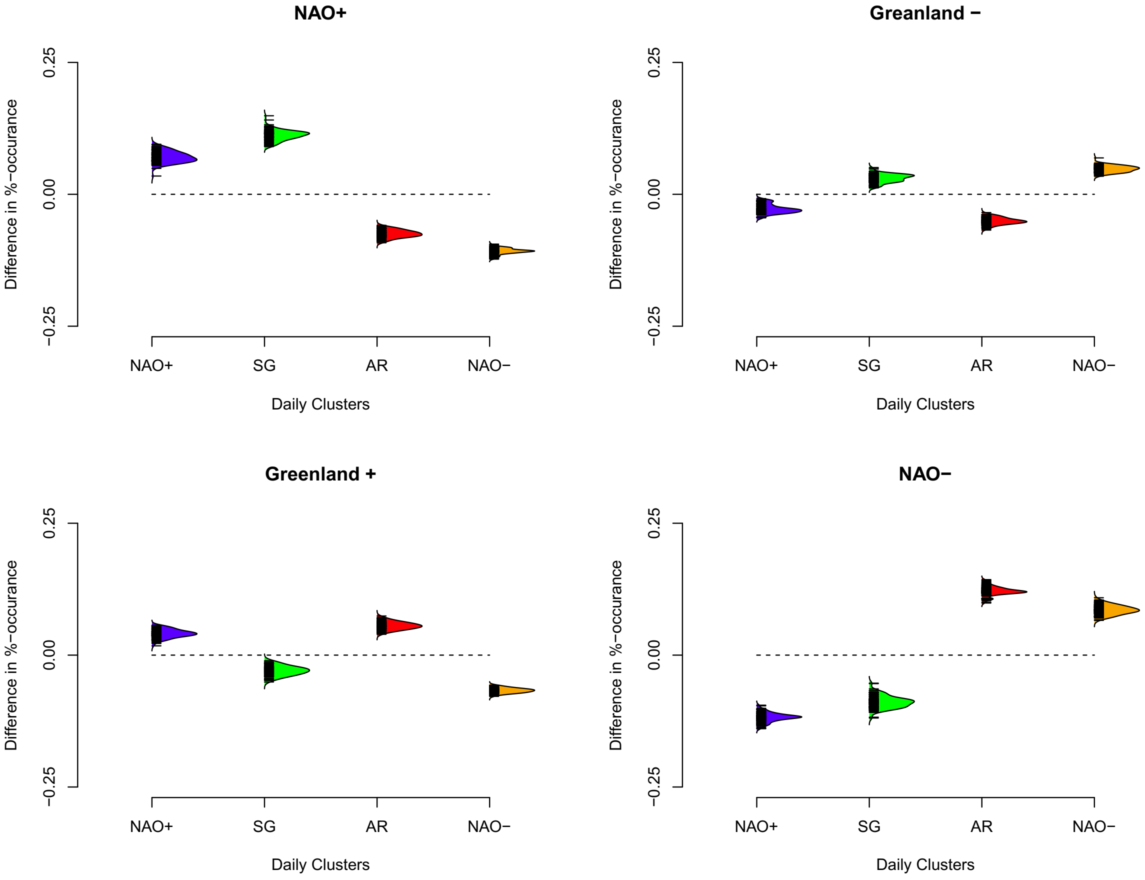


Figure 3.17: The frequency of occurrence for the sub-seasonal data for each interannual regime. The beanplots show how this frequency is distributed for different samples from the posterior estimate of the sub-seasonal classification. Dashed line shows the 25% occurrence mark; i.e., the line the sub-seasonal regimes would follow if all contributed equally to a given interannual regime.

variability of Hasselmann (1976) assumes that the climate system variability consists of two parts: (1) a fast component which is the atmosphere, and (2) a slow component - the oceans. The effects of the fast atmospheric component on the surface ocean mixed layer is represented by white noise. In this case the SST variability is a result of the integration of surface heat fluxes and is an auto-regressive processes of first order. Within this paradigm, the ocean is a passive element of the system which is forced by the atmosphere and influences the long term climate variability mostly through its large heat capacity and the dynamical processes in the ocean are not considered. The results of Frankignoul and Hasselmann (1977) demonstrated that on time scales shorter than a decade, SST variation can be approximated as a first order auto-regressive process.

Dommenget and Latif (2002) studied the spectrum of long term observed and simulated SST. They found that neither spectrum corresponded to that of a first order auto-regressive process. Dommenget and Latif (2002) explained these deviations as an increase in the variance of SST oscillations on interannual to decadal time scales, triggered by internal ocean dynamics, specifically lateral heat transport.

The idea that decadal variations of SST are driven by the variability in the meridional heat transport dates back to the work of Bjerknes (1964). Delworth et al. (1993) found that the AMOC exhibits oscillations with a period of about 50 years, which lead to variance of SST of approximately 0.5C in the sub-polar North Atlantic. Delworth et al. (1993) suggested a mechanism of decadal AMOC variability explaining the connection between variation in horizontal ocean transport, properties of the water masses, and deep convection in the Subpolar North Atlantic. A weakening of the AMOC reduces heat transport, which, following decadal scale lags, causes cooling of the water masses and anomalously high salinity in the region of deep water formation. This strengthens the AMOC, leading to an increase in transport of warm and salty waters into the Subpolar North Atlantic. This in turn reduces the intensity of vertical convection, and as a result produces a new decay in the AMOC.

Delworth and Greatbatch (2000) have demonstrated that there is no evidence that the AMOC is a part of a dynamically coupled mode of atmosphere and ocean, nor that the AMOC driven variations in the SST have any significant impact on the atmospheric circulation. Rather, the long term meridional heat flux in the ocean and atmosphere is positively correlated with ocean horizontal heat transport, and negatively correlated with transport in the atmosphere (Delworth and Greatbatch, 2000). Thus the ocean is the driver of the variability in the AMOC, while the atmosphere compensates for the long term changes in the ocean heat transport.

On time scales shorter than the period of the AMOC, the SST anomalies are forced by the atmospheric variability. Like in the Hasselman-1976 model, this forcing is triggered mostly by fast weather systems with time scales smaller than the one of the surface mixed layer variability. The type of this forcing, however, depends on the frequency of occurrence of the dominating weather regimes (Figure 3.7). The NAO+ and GS regimes are related to deepening of the Icelandic minimum. This favors colder than average winters in the Subpolar North Atlantic. The NAO- and AR regimes favor blocking events in the region, which divert the storm tracks from their climatological path. When these two patterns dominate the winters are warmer than average. Within the paradigm of atmosphere as a multiscale nonlinear system, the probability distribution of these regimes is conditioned upon the interannual patterns of variability (see Figure 3.17).

Here we study the correlation patterns between monthly mean interannual patterns memberships and SST (Figure 3.18). Note that high temporal correlation in the data reduces the significance of correlations found between data sets (Zwiers, 1990). To determine significance we test correlating the data sets against red-noise simulations constructed to have the same temporal auto-correlations as the membership indexes (Ebisuzaki, 1997). Correlations were only considered significant if they were higher than those found when comparing red noise simulations of the index time series to the SST fields, so as to maintain the spatial autocorrelation structure of the data. We find that accepting correlations above values of 0.3 allows for a very conservative claim that the correlations are above 95% significant.

The correlation maps between SST and monthly means fuzzy membership (Figure 3.18) show two oscillatory patterns of SST that are forced by the G+/G- and NAO+/NAO-. The NAO+/NAO- driven SST anomalies resemble the well known tripole pattern; cf., Visbeck et al. (2003). The tripole pattern has a cold anomaly in the subpolar ocean during positive NAO phase in the sub-polar and equatorial North Atlantic ocean and warm anomaly in the subtropical sector. The spatial pattern for the negative NAO phase is symmetric to the NAO+ pattern with opposite signs of the SST anomalies.

The G+/G- driven SST anomaly is less symmetric than the tripole pattern and has magnitude of the SST anomaly associated with the G+ pattern much stronger than the one for G-. The SST spatial pattern for G+ in the subpolar North Atlantic has a spatial structure similar to the one for NAO- with the warm centre in the subpolar ocean shifted towards the Eastern North Atlantic. Both interannual patterns G+ and NAO- (Figure 3.14) are related to higher occurrence of the AR and NAO- weather regimes (see Figure 3.17). The latter favor atmospheric blocking over the sub-polar North Atlantic, which can promote anomalous distributions of heat within the region.

While in general the spatial structure G- driven SST pattern is antisymmetric to the G+ pattern in most of the regions the magnitude of SST correlations to the G- membership are insignificant except in the Eastern Subpolar North Atlantic. The low temperature anomaly in this region is concomitant with a maximum in the correlation of the G- membership and near surface wind stress curl (not shown here). One possible explanation of this pattern is that it favors local intensification of the Ekman pumping triggered by local strengthening of cyclonic wind vorticity.

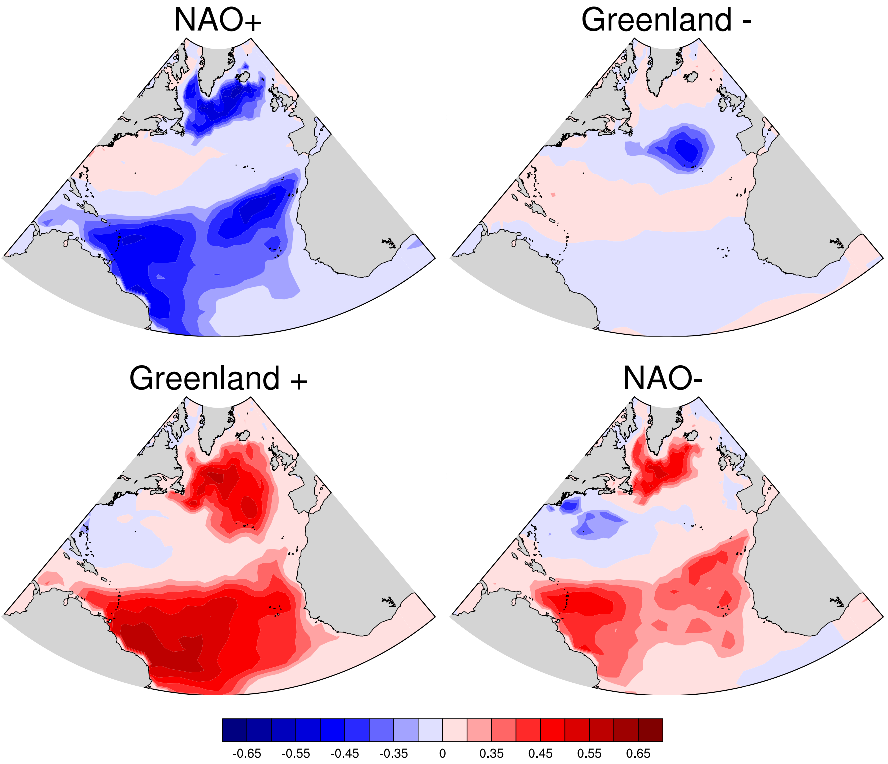


Figure 3.18: Map of correlations between monthly mean interannual H5 fuzzy memberships and SST values.

3.8 Conclusions

Four patterns of interannual and decadal variability for the North Atlantic are identified in this study. They display the spatial structure of two oscillatory patterns, referred to here as the NAO+/NAO and G+/G-. The NAO+/NAO- represents the regional manifestation of long term variability of the NAO/NAM mode. The G+/G- oscillatory pattern suggests the elements of a regional manifestation of the PNA mode.

The four interannual patterns ,NAO+/NAO- and G+/G-, have many similar elements with the dominant PCs of the H5 field, the PNA (Figure 3.12a), and NAO/NAM (Figure 3.12b), which have extrema in the North Pacific. Our analysis (not shown here), however, did not show significant correlation between the interannual patterns, NAO+/NAO- and G+/G-, and the SLP variability over the North Pacific. Previously, Wallace and Thompson (2002) have shown that the SLP variations in the Atlantic and Pacific centres of NAM are uncorrelated, due to the PNA related variability, whose pattern in the Pacific is inversely related to that of the NAM (see also Figure 3.12).

The same notations (NAO+/NAO-) are used here for the NAO+/NAO- weather regimes shown on Figure 3.7, cf., Cassou et al. (2004), and the NAO+/NAO- interannual patterns of Figure 3.14; cf., Esbensen (1984). They, however, refer to patterns that differ in terms of their spatial structure and temporal variability. The NAO+/NAO- weather regimes are associated with the local intensification of the Icelandic low and severe/mild winters in the Subpolar North Atlantic, as described by Walker and Bliss (1932). The interannual NAO+/NAO- patterns are defined by the frequency of occurrence of the four weather regimes and their amplitudes therein (Hurrell and Deser (2009)). Hence, the spatial structure of the interannual NAO+/NAO- patterns are elongated in zonal directions (see Figure 3.14) compared to the corresponding NAO+/NAO- weather regimes; as expected from previous studies; cf., Esbensen (1984). In this study our focus is on the long term interannual and decadal atmospheric shifts represented by these patterns.

Figure 3.17 associates each interranual pattern with specific distribution of the weather regimes. The NAO+ interranual regime favors weather NAO+ and GS weather regimes. In the late 1980s-early 1990s when the NAO index was predominantly positive, the NAO+ interannual pattern was dominant (see Figure 3.15), and the NAO+ and SG weather regimes were the most frequent over the North Atlantic (see Figure 3.9). Both weather regimes favor intensification of the pressure low over the Greenland and Iceland regions, and colder than normal winters over the Subpolar North Atlantic. The NAO- interannual pattern is associated with NAO- and AR+ weather patterns (see Figure 3.17). In the 1960s when the NAO index was negative, the NAO-interannual was dominating (see Figure 3.15), and the NAO- and AR+ weather regimes were the most frequent over the North Atlantic (see Figure 3.9). Both weather regimes favor development of blocking structures over Greenland, and deviation of the storm tracks from their climatological positions.

The G+/G- patterns dominate in the time before the 1960s and after the early 1990s. G+/G- are associated to more evenly distributed weather regimes (see Figure 3.17). Our results suggest that, in general, the G+ related atmospheric forcing is warmer than normal and winter SST is higher than average in the Subpolar North Atlantic (see Section 3.7). The G- pattern temporal variability correlates with a cold SST anomaly in the subpolar ocean, with a maximum in the Eastern North Atlantic (see Figure 3.18). There are strong indications; cf., Robson et al. (2012) and Marsh et al. (2008), that the warming of the Subpolar Ocean in the past two decades was triggered by intensified AMOC. The frequent occurrence of the G+ in that period (see Fig. 3.16) may have additionally contributed to this warming.

There was a significant shift of the NAO index in the 1970s-1980s (see Figures 3.3) from strongly negative phase in the 1960 to high positive the late 1980s (Hurrell and Van Loon, 1997; Dickson et al., 2000). Figure 3.19 shows the impact of this shift on the probability distribution of the four interannual regimes, calculated for two separate time periods. Figure 3.19a shows the distribution for the period from 1952 to 1971 when NAO index was negative. Figure 3.19b shows the distribution for the period of positive high NAO index from 1983 to 2008. These figures show a decadal shift similar to the low frequency shifts described in the example of the Lorenz

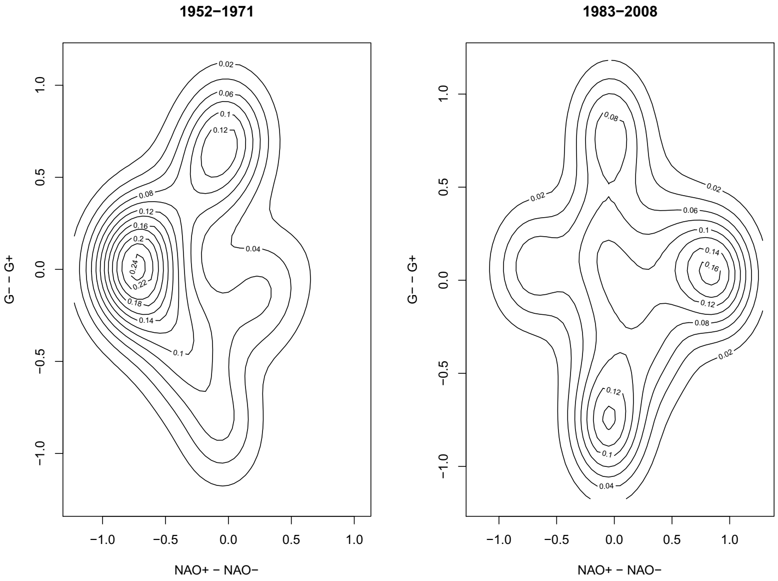


Figure 3.19: Contours showing percent probability (derived from kernel density estimates) of the interannual regime difference indexes for the time period of (a) 1952-1972 (left) and (b) 1983-2008 (right)

system (see section 3.6.1). While all of the regimes occur during the two periods, in the first period years with strong NAO+ are less frequent than for the other three patterns. Correspondingly, the NAO- interannual pattern has a low frequency of occurrence in the second period. Palmer (1999) suggested that similar shifts in the probability distribution of the regimes of a nonlinear climatic dynamical system can be triggered by weak and persistent forcing. One possible mechanism of the shift observed in Northern Hemisphere atmosphere in the 1980s is given by Shindell et al. (1999). These authors demonstrated that the atmospheric circulation and mean-eddy interaction from stratosphere to the ground can be sensitive the changes in the external radiative forcing. This study also demonstrated that the impact of Northern Hemisphere atmospheric response to the greenhouse gases is “manifested by a gradual reduction in high-latitude sea-level pressure, and an increase in mid-latitude sea-level pressure associated with one phase of the Arctic Oscillation excitation of the positive phase of the NAO”. Whether the expected increase in greenhouse gas forcing in the mid to late 21st century will result in increased occurrence of NAO+ related activity is still an open question. Many, but not all climate model studies suggest this is the case (Gillett et al., 2003), although there is debate as to how well such models address the range of processes that have been postulated to affect regime behaviour (Shindell et al., 1999).

3.9 Bibliography

Aagaard, K., Carmack, E., 1989. The role of sea ice and other fresh water in the arctic circulation. Journal of Geophysical Research: Oceans (1978–2012) 94 (C10), 14485–14498.

Baur, F., Hess, P., Nagel, H., 1944. Kalender der grosswetterlagen europas 1881-1939. Bad Homburg 500.

Berrisford, P., Hoskins, B. J., Tyrlis, E., Aug. 2007. Blocking and rossby wave breaking on the dynamical tropopause in the southern hemisphere. Journal of the Atmospheric Sciences 64 (8), 2881–2898.

Bezdek, J. C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, New York.

Bjerknes, J., 1964. Atlantic air-sea interaction. Advances in geophysics 10 (1), 82.

Boe, J., Terray, L., Habets, F., Martin, E., 2006. A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling. J. Geophys. Res 111, D23106.

Cassou, C., Sep. 2008. Intraseasonal interaction between the Madden–Julian oscillation and the north atlantic oscillation. Nature 455 (7212), 523–527.

Cassou, C., Terray, L., Hurrell, J. W., Deser, C., 2004. North atlantic winter climate regimes: Spatial asymmetry, stationarity with time, and oceanic forcing. Journal of Climate 17, 1055–1068.

Casty, C., Handorf, D., Raible, C. C., Gonzalez-Rouco, J. F., Weisheimer, A., Xoplaki, E., Luterbacher, J., Dethloff, K., Wanner, H., Mar. 2005. Recurrent climate winter regimes in reconstructed and modelled 500 hpa geopotential height fields over the north atlantic/european sector 1659–1990. Climate Dynamics 24 (7-8), 809–822.

Cattiaux, J., Vautard, R., Cassou, C., Yiou, P., Masson-Delmotte, V., Codron, F., Oct. 2010. Winter 2010 in europe: A cold extreme in a warming climate. Geophysical Research Letters 37 (20).

Cheng, X., Wallace, J., 1991. Cluster analysis of the northern hemisphere 500-hPa height field: Spatial patterns. Journal of the Atmospheric Sciences 50 (16), 2674–2696.

Christiansen, B., May 2007. Atmospheric circulation regimes: Can cluster analysis provide the number? Journal of Climate 20, 2229–2250.

Corte-Real, J., Quian, B., Xu, H., 1999. Circulation patterns, daily precipitation in portugal and implications for climate change simulated by the second hadley centre GCM. Climate Dynamics 15 (12), 921–935.

Corti, S., Molteni, F., Palmer, T., 1999. Signature of recent climate change in frequencies of natural atmospheric circulation regimes. Nature 398.

Croci-Maspoli, M., Schwierz, C., Davies, H. C., 2007. Atmospheric blocking: space-time links to the nao and pna. Climate Dynamics 29 (7-8), 713–725.

Curry, R. G., McCartney, M. S., 2001. Ocean gyre circulation changes associated with the north atlantic oscillation. Journal of Physical Oceanography 31 (12), 3374–3400.

Delworth, T., Manabe, S., Stouffer, R. J., 1993. Interdecadal variations of the thermohaline circulation in a coupled ocean-atmosphere model. Journal of Climate 6 (11), 1993–2011.

Delworth, T. L., Greatbatch, R. J., 2000. Multidecadal thermohaline circulation variability driven by atmospheric surface flux forcing. Journal of Climate 13 (9), 1481–1495.

Dennett, D. C., 1991. Real patterns. The Journal of Philosophy 88 (1), 27–51.

Deser, C., 2000. On the teleconnectivity of the “arctic oscillation”. Geophysical Research Letters 27 (6), 779–782.

Dickson, B., Yashayaev, I., Meincke, J., Turrell, B., Dye, S., Holfort, J., 2002. Rapid freshening of the deep north atlantic ocean over the past four decades. Nature 416 (6883), 832–837.

Dickson, R., Osborn, T., Hurrell, J., Meincke, J., Blindheim, J., Adlandsvik, B., Vinje, T., Alekseev, G., Maslowski, W., 2000. The arctic ocean response to the north atlantic oscillation. Journal of Climate 13 (15), 2671–2696.

Dickson, R. R., Meincke, J., Malmberg, S.-A., Lee, A. J., 1988. The “great salinity anomaly” in the northern north atlantic 1968–1982. Progress in Oceanography 20 (2), 103–151.

Dommenget, D., Latif, M., 2002. Analysis of observed and simulated sst spectra in the midlatitudes. Climate dynamics 19 (3-4), 277–288.

Duchon, C. E., 1979. Lanczos filtering in one and two dimensions. Journal of Applied Meteorology 18, 1016–1022.

Ebisuzaki, W., 1997. A method to estimate the statistical significance of a correlation when the data are serially correlated.

Esbensen, S. K., 1984. A comparison of intermonthly and interannual teleconnections in the 700 mb geopotential height field during the northern hemisphere winter. Monthly weather review 112 (10), 2016–2032.

Feldstein, S. B., 2007. The dynamics of the north atlantic oscillation during the summer season. Quarterly Journal of the Royal Meteorological Society.

Feliks, Y., Ghil, M., Robertson, A. W., Aug. 2010. Oscillatory climate modes in the eastern mediterranean and their synchronization with the north atlantic oscillation. Journal of Climate 23 (15), 4060–4079.

Frankignoul, C., Hasselmann, K., 1977. Stochastic climate models, part II application to sea-surface temperature anomalies and thermocline variability. Tellus 29 (4), 289–305.

Gillett, N. P., Graf, H. F., Osborn, T. J., 2003. Climate change and the north atlantic oscillation. In: Hurrell, J. W., Kushnir, Y., Ottersen, G., Visbeck, M. (Eds.), Geophysical Monograph Series. Vol. 134. American Geophysical Union, Washington, D. C., pp. 193–209.

Greatbatch, R. J., 2000. The north atlantic oscillation. Stochastic Environmental Research and Risk Assessment 14 (4-5), 213–242.

Hakkinen, S., Rhines, P. B., 2004. Decline of subpolar north atlantic circulation during the 1990s. Science 304 (5670), 555–559.

Hakkinen, S., Rhines, P. B., Worthen, D. L., Nov. 2011. Atmospheric blocking and atlantic multidecadal ocean variability. Science 334 (6056), 655–659.

Hale, J., Kocak, H., 1991. Dynamics and Bifurcations. Springer-Verlag, New York.

Hasselmann, K., 1976. Stochastic climate models part i. theory. Tellus 28 (6), 473–485.

Hatun, H., Sando, A. B., Drange, H., Hansen, B., Valdimarsson, H., 2005. Influence of the atlantic subpolar gyre on the thermohaline circulation. Science 309 (5742), 1841–1844.

Hess, P., Brezowsky, H., 1952. Katalog der grosswetterlagen Europas. Deutscher Wetterdienst in d. US-Zone.

Horenko, I., Apr. 2010. On clustering of non-stationary meteorological time series. Dynamics of Atmospheres and Oceans 49 (2-3), 164–187.

Hoskins, B. J., Hodges, K. I., 2010. New perspectives on the northern hemisphere winter storm tracks.

Hurrell, J. W., 1995. Decadal trends in the north atlantic oscillation, regional temperatures and precipitation. Science 269, 67–69.

Hurrell, J. W., Deser, C., 2009. North atlantic climate variability: the role of the north atlantic oscillation. Journal of Marine Systems 78 (1), 28–41.

Hurrell, J. W., Kushnir, Y., Ottersen, G., Visbeck, M., 2003. An overview of the north atlantic oscillation. GEOPHYSICAL MONOGRAPH-AMERICAN GEOPHYSICAL UNION 134, 1–36.

Hurrell, J. W., Van Loon, H., 1997. Decadal variations in climate associated with the north atlantic oscillation. Climatic change 36 (3-4), 301–326.

Kalnay, E., 2002. Atmospheric modeling, data assimilation and predictability. Cambridge university press.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gadin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., Joseph, D., 1996. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteoroligcal Society 77, 437–470.

Kannan, S., Ghosh, S., Jul. 2010. Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output. Stochastic Environmental Research and Risk Assessment 25 (4), 457–474.

Kaufman, L., Rousseeuw, P. J., 1990. Finding Groups in Data; An Introduction to Cluster Analysis. John Wiley & Sons.

Klawonn, F., Hoppner, F., 2003. What is fuzzy about fuzzy clustering? understanding and improving the concept of the fuzzifier. In: Advances in Intelligent Data Analysis. V. Springer, Berlin, pp. 254–264.

Kosko, B., 1990. Fuzziness vs. probability 17 (1), 211–240.

Kushnir, Y., Robinson, W., Bladé, I., Hall, N., Peng, S., Sutton, R., 2002. Atmospheric gcm response to extratropical sst anomalies: Synthesis and evaluation. Journal of Climate 15 (16), 2233–2256.

Kwok, R., Rothrock, D. A., 1999. Variability of fram strait ice flux and north atlantic oscillation. Journal of Geophysical Research: Oceans (1978–2012) 104 (C3), 5177–5189.

Lee, S., Feldstein, S., 1996. Mechanism of zonal index evolution in a two-layer model. Journal of the atmospheric sciences 53 (15), 2232–2246.

Levitus, S., Antonov, J., Boyer, T., Locarnini, R., Garcia, H., Mishonov, A., 2009. Global ocean heat content 1955–2008 in light of recently revealed instrumentation problems. Geophysical Research Letters 36 (7).

Lohmann, K., Drange, H., Bentsen, M., 2009. A possible mechanism for the strong weakening of the north atlantic subpolar gyre in the mid-1990s. Geophysical Research Letters 36 (15).

Lorenz, E., 1963. Deterministic nonperiodic flow. Journal of the atmospheric sciences 20, 130–141.

Lorenz, E. N., Aug. 2006. Regimes in simple systems. Journal of the Atmospheric Sciences 63 (8), 2056–2073.

Lozier, M. S., Roussenov, V., Reed, M. S., Williams, R. G., 2010. Opposing decadal changes for the north atlantic meridional overturning circulation. Nature Geoscience 3 (10), 728–734.

Lundrigan, S., Demirov, E., 2012. Long-term variability of volume and heat transport in the nordic seas: A model study. Atmosphere-Ocean 50 (2), 156–168.

Luo, D., Diao, Y., Feldstein, S. B., Mar. 2011. The variability of the atlantic storm track and the north atlantic oscillation: A link between intraseasonal and interannual variability. Journal of the Atmospheric Sciences 68 (3), 577–601.

M. P. Baldwin, L. J. Gray, T. J. Dunkerton, K. Hamilton, P. H. Haynes, W. J. Randel, J. R. Holton, M. J. Alexander, I. Hirote, T. Horinouchi, D. B. A. Jones, J. S. Kinnersley, C. Marquardt, K. Sato, M. Takahashi, 2001. The quasi-biennial oscillation. Reviews of Geophysics 39 (2), 179—229.

MacKay, D., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., 2005. Cluster analysis basics and extensions.

Marsh, R., Josey, S. A., de Cuevas, B. A., Redbourn, L. J., Quartly, G. D., 2008. Mechanisms for recent warming of the north atlantic: Insights gained with an eddy-permitting model. Journal of Geophysical Research 113 (C4), C04031.

Molteni, F., Kucharski, F., Corti, S., 2006. On the predictability of flow-regime properties on interannual to interdecadal timescales. In: Predictability of Weather and Climate. Cambridge University Press.

Monahan, A. H., 2002. Stabilization of climate regimes by noise in a simple model of the thermohaline circulation. Journal of physical oceanography 32 (7), 2072–2085.

Monahan, A. H., Alexander, J., Weaver, A. L., 2010. Stochastic models of the meridional overturning circulation: time scales and patterns of variability. In: Stochastic Physics and Climate Modelling, 1st Edition. Cambridge University Press, pp. 266–286.

Monahan, A. H., Pandolfo, L., Fyfe, J. C., 2001. The preferred structure of variability of the northern hemisphere atmospheric circulation. Geophysical Research Letters 28 (6), 1019–1022.

Neal, R., 1991. Bayesian mixture modeling by monte carlo simmulation. Tech. Rep. CRG-TR-91-2, University of Toronto.

OrtizBevia, M. J., SanchezGomez, E., Alvarez-Garcia, F. J., Mar. 2011. North atlantic atmospheric regimes and winter extremes in the iberian peninsula. Natural Hazards and Earth System Science 11 (3), 971–980.

Palmer, T. N., 1999. A nonlinear dynamical perspective on climate prediction. Journal of Climate 12 (2), 575–591.

Panetta, R., Held, I., 1988. Baroclinic eddy fluxes in a one-dimensional model of quasi-geostrophic turbulence. Journal of the atmospheric sciences 45 (22), 3354–3365.

R Development Core Team, 2011. R: A language and environment for statistical computing. Manual, R Foundation for Statistical Computing, Vienna, Austria.

Reusch, D., Alley, R., Hewitson, B., 2007. North atlantic climate variability from a self-organizing map perspective. Journal of Geophysical Research 112, 1–20.

Robson, J., Sutton, R., Lohmann, K., Smith, D., Palmer, M. D., 2012. Causes of the rapid warming of the north atlantic ocean in the mid-1990s. Journal of Climate 25 (12), 4116–4134.

Rodwell, M. J., 2003. On the predictability of north atlantic climate. Geophysical Monograph Series 134, 173–192.

Shindell, D. T., Miller, R. L., Schmidt, G. A., Pandolfo, L., 1999. Simulation of recent northern winter climate trends by greenhouse-gas forcing. Nature 399 (6735), 452–455.

Simmons, A. J., Hoskins, B. J., 1978. The life cycles of some nonlinear baroclinic waves. Journal of the Atmospheric Sciences 35 (3), 414–432.

Solomon, S., Qin, D., Manning, M., Alley, R., Berntsen, T., Bindoff, N., Chen, Z., Chidthaisong, A., Gregory, J., Hegerl, G., et al., 2007. Climate change 2007: The physical science basis, contribution of working group 1 to the fourth assessment report of the intergovernmental panel on climate change.

Sparrow, C., 1982. The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors. Vol. 64. Berlin-Heidelberg-New York.

Spekat, A., Heller-Schulze, B., Lutz, M., 1983. Uber grosswetter und markov-ketten. Meteorologische Rundschau 36 (6), 243–248.

Stephenson, D. B., Hannachi, A., O'Neill, A., Jan. 2004. On the existence of multiple climate regimes. Quarterly Journal of the Royal Meteorological Society 130 (597), 583–605.

Sura, P., Newman, M., Penland, C., Sardeshmukh, P., 2005. Multiplicative noise and non-gaussianity: A paradigm for atmospheric regimes? Journal of the atmospheric sciences 62 (5), 1391–1409.

Teng, Q., Monahan, A. H., Fyfe, J. C., 2004. Effects of time averaging on climate regimes. Geophysical Research Letters 31 (22).

Terray, L., Demory, M. E., Deque, M., de Coetlogon, G., Maisonnave, E., 2004. Simulation of late-twenty-first-century changes in wintertime atmospheric circulation over europe due to anthropogenic causes. Journal of climate 17 (24), 4630–4635.

Thompson, D. W., Lee, S., Baldwin, M. P., 2003. Atmospheric processes governing the northern hemisphere annular mode/north atlantic oscillation. Geophysical Monograph - American Geophysical Union 134, 81–112.

Thompson, D. W., Wallace, J. M., 2000. Annular modes in the extratropical circulation. part i: month-to-month variability*. Journal of Climate 13 (5), 1000–1016.

Tibaldi, S., Molteni, F., 1990. On the operational predictability of blocking. Tellus A 42 (3), 343–365.
URL http://dx.doi.org/10.1034/j.1600-0870.1990.t01-2-00003.x

Vage, K., Pickart, R. S., Thierry, V., Reverdin, G., Lee, C. M., Petrie, B., Agnew, T. A., Wong, A., Ribergaard, M. H., 2008. Surprising return of deep convection to the subpolar north atlantic ocean in winter 2007–2008. Nature Geoscience 2 (1), 67–72.

Vallis, G. K., 2009. Mechanisms of climate variability from years to decades.

Vallis, G. K., Gerber, E. P., Kushner, P. J., Cash, B. A., 2004. A mechanism and simple dynamical model of the north atlantic oscillation and annular modes. Journal of the atmospheric sciences 61 (3), 264–280.

Visbeck, M., Chassignet, E. P., Curry, R. G., Delworth, T. L., Dickson, R. R., Krahmann, G., 2003. The ocean's response to north atlantic oscillation variability. Geophysical Monograph - American Geophysical Union 134, 113–146.

von Storch, H., Zwiers, F. W., 1984. Statistical Analysis in Climate Research. Cambridge University Press, Cambridge.

Walker, G. T., Bliss, E., 1932. World weather. V. Mem. Roy. Meteor. Soc. 4, 53–84.

Walker, S. G. T., 1922. Correlation in Seasonal Variations of Weather, VIII: A Preliminary Study of World-weather. Meteorological Office.

Walker, S. G. T., 1924. Correlation in Seasonal Variations of Weather, IX: Further Study of World-weather. Meteorological Office.

Wallace, J. M., Gutzler, D. S., 1981. Teleconnections in the geopotential height field during the northern hemisphere winter. Monthly Weather Review 109, 784–812.

Wallace, J. M., Thompson, D. W., 2002. The pacific center of action of the northern hemisphere annular mode: Real or artifact? Journal of Climate 15 (14), 1987–1991.

Wallace, J. M., Zhang, Y., Bajuk, L., 1996. Interpretation of interdecadal trends in northern hemisphere surface air temperature. Journal of Climate 9 (2), 249–259.

Willet, H. C., Sanders, F., 1952. Descriptive Meteorology, 3rd Edition. Academic Press Inc.

Woollings, T., Hoskins, B., Blackburn, M., Berrisford, P., 2008. A new rossby wave-breaking interpretation of the north atlantic oscillation. Journal of the Atmospheric Sciences 65 (2), 609–626.

Yashayaev, I., 2007. Hydrographic changes in the labrador sea, 1960–2005. Progress in Oceanography 73 (3), 242–276.

Yashayaev, I., Clarke, A., 2006. Recent warming of the labrador sea. Gulf Region/Région du Golfe Joël Chassé1, Doug Swain6 Marine Environmental Data Service/Service des données sur le milieu marin Bob Keeley3, Mathieu Ouellet3, 7, 8, Don Spear3, 12.

Yashayaev, I., Loder, J. W., 2009. Enhanced production of labrador sea water in 2008. Geophysical Research Letters 36 (1).

Yiou, P., 2004. Extreme climatic events and weather regimes over the north atlantic: When and where? Geophysical Research Letters 31 (7).

Zhu, J., Demirov, E., Mar. 2011. On the mechanism of interannual variability of the irminger water in the labrador sea. Journal of Geophysical Research 116 (C3).

Zwiers, F. W., 1990. The effect of serial correlation on statistical inferences made with resampling procedures. Journal of Climate 3, 1452–1461.

Connecting Text

One uncertainty not addressed in the Chapters 2 and 3 is residual variability, as defined in Chapter 1. Chapter 3 examines mean states at different spectral intervals, not day to day weather events themselves. The EMIC used in Chapter 2 does simulate daily weather (at a very coarse resolution), however these features are not examined, as the low resolution of the simulator prevents it from accurately reproducing the full structure and variability of these events. The following article develops an approach to describing residual variability in the form of a stochastic weather generator for the Subpolar North Atlantic, a sub-region of that outlined in Chapter 3. For this smaller region the sub-seasonal patterns of Chapter 3 are not defined, and as well do not approximate day to day circulation events. A more detailed analysis using SOMs, as discussed in Chapter 1, is used to describe activity within the area of interest. The weather generator is conditioned on the interannual patterns defined in Chapter 3. That these patterns can be used to define longer term shifts in the distribution of weather events was brought to attention in the analysis of the sub-seasonal regimes of Chapter 3. Weather generator construction is an area of ongoing investigation, with less development in the area of regional, rather than site specific, models. The design questions addressed in this article follow objectives (6)-(8). This article has appeared as Hauser and Demirov (2013) in the journal *Stochastic Environmental Research and Risk Assessment*. Additional discussion can be found in Section C.1.

A long term aim of the work presented in Chapter 4 is to examine ways of providing detailed realistic atmospheric forcing for ocean models that is not constrained by the length of historical records. At present it is not known how responsive ocean variability is to subtleties in atmospheric forcing. Studies conducted using forcing typical to the long-term variation identified in Chapter 3 would help clarify this, particularly for neutral periods in long term evolution of the NAO.

There are two inconsistencies in presentation between the preceding and the following chapter. In Chapter 3 the term “weather regimes” was used to refer to the sub-seasonal modes. In Chapter 4 the term refers to the interannual patterns of Chapter 3, as these are the only regime type features discussed. Also, the interannual pattern “G-” from Chapter 3 is referred to as “Eastern Blocking” in Chapter 4 and the pattern “G+” is referred to as “Western Blocking” in Chapter 4. The change in naming convention is due to the different context and target audiences of each article. While “blocking” is visually informative about the structure of the features, blocking events are not interannual phenomena and so would confuse the presentation of Chapter 3.

Chapter 4

Development of a Stochastic Weather Generator for the Sub-polar North Atlantic

4.1 Abstract

The article presents an approach for creating a computationally efficient stochastic weather generator. In this work the method is tested by the stochastic simulation of sea level pressure over the sub-polar North Atlantic. The weather generator includes a hidden Markov model, which propagates regional circulation patterns identified by a self organising map analysis, conditioned on the state of large-scale interannaul weather regimes. The remaining residual effects are propagated by a regression model with added noise components. The regression step is performed by one of two methods, a linear model or artificial neural networks and the performance of these two methods is assessed and compared. The resulting simulations express the range of the major regional patterns of atmospheric variability and typical time scales. The long term aims of this work are to provide ensembles of atmospheric data for applied regional studies and to develop tools applicable in down-scaling large-scale ocean and atmospheric simulations.

4.2 Introduction

A stochastic weather generator (WG) produces synthetic time series of weather data based on the statistical characteristics of weather at that location. As such, WGs are not designed to forecast individual events; i.e., there is no expectation that the value of the generated variables for a given date/time will match those observed. Rather, they create time series of atmospheric variables with statistical characteristics that resemble those of observations (Jones et al., 2009). These are empirical models based on statistical relations rather than the equations of earth system dynamics (Benestad et al., 2008).

WGs are often used as a downscaling technique, as they can simulate values at scales below the resolution of most dynamical circulation models. Most commonly they are used to produce precipitation values for agricultural and hydrological models (Maraun et al., 2010). Such WGs can provide ensembles of long time series for use in uncertainty analysis and are often used in climate projections (Semenov and Barrow, 1997).

The primarily goal of this work is to develop a tool which is capable of producing synthetic atmospheric fields that have characteristics typical for the relevant trends of a specific period of time; using the long-term tendencies of the atmospheric circulation as inputs. The need for such a WG arises from recent studies of climate change for the North Atlantic Ocean. Through the late 1980s and early 1990s the sub-polar North Atlantic has shown a tendency towards cooling temperatures which have changed towards rapid warming since the 1990s (Marsh et al., 2008). The mechanism of this change has strong links with the North Atlantic Oscillation (NAO) interannual atmospheric variability (Bersch et al., 2007) and the Atlantic Meridional Overturning Circulation (AMOC) (Robson et al., 2012) and has received strong attention in recent publications; e.g., Marsh et al. (2008), Sarafanov et al. (2008), Lohmann et al. (2008) and Hakkinen et al. (2011). The NAO is understood to be indicative of trends in the west-east tracks of extra-tropical cyclones (Hurrell et al. (2003), Luo et al. (2011)). These storms are the major source of atmospheric flux anomalies relevant to ocean circulation, namely surface winds and precipitation, for the region. As such, shifts in the dominant pathways of these storms; whether tending towards Central Europe, Iceland, or being diverted northward towards the Labrador Sea, can have significant effects, especially given the highly local nature of the deep water formation which occurs within the region.

For the purposes of ocean modelling, Lohmann et al. (2008) and Zhu and Demirov (2011) created deterministic forcing series typical of positive and negative phase NAO atmospheric variability. The forcing was constructed as a sum of the monthly mean characteristics for the specific phase of NAO and the deviation from the monthly mean taken from a specific subjectively chosen year. The weakness of this method is that atmospheric weather conditions in years with persistent NAO phase are highly variable and change from year to year. The sign of the phase itself is also a rather qualitative characteristic and the results of calculating monthly averaged characteristics over years of persistent phase is sensitive to the specific years used. In this article we describe an alternative approach for providing series of synthetic atmospheric data that are representative of the statistical characteristics of the atmospheric interannual variability.

Previously a simple generator of atmospheric characteristics in ocean modelling applications was used by Tang et al. (2001). Their approach was based on so called Hybrid Coupled Models (HCM) which include two way coupling between a general circulation ocean model and an empirically derived atmospheric regression model. The HCM proved to be an efficient tool for representing the air-sea interaction in the equatorial area and its impact on coupled atmospheric and ocean dynamics. In the North Atlantic simulations, however, a different approach is required because the dominant atmospheric variability in this region does not owe its existence directly to the air-sea interaction (Valis, 2007).

The atmospheric variability in the midlatitudes of the North Atlantic is triggered by transient waves, meandering of the jet stream, and the instability and breaking of planetary waves (Thompson et al., 2003). The local processes that govern the mechanisms of circulation regimes such as the NAO are nonlinear and chaotic and show strong links to global planetary circulation variability. The main purpose of this work is to develop a stochastic tool; i.e, a WG, that will be able to represent properly the statistical characteristics of these processes. In this work we focus on developing a stochastic model capable of reproducing the intrinsic variability of dominant regional patterns in climate, unresolved processes, and the interactions between them.

Weather generators are typically designed to generate values for specific locations, e.g., Oelschlagel (1995), or multiple sites (Maraun et al., 2010). Whole field simulating weather generators are mostly developed for precipitation modelling over local regions. A comparison of common approaches is given by Ferraris et al. (2003). These methods are designed to mimic spatially discontinuous time series of small scale convective precipitation, rather than the smoothly evolving, large scale weather systems considered here. In this article we present a method for generating spatially continuous atmospheric fields that reflect dominant patterns of atmospheric variability at interannual, seasonal, and intra-seasonal scales.

Many weather generation methods use a two layer approach. First a general weather state is selected, which in turn defines the parameters used to generate the model output. A common example is a precipitation model which first selects between states of precipitation occurrence (or non-occurrence), with each state being associated with its own probability distribution from which rainfall amounts can be sampled; e.g., Furrer and Katz (2007), and Baigorria and Jones (2010). The method used in our work is also based on a multi-layer approach to represent the multi-scale dynamics of the atmosphere and the interaction between the scales in propagating the atmospheric fields. Here, we first model the progression of expected general states within individual interannual regimes. The status of these states then influences a residual model which simulates variability unaddressed by the initial classification. In this article we present the results from testing different methods of describing these components. Since these are statistical rather than physically based models, it is often indeterminable *a priori* what the most appropriate and/or effective approach will be. The importance of considering multiple approaches is often stated in the literature;

e.g., Hashmi et al. (2011).

This article outlines a WG designed for a limited area over the sub-polar North Atlantic. The initial experiments presented here provide a proof of concept for the method and compare two potential approaches to implementation. The article is organised as follows. The WG is described in detail in the following section. This is followed by a presentation of the results from simulations, along with some concluding comments on the results and required further developments.

4.3 Methods

4.3.1 Region and Data

The region of interest is the portion of the Sub-Polar North Atlantic from 45°N : 67°N and 4°W : 66°W, shown in Figure 4.1. This is an area of active atmosphere - ocean interaction which has a strong impact on the AMOC and global climate. Gulf Stream water of sub-tropical origin is carried north by the North Atlantic Current, a branch of which carries this warm and salty water to the Irminger Sea. This water mass is then cooled through interaction with the cold sub-polar atmosphere and mixing with the surrounding waters, as a result sinking to intermediate depths to form so called Irminger Water. The abrupt warming of the sub-polar ocean observed since the mid 1990s has been related in recent studies (see for review Zhu and Demirov (2011) and Robson et al. (2012)) to variations of the inflow of Irminger Waters in the sub-polar North Atlantic. These are ultimately attributed to the impact of variations in surface forcing for this region. Our current study is motivated by a desire to allow further investigation of this process.

The atmospheric data used in this work is taken from the NCEP reanalysis project (Kalnay et al., 1996) daily Sea Level Pressure (SLP) and 500 mb geopotential height

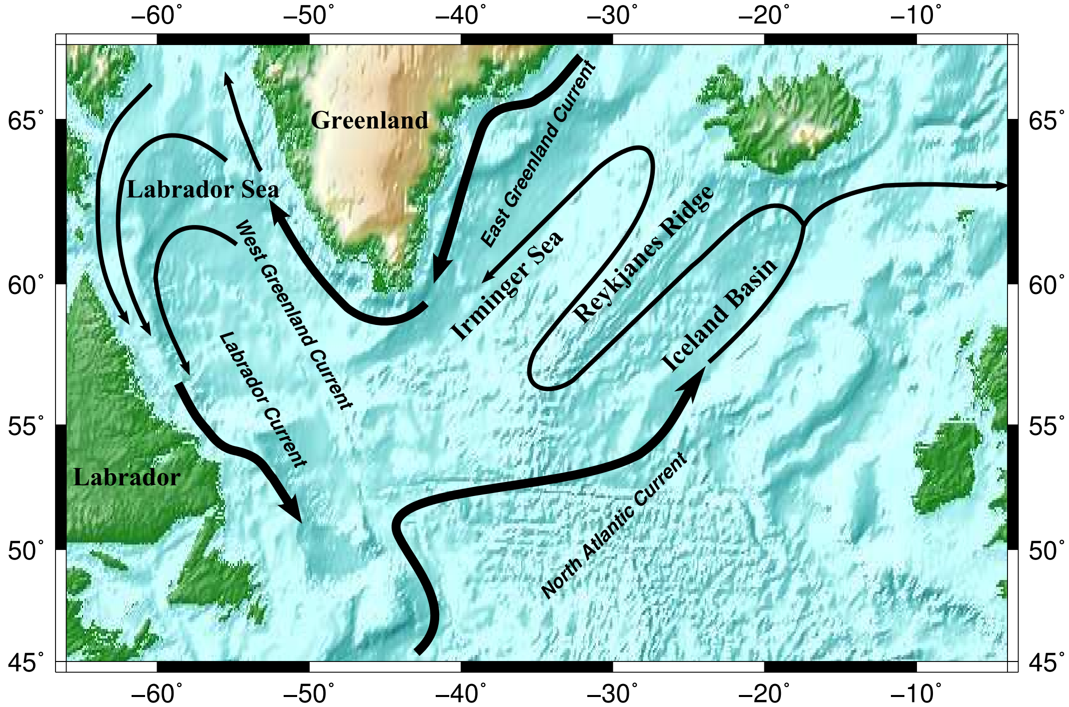


Figure 4.1: Map of the region of interest, displaying ocean currents that compose the North Atlantic Supolar Gyre.

fields (H5). The data is de-trended and high and low pass filtered respectively, to isolate the SLP intra-annual components and the interannual H5 components respectively, using a Lanczos filter; see Duchon (1979). This results in daily fields which express the comparative high and low frequency portions of the data set, respectively. To remove boundary filter effects the resulting data sets are trimmed to encompass the years 1951-2008. Currently simulations are only created for Northern Hemisphere winter (DJF). Future development will include repeating the process described below for each season.

For the experiments described here, the focus is on simulating intra-seasonal SLP anomaly fields. SLP is selected for these initial tests as the surface lows and highs it depicts are representative of factors composing and influencing the storm tracks discussed above. If these features can be represented then the method has the potential to be extended to related variables. This method can also potentially be used to downscale General Circulation Model (GCM) results. Current GCMs are typically able to simulate large (spatial and temporal) scale climate features; e.g., Severijns and Hazeleger (2010), Rust et al. (2010), and the uncertainties at these scales can be quantified to a certain degree; e.g., Sexton and Murphy (2011) and Hauser et al. (2011). Stochastic downscaling techniques can serve as a method to better estimate smaller scale variability often missing from such simulations (von Hardenberg et al., 2007; Minville et al., 2008).

4.3.2 The weather generator

Atmospheric circulation displays variability on multiple scales. On time scales longer than a single season the dynamics of a limited region, such as the one which is of interest for the present study, are influenced by the interactions at its boundaries. The most energetic variability in the extratropics is driven by synoptic eddies and weather systems, with time scales on the order of a few days. Variability with time scales between ten and one hundred days typically has lower amplitude than that seen at shorter time scales. While the dynamics of this variability are not well understood, they are considered to be primarily atmospherically driven as the scales of interannual variations in SST are significantly longer and not likely to affect strongly the atmosphere intra-annual variations (Vallis et al., 2004). In the present study we separate the variability of the atmospheric field in three components - (a) large-scale for the region at interannual time scales, (b) seasonal, and (c) regional limited area scale intra-seasonal. Component (c) is what we seek to simulate with the presented weather generator. The daily signature of the large-scale component is the input for the model and separate models are considered for different seasons.

As mentioned in the introduction, the weather generator is based on a multilevel approach. The primary level is essentially a Hidden Markov Model (HMM); see Cappe (2005) for a thorough overview of this technique. The HMM makes discrete stochastic transitions between predetermined states based on prescribed transition probabilities. These transition probabilities are determined by a "hidden state". Such a model, determined by lagged internal processes affected by external forcing matches our concept of the examined system. For the model presented here the "hidden states" are determined by classification of the large-scale interannual trends over the North Atlantic as described in section 4.3.3. Studies performed in other regions have argued that interannual variability, though often comparatively small in amplitude, can be an important modulator of behaviour on shorter time scales; e.g., Grimm (2011). For this model there are four possible "hidden states" which in the following are referred to as weather regimes. Each weather regime has an associated set of discrete states and transition probabilities for the intra-seasonal behaviours of the sub-polar region described above; i.e., the fields we are looking to simulate. These states and transitions are determined by a Self Organising Map (SOM) analysis of the region (performed separately for time-periods dominated by different weather regimes), as described in section 4.3.4.

The use of a limited number of discrete states in the HMM gives a limited representation of field evolution. More nuanced effects are described using Principal Component Analysis (PCA) to represent the leading variability modes of the residuals between the full intra-seasonal state and that described by the HMM. The strength of these modes are propagated by lagged regressions which are specific to the large-scale weather regimes and conditional on the current state of the HMM. The lagged regression methods tested here are similar to those used by Kravtsov et al. (2005) and Aguilar-Martinez and Hsieh (2009). Additional variability not captured by the HMM and the residual model is represented as spatially correlated noise, the parameters of which are also dependent on the higher-level states. Samples from these distributions are used to perturb the estimated system state to account for unrepresented processes.

Two methods for performing the lagged regression step described above are tested. The first is based on the empirical model reduction approach as described by Kravtsov et al. (2010). For this method a polynomial regression is fit using training data, along with additional modelling of residuals. The method also uses stochastic terms to account for unresolved processes, as also seen in Guo et al. (2012). The second approach creates a non-linear regression by using Bayesian methods to generate an ensemble of Artificial Neural Networks (ANNs) with additional stochastic noise models that account for unexplained variability. ANN based regressions have been shown to improve on linear approaches in similar contexts; e.g., Tang and Hsieh (2002) and Chowdhury et al. (2010). However, they are much more involved to implement and provide less transparent results (Heckerling et al., 2003). As well, it is not clear *a priori* if the high sensitivity of the ANNs to non-linearities will offer an advantage when modelling a large-scale flow determined variable, which should respond well to the empirical reduction technique.

Stochastic simulations are created only for possible trajectories of the intra-seasonal component of the described system, given indicators of the long term trend of interannual variability, which are defined from the original data. Should we wish to reconstruct the full vector of the unfiltered system state $\boldsymbol{x}_{\mathrm{full-state}}$ for any time t, the interannual $\boldsymbol{x}_{\mathrm{interannual}}$ and seasonal $\boldsymbol{x}_{\mathrm{seasonal}}$ components can be taken as given, prescribed by the observations or GCM which provide the model predictors, such that,

$$\boldsymbol{x}^{(t)}_{\mathrm{full-state}} = \underbrace{\boldsymbol{x}^{(t)}_{\mathrm{interannual}}}_{\mathrm{GCM}} + \underbrace{\boldsymbol{x}^{(t)}_{\mathrm{seasonal}}}_{\mathrm{GCM}} + \underbrace{\boldsymbol{x}^{(t)}_{\mathrm{intra-seasonal}}}_{\text{"generated"}} . \tag{4.1}$$

These intra-seasonal simulations are designed to reflect the daily signature of the large scale, lower frequency components of the system which are classified into interannual regimes. These are defined based on clustering of H5 interannual fields over the North Atlantic (see section 4.3.3). This region and variable are considered to be representative of the dominant processes affecting weather patterns for the sub-polar North Atlantic (Hakkinen et al., 2011). Separate stochastic simulations of the intraseasonal data are created using data specific to the dates considered as part of a given, season specific, interannaul weather regime.

4.3.3 Weather Regimes (Large-scale Interannual Modes) - Fuzzy Clustering

The large-scale atmospheric variability of the North Atlantic is dominated by the NAO (Hurrell et al., 2003). The NAO is defined as an oscillatory spatial pattern which appears in multiple layers of the atmosphere (Wallace and Gutzler, 1981). It is a robust pattern which can be easily identified using linear methods such as correlation maps (Wallace and Gutzler, 1981) or PCA techniques (Feldstein, 2000). By necessity these linear methods assume a spatial symmetry of distribution in the phase space between opposing phases of the oscillation (Cassou et al., 2004). An alternative nonlinear approach for describing the dominant patterns of atmospheric variability is based on the concept of weather or climate regimes, defined as peaks in the probability density of the climate phase space (Palmer, 1999). Long term climate changes are then defined as a shifts in the amplitude of these peaks due to changes in frequency of regime appearances (Corti et al., 1999). For example, Corti et al. (1999) showed that recent temperature trends over the Northern Hemisphere could be described by an increased occurrence of the so called cold-ocean-warm-land regime. One method for defining regimes which can take into account spatial asymmetry and temporal shifts in dominant modes is cluster analysis (Cassou et al., 2004).

Cluster analysis is a classification method which divides a data set into a predefined number of subsets (clusters) of similar elements. The cluster centroids describe the characteristic pattern common to their elements. Often these are defined as the mean of all elements assigned to the same cluster. The goal is to subdivide the data in a way that maximises the distance/difference between centres, while also maximising the similarity of the members within individual clusters. The method has been used to describe large scale atmospheric circulation regimes over hemispheric domains; e.g., Cheng and Wallace (1991), Corti et al. (1999), and Molteni et al. (2006), as well as the North Atlantic region; e.g., Cassou et al. (2004), Yiou (2004)and Cassou (2008). Once defined these climate regimes can be statistically linked to mesoscale local phenomena; e.g., Cattiaux et al. (2010) and OrtizBeviá et al. (2011). Fuzzy clustering provides an alternative to the more typical binary clustering, so called as the membership to a cluster can be considered to be equal to zero or one. Rather, fuzzy clustering defines a continuous range of membership to a cluster on the range [0,1], with the condition that memberships across all clusters sum to one. These requirements result in this method sometimes being referred to as “probabilistic clustering”, with the interpretation that the membership degree represents a (subjective) probability that one would assign a given data point to a given cluster (Bezdek, 1981).

The weather regimes used here are categorised by using fuzzy clustering analysis on the interannual H5 fields for the region from 25°N : 75°N (Bahamas and Canary Islands to Baffin Bay and Barents Sea) and from -105°E : 45°E (Gulf of Mexico and Hudson Bay to edge of Scandinavia and Mediterranean), so as to capture the extent of the relevant weather modes (Hoskins and Hodges, 2010). Taking predictors from beyond the simulated region is common practice for weather generation; e.g., Guo et al. (2012). For this investigation, membership degrees are determined using the “FANNY” algorithm (Kaufman and Rousseeuw, 1990) as implemented in by Maechler et al. (2005) for the software package R. For this application this amounts to a variant of the "fuzzy c-means" algorithm (Bezdek, 1981) using Euclidean distance, rather than the traditional squared Euclidean distance, since the former method has less outlier sensitivity, and better represents non-spherical clusters (Kaufman and Rousseeuw, 1990). The results for the DJF season are shown in Figure 4.2.

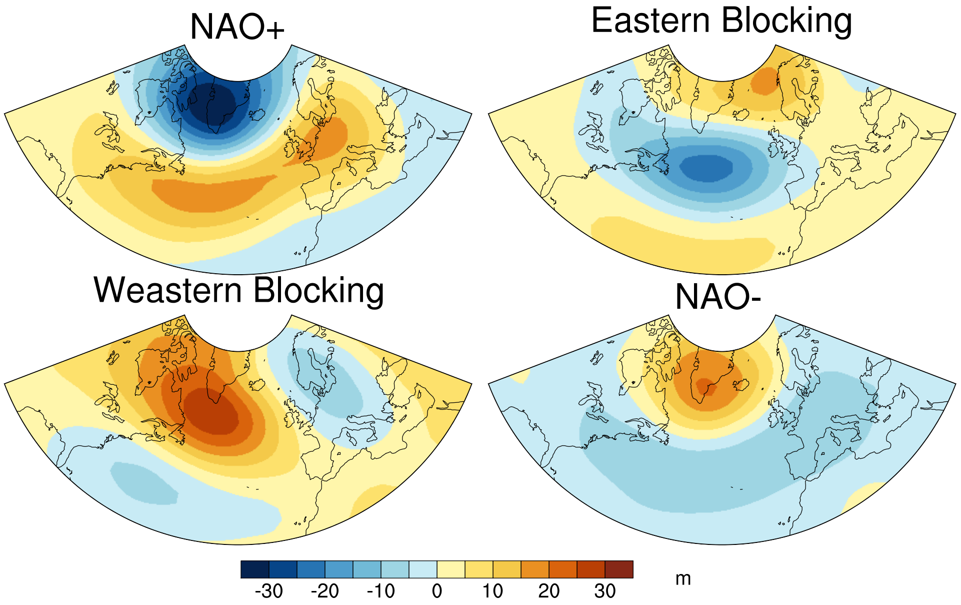


Figure 4.2: Cluster centres produced using the PC time series of the 10 leading eigenvectors for interannual 500 mb geopotential height.

Initial investigations of the phase space using kernel density estimates; following Molteni et al. (2006), support the multi-modal hypothesis which justifies the use of the clustering approach. This analysis also suggests four as the operative number of centres. These results are supported by an analysis using Bayesain Infinite Gaussian Mixture Models as developed by Neal (1991). Note that the use of interannual rather than sub-seasonal data means the results are not directly comparable to the four regimes commonly identified with the North Atlantic; cf., Cassou (2008), which more closely correspond to observable events. The first and fourth clusters represent the positive and negative phase of the NAO. The maximums and minimums of the meridional dipoles of these two clusters are zonally elongated. The second mode depicts trends towards a north-eastern high pressure feature by Scandinavia and a low in the Newfoundland Basin. The third mode displayed depicts a western blocking feature. The occurrence of the first and fourth regimes over the historical record correspond with the behaviour of the classical NAO index and the published results of associated sub-seasonal data classifications. Years dominated by the second mode show a preferential, but not exclusive, occurrence of the so call Scandinavian Blocking feature (Cassou et al., 2004). Years dominated by the third mode show an increase in the so called Atlantic Ridge (Cassou et al., 2004) anti-cyclone events compared to other regimes, but there is no dominant sub-seasonal scale feature for this regime.

4.3.4 Regional Intra-Seasonal Modes - Self Organising Maps

A time decorrelation analysis of principal components of the intra-seasonal field (not shown here) for the limited area shown on Fig 4.1 suggests that there are dominant regional modes of variability on intra-seasonal time scales. The spatial structure of these modes is described by classifying the intra-seasonal field using Self Organizing Maps (SOMs).

Self Organising Maps[1] are a form of non-linear regression (machine learning). Where many clustering techniques (as above) attempt to define the most distinct groups within the data set, SOMs create sets of neighbouring states to highlight more subtle differences and transitions in behaviour. An overview of the application of the method to the field of synoptic climatology is given by Hewitson and Crane (2002). The analysis is performed in the following stages as described by Kohonen et al. (1996), Cassano et al. (2005), and Reusch et al. (2007). First a preselected number of “maps”; i.e., vectors whose length corresponds to the number of data points per time step of the

[1] The software used here is freely available from: www.cis.hut.fi/research/som_pak/.

fields being investigated, are initialised. Here smooth transitions between the opposite phases of the leading two PC eigenvectors are used. These maps are considered to be arranged on a two-dimensional grid so that the individual maps can be thought of as having neighbouring states. In the training phase each time step of the investigated data set is repeatedly (as the routine progresses) assigned to the map it is currently most similar to (in terms of Euclidean distance). This map and its neighbours are nudged to pre-specified degrees towards the state of the associated data. The number of training steps and the order that the data fields (re)appear within the training is also user specified, with the degree of assimilation between data and reference map decaying over time. Once training is finished the SOM grid is thought to describe "typical" system states and so can be used to identify main features of the data and to classify meta data associated with the individual fields and/or additional data. Here, the parameters which control the training process are determined by gridded searches; c.f. MacKay (2003). The optimal SOM grid is determined by first checking that states are in fact more similar to their defined neighbours than any other states in the grid, and then selecting the SOM grid which minimises the mean difference between the generated reference maps and the training data.

Elements of the intra-seasonal data set are segregated by "crisp" cluster membership; i.e., the cluster which a given time step has the highest probability of belonging to, based on the results of the fuzzy clustering described above. These four data sets are separately analysed using SOM analysis. The resulting classifications and their highest probability transitions (ignoring for the moment the probability of repeating the current state, which for all the SOM clusters is the highest probability transition) are shown in Figure 4.3 and 4.4.

Additional tests are performed to check the robustness of the classification. The analysis is repeated using different subsections of the data to check the optimal pa-

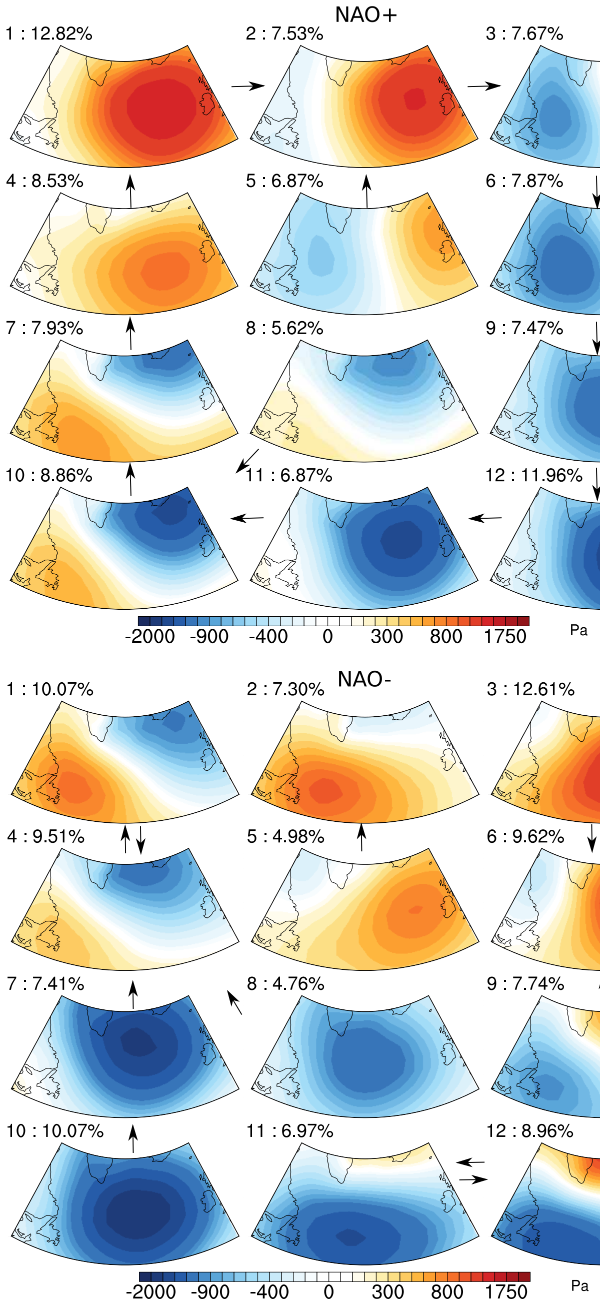


Figure 4.3: SOM grids generated from data within two of the identified clusters, percentages describe frequency of occurrence, arrows indicate which neighbour a given cluster is most likely to transition to. The labels 1-12 are used within the text to reference individual modes.

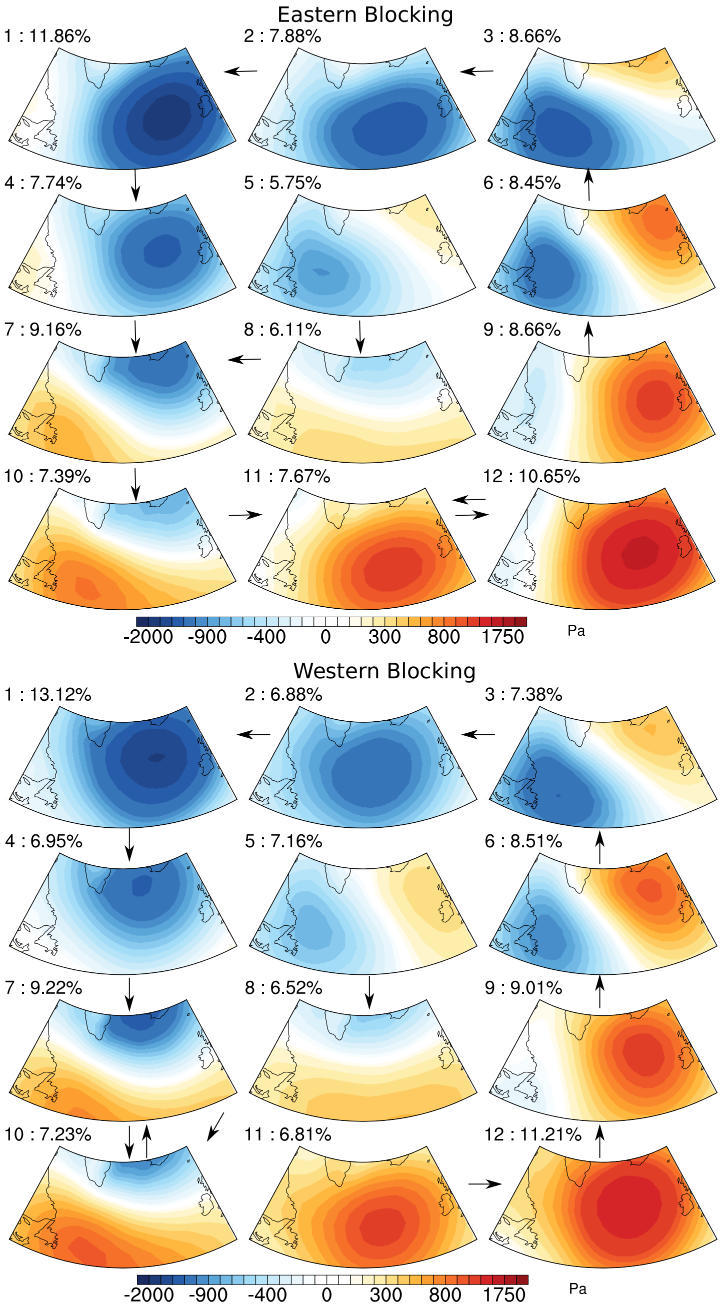


Figure 4.4: Same as Figure 4.3 but for the other two interannual clusters.

rameter settings are not overly sensitive to the range of data. The analysis is also repeated for different size grids. The 3 × 4 grids presented here are considered preferable. Smaller grids do not reproduce the full range of basic patterns seen with the twelve member grids, while the use of larger grids does not produce any new "behaviours" but simply creates more subtle distinctions between previously observed features.

The SOMs for all of the four large regimes describe similar regional patterns. They all contain an oscillating mode with a pole west of Great Britain (patterns 1 and 12 for NAO+, Eastern and Western Blocking and patterns 3 and 10 for NAO-). This corresponds to a anticylonic blocking pattern in the positive phase of the mode and cyclonic through for the negative phase. This oscillating mode has the structure of so called East Atlantic pattern described by Wallace and Gutzler (1981). The three other patterns are present in the SOMs, corresponding to a regional projection of the positive/negative NAO and the Eastern Blocking on Fig. 4.2. A pattern similar to the Western Blocking is not present in the SOMs.

4.3.5 Residual Modelling

The residual between the intra-seasonal field and the SOM derived modes above are then described using the leading eigenvectors derived from PCA. The final residual resulting from using only a limited number of eigenvectors is treated as spatially correlated white noise. As such the intra-seasonal field is constructed as,

$$\boldsymbol{x}^{(t)}_{\text{intra}-\text{seasonal}} = \underbrace{\boldsymbol{x}^{(t)}_{\text{regional}}}_{\text{SOM}} + \underbrace{\boldsymbol{x}^{(t)}_{\text{residual}}}_{\text{PCA}} + \underbrace{\boldsymbol{\xi}^{(t)}}_{\text{noise}}, \tag{4.2}$$

where $x_{\text{intra}-\text{seasonal}}$ is the vector giving the intra-seasonal state, as in Equation (1), at time t, $\boldsymbol{x}_{\text{regional}}$ is the portion of the intra-seasonal state described by the applying the

HMM to the SOM derived modes, $\boldsymbol{x}_{\text{residual}}$ is the portion of the intra-seasonal state described by the residual model, and $\boldsymbol{\xi}$ is a vector of additional noise used to represent the information lost by truncating the number of residual PCs (as described below).

The residuals are described using the nine leading PCs of the residual between the original intra-seasonal field and the defined SOM sequence. These modes describe 92% of the variance of the residual field. The PCs used here come from performing the PCA over the entire time period, although (as described above) separate regressions are fitted for data relating to separate weather regimes. It was found that subdividing the data by weather regime does not fundamentally change the structure of the leading PCs, and as well, using common PCs simplifies the use of time-lagged information during regime transitions. The regression outputs are the PC coefficients for the current day. The regression inputs are as follows:

- Which SOM grid element is prescribed as representative of the given day.
- The "strength" of the current weather regime; i.e., the amplitude of that days given regime centre when projected onto the 500 mb geopotential height field for that day.
- The PC coefficients from the two previous days. Experiments have been performed using lags from day one to three, with a two day lag producing the best results with the simulation assessment metrics.

The remaining residual resulting from using only a limited number of eigenvectors is treated as spatially correlated white noise.

4.3.5.1 Linear Inverse Model

The Linear Inverse Models (LIMs) are constructed following Kravtsov et al. (2005). This approach is essentially an iterative method for constructing an auto regressive moving average model which includes second order; i.e., interaction, terms (Strounine et al., 2010). This empirical method has been shown to perform well compared to more formal stochastic dynamical models, which typically require a more distinct scale separation between modelled and sub-scale processes (Strounine et al., 2010).

The initial implementation step is to fit a regression (through the method of least squares), including interaction terms, for the predictors described above, with categorical values identifying the current SOM cluster. The decision to use (up to) second order interaction terms in the initial model is based on the order of the equations used to model atmospheric flow (Kravtsov et al., 2005). Note that for the LIM is fit to the change between the previous and current system state (dx_i) to further mimic the atmospheric flow equations. Additionally linear residual models may also be fit using the information above (but with no interaction terms), plus the residual values forecast for the previous time step. The number of residual models is selected such that the final residual is white in time, and so in simulation can be described using a spatially correlated noise term. No such residual modelling (beyond the noise term) is needed here. This is likely due to the decision to use inputs beyond a time lag of one, which is a departure from the original Kravtsov et al. (2005) implementation. For both regressions the set of predictors used for each individual PC coefficient is pruned to find the regression resulting in the best Akaike Information Criterion (AIC) score. A conceptual overview of this approach to model selection can be found in Burnham (2004). This is an additional modification of method of Kravtsov et al. (2005). Stochastic noise is added to each forecast, with these terms generated from a multivariate normal distribution, using the sample covariance matrix calculated from the observed errors. The use of stochastic terms will occasionally create unstable simulations. This is addressed by truncating the sampling to not allow any values which would result in predictions that surpass observed minimum or maximum values. This is justified as the simulation only seeks to reproduce observed statistics, rather than investigate the possibility of new behaviours.

4.3.5.2 Bayesian Artificial Neural Networks

Artificial Neural Networks (ANNs) are a method of non-linear regression, using a network of functions linked by prescribed weights and biases to map given input to an expected output. These networks are more flexible than many empirical regression methods that are based on linear correlations, as they contain so called "hidden layers", composed of nonlinear transform functions[2] (referred to as "nodes"). As well, the Bayesian ANNs (BANNs), as developed by Neal (1996), that are employed here[3], represent an ensemble of ANNs. This ensemble is defined by posterior probability distributions for the model parameters derived by training the network against observed input-output sets. The resulting neural network is considered a nonparametric model, as it does not assume any type of statistical distribution when fitting to data (Lee, 2006). Individual "hyper-parameters" are allowed to modify the contribution of individual inputs to the network. This is know as Automatic Relevance Detection (ARD); and is described by Neal (1996) and MacKay (2003). This feature potentially provides a counterpart to the AIC pruning used in selecting the LIMs. The BANN training procedure estimates not only network parameters but also fits a noise model; i.e., each prediction target is considered to be a sample from a Gaussian distribution whose standard deviation is estimated along with the other network parameters. The networks are designed to predict the mean of this distribution.

Initially a variety of designs; i.e., the grouping and interconnectedness of inputs, nodes and outputs, as well as different sampling parameters are tested. This is done

[2] For the ANNs used in this project the non-linear transforms are the arctan function.

[3] The software used for these experiments is freely available from: `www.cs.toronto.edu/\$$\ sim$\$radford/fbm.software.html`.

by training networks on a subset of the data and testing their predictive skill on the remaining data. Time steps used for testing are selected randomly from throughout the entire observation period. The best performing design uses an initial hidden layer of a few nodes to process the SOM and regime strength information, and then joins this information with that of the lagged coefficients in a larger second layer. The actual size of the hidden layers varies between the regressions fit for separate regimes. This design is diagrammed in Figure 4.5.

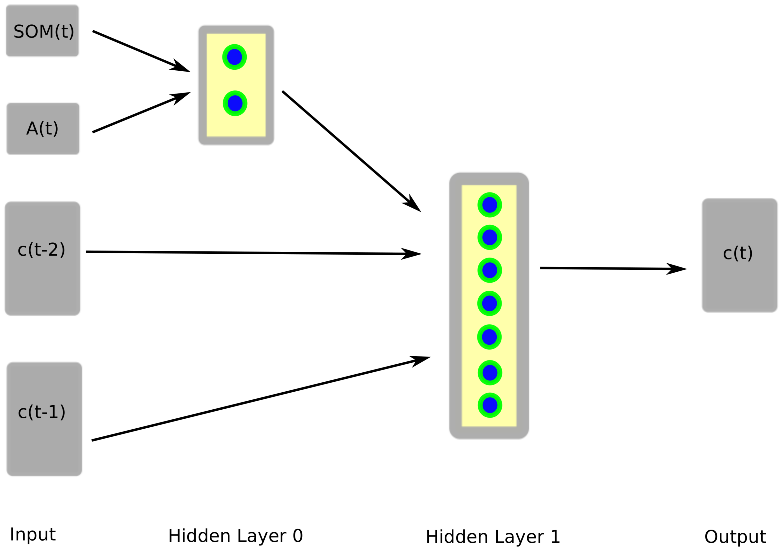


Figure 4.5: Conceptual diagram of the ANN architecture used for the experiment. The actual size of the hidden layers varies between the regressions fit for separate regimes. The top two inputs are the SOM map selected for the given day, and the strength of the interannaul regime. The bottom two are the PC coefficients for lags 2 and 1. The outputs are the PC coefficients describing the residual for the current simulation day.

This process is performed for data relating to each “crisp” weather regime cluster. As such, as with the LIMs above, each regime has its own set of regression equations for propagating the residual PCs. The final design for each regime is retrained using the entire data set for the given regime. Several network ensembles are created using different initial seeds for the parameter sampling procedure. Samples from all these ensembles are combined to create the final BANN. Propagating the model using the final BANNs is done by at each time step first selecting a random network output from within the ensemble, and then perturbing this prediction by a random variable generated from that network's noise model.

4.3.5.3 White Noise Sampling

The final residual between the retained and truncated PCs of the SOM residuals is assumed to be white in time (there is a significant drop off in lag ≥ 1 correlation following the ninth PC). A random field is generated for each time step from a multivariate Gaussian distribution, using a mean vector and covariance matrix particular to the selected weather regime and SOM mode for the given day. Often there are fewer observations mapped to a given SOM than needed to properly estimate the covariance matrix. For these cases a covariance estimate using the shrinkage method of Schaefer and Strimmer (2005) is used rather than the sample covariance matrix.

4.3.5.4 Propagation Scheme

The implementation for the stochastic simulation of the intra-seasonal state is as follows, the described procedure is outlined in Figure 4.6:

For each time step (day) t:

1. Select categorically which interannaul regime, R, the day is considered to belong to using the probabilities defined by the cluster membership $\boldsymbol{M}$ (as described above) for that day; $R^{(t)} \sim P(R|\boldsymbol{M}^{(t)})$. E.g.; if the current day is classified as 0.6 Regime 1, 0.2 Regime 2, 0.1 Regime 3, and 0.1 Regime 4, then there is a 60% probability of selecting regime 1, 20% probability of selecting regime 2, etc. The strength $A(R)$ of the chosen regime is estimated by then projecting

this regime onto the interannual data for that day and recording the resulting amplitude.

2. The choice of R in Step (1) defines which set of SOMs to select from. The SOM cluster S for time (t) is selected based on the transition probability of the SOM cluster of the previous day $(t-1)$; $S^{(t)} \sim P(S|S^{(t-1)}, R^{(t)})$. If the SOM cluster at time $(t-1)$ is from a different regime, then the transition probability is taken from the map within the set for the current regime which is most similar (by Euclidean distance) to $S^{(t-1)}$.

3. The coefficients of the n leading PCs ($\boldsymbol{c} = \{c_1, \ldots, c_n\}$) which describe the residuals are propagated by a regression, such that

$$c_j^{(t)} = f_R(S^{(t)}, A(R)^{(t)}, \boldsymbol{c}^{(t-1)}, \boldsymbol{c}^{(t-2)}) + \epsilon,$$

where a separate regression (f) is fit for each regime (R). The $S^{(t)}$ terms are taken into account using categorical regression; i.e., there are twelve additional predictors only one of which can be non-zero (actually $12-1$ predictors are used, if all eleven have a value of null, then the state is assumed to be that of the twelfth), ϵ is a stochastic term representing the regression residual.

4. Spatially correlated noise fields are added at each time step. Separate spatial means and covariance matrices are defined for each weather regime and SOM. As Step (3) is performed using modes that are orthogonal (at time (t)) to the information being modelled here, this noise and the state of Step (3) are independent.

Note that initial values are needed for the first iteration of Steps (2) and (3). The initial values are taken from the observational data so as to provide realistic combinations of components. For subsequent winters the initial values are taken from the end of the previous year to simplify implementation. As the simulations have very limited memory of initial conditions, this should not affect the results which are presented here.

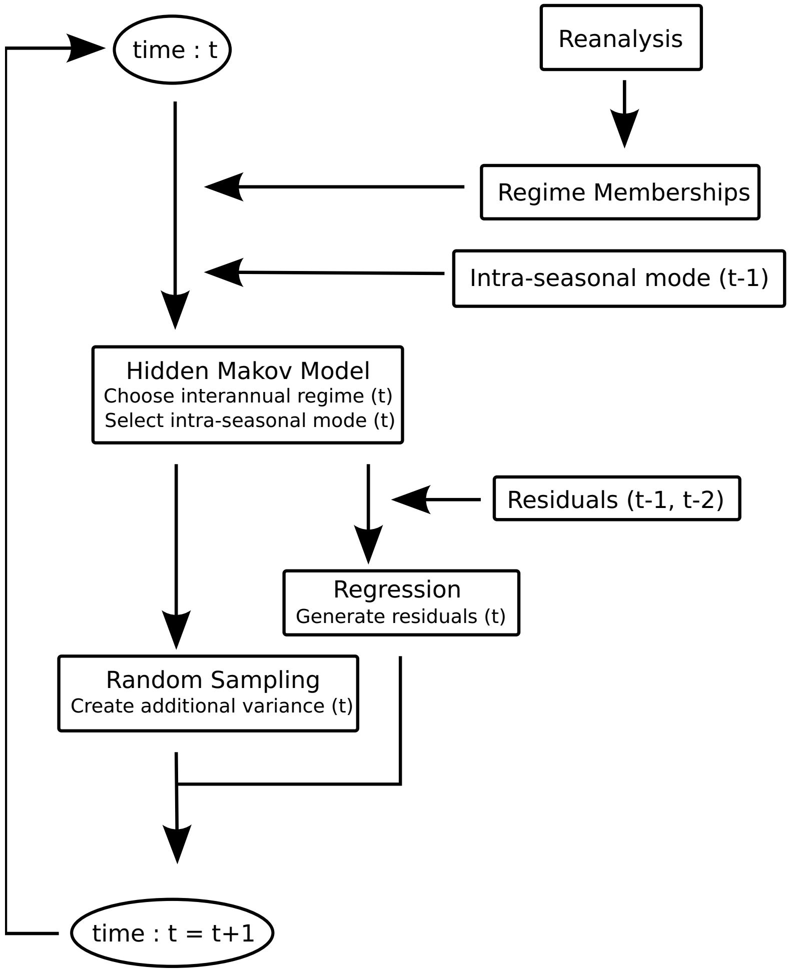


Figure 4.6: Conceptual diagram of the implementation scheme for the presented Weather Generator.

4.3.6 Results

Following Furrer and Katz (2007), the weather generator is evaluated by running simulations of the same duration as the training data series, and calculating various gross statistics for comparison against the respective results from this data set. Ensembles of simulations are created to check the range of fit to observations between model runs. Evaluation metrics are the same as in Strounine et al. (2010). To reduce dimensionality of the simulated data is projected onto the leading PCs (95% of the variance of the data set) of the training set; i.e., the NCEP intra-seasonal data for the sub-polar region. Note that these are separate PCs than those used above in constructing the residual models in the weather generator.

Comparisons are made between the distribution of the amplitudes of these PCs in the observations and amongst the ensemble members. This checks that the range of observed behaviours is reproduced by the simulations. Results for the LIM experiment are given in Figure 4.7 and for the BANN experiment in Figure 4.8. Distributions are depicted using kernel density plots to add visual comparison, and show similar information to those obtained using other non-parametric comparison methods (not shown). The LIM based model performs quite well, with the observation distribution within the span of the ensemble for most of the PCs except for isolated areas of the first two PCs. As well the ensemble range is comparatively small and the members quite self similar. This reassures that the highly stochastic nature of the simulations does not permit so much variability as to allow unrealistic results. The BANN based model appears to be outperformed by the LIM model. None of the simulated distributions match the observations as closely as seen for the LIM, with the errors appearing to be consistent through out the ensemble members. PCs 4 and 6 appear especially troublesome, and although they account for only 10% and 3% of the data respectively they seem to be extreme cases of a consistent tendency for the BANN based model to underestimate the variability seen in the observations, with all but PCs 3 and 5 having shallower tails than the observed distributions.

Comparisons are also made between the Auto Correlation Functions (ACFs) of the

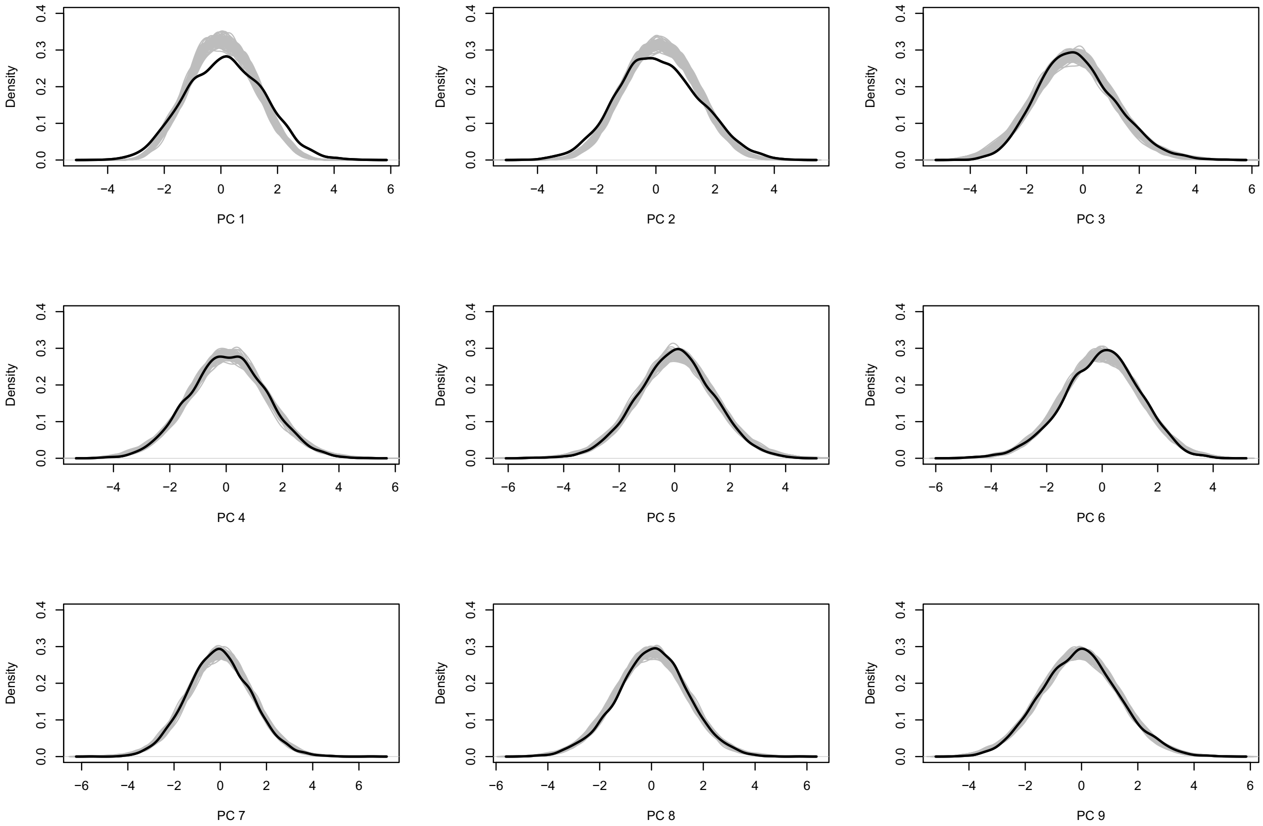


Figure 4.7: Distribution of the values for the expansion coefficients of the leading nine PCs of the observational data set (black), and the distributions of expansion coefficients obtained by projecting an ensemble of weather generator simulation obtained using the LIM (grey) onto the same PCs. Distributions are depicted using kernel density plots.

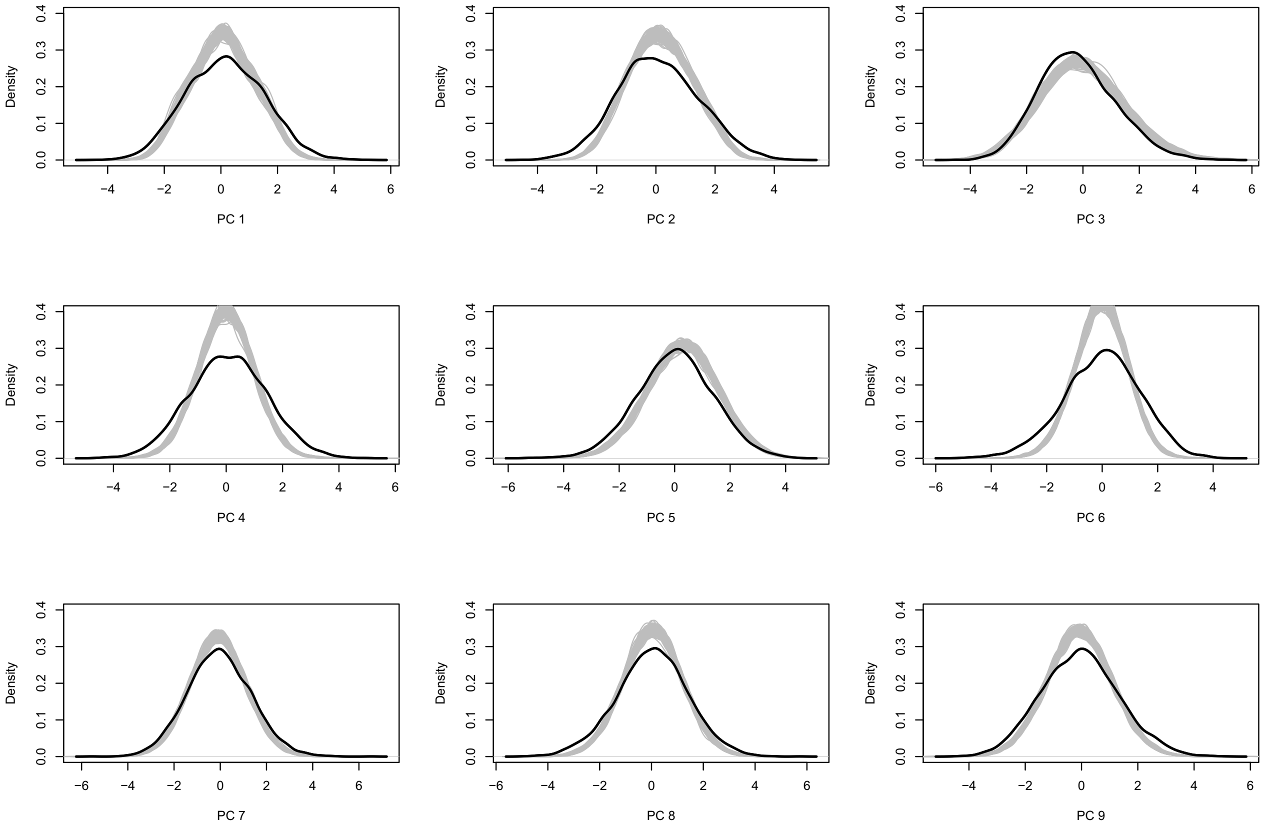


Figure 4.8: Distribution of the values for the expansion coefficients of the leading nine PCs of the observational data set (black), and the distributions of expansion coefficients obtained by projecting an ensemble of weather generator simulation obtained using the BANNs (grey) onto the same PCs. Distributions are depicted using kernel density plots.

observed and simulated expansion coefficients. This is to check that the simulations have temporal structures similar to the observations. This is important as the simulations are meant to recreate transient features. Results for the LIM experiment are shown in Figure 4.9 and for the BANN experiment in Figure 4.10. Here both methods seem to match well with the observations, although both underestimate the temporal correlation of the first two PCs. Over all the LIM method matches the observations closer, and avoids the issues with PC 6 shown by the BANNs.

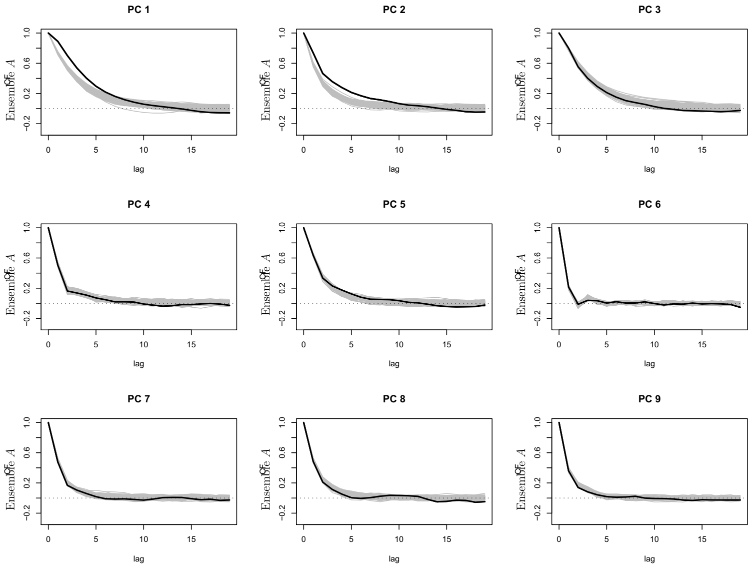


Figure 4.9: Auto Correlation Functions for the expansion coefficients of the leading nine PCs of the observational data set (black), and those of expansion coefficients obtained by projecting an ensemble of weather generator simulation obtained using the LIM (grey) onto the same PCs.

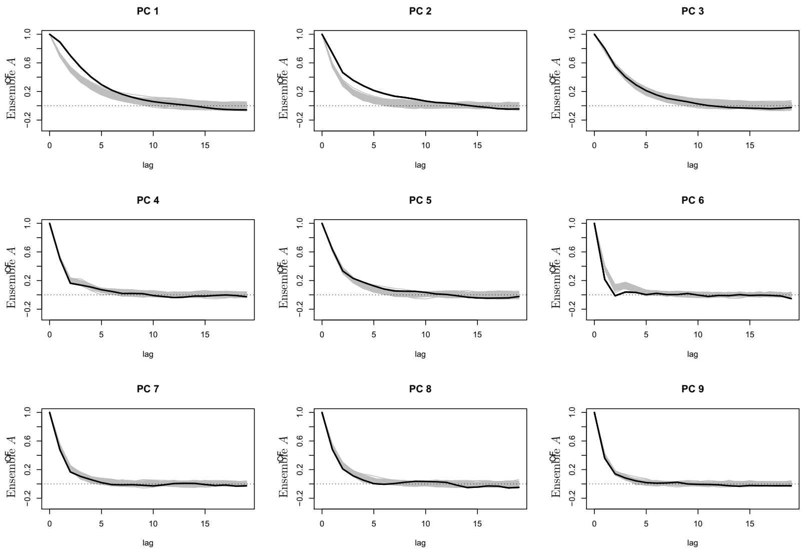


Figure 4.10: Auto Correlation Functions for the expansion coefficients of the leading nine PCs of the observational data set (black), and those of expansion coefficients obtained by projecting an ensemble of weather generator simulation obtained using the BANNs (grey) onto the same PCs.

To assess the over all phase space of the simulations, rather than individual PCs, Gaussian Mixture Models (GMMs) are fit using the joint distributions of the same expansion coefficients investigated above; following Fraley and Raftery (2002) and using the software provided by Fraley and Raftery (2006). This is to investigate if the multi-modal and non-linear behaviour of the data set is reproduced by simulations, again following Strounine et al. (2010). Using the method of Fraley and Raftery (2002) the Bayesian Information Criterion (BIC) is used to rank models composed of different numbers of spherical multivariate Gaussian distributions. The analysis is applied to the observational intra-seasonal data as well as to each ensemble member of the LIM and BANN experiments. The optimal GMM for the observational data contains fifteen distributions. A number this high is most likely indicative of the non-spherical nature of the data set, rather than the number of peaks in its probability density function. No simulation from either ensemble displayed this level of multi-modal behaviour. For both methods the ensemble mode for the optimum number of distributions is six. While this is less than seen in the observations it still shows that the models are not simply red-noise generators but incorporate some degree of system dynamics. It is expected that the simulations would be more Gaussian than the reanalysis data set, given the number of Gaussian noise components they contain. These represent only a portion of the stochastic elements of the weather generator, and their use was justified by testing the structure of the residuals they describe. However, they still contribute to a smoothing of the resulting system compared to the raw data. The BANN regression produced more (comparatively) highly multi-model ensemble members than the LIM, This is also expected given the higher degree of non-linearity in the regression and the non-parametric distribution of network parameters resulting from the Bayesian implementation.

By the applied metrics the weather generator with the LIM component appears to have outperformed that with the BANN component. This is not unexpected, as the former approach is designed to mimic flow dynamics with quadratic nonlinearities with a limited number of parameters (Kravtsov et al., 2010). The creation of a comparable result using the BANNs may require more complex models than the amount of available training data allows. There is however, no reason to expect that this result would stand for fields with a more non-linear dynamical structure, such as precipitation, especially given the more multi-modal behaviour of the BANN derived simulations.

To get an idea of how realistically the favoured weather generator describes the day to day evolution of the SLP field sample weather events are examined. To illustrate the model performance figures 4.11, 4.12 and 4.13 show three sequences from simulations. Figure 4.11 shows a low pressure system passing over the North Atlantic towards Iceland, typical of such systems which travel northeast from Newfoundland to Scandinavia by this route. Figure 4.13 shows a system taking a more southerly path while high pressure anomalies dominate the northern latitudes. Figure 4.12 shows another system originating south of Newfoundland, but being diverted by a high pressure system northward to the Labrador Sea. All three are examples of typical meteorological occurrences, with variants being generated frequently within the simulations. Figure 4.11 and 4.13 show typical NAO+ and NAO- behaviour, respectively. Figure 4.12 provides a good example of a blocking event, which typically bring wet and mild weather to the Labrador region. All of these events have important meteorological and ocean circulation associations. The occasional "wispy" appearance of the simulated systems is an artifact of the final noise terms being generated from distributions defined using shrinkage estimated covariance matrices (as described above). When sample covariance matrices are used for this component such "blurred edges" do not occur.

We are confident that the weather generator is able to reasonably emulate the system and conclude by summarising some further details of its empirical structure. The components of the HMM have been presented above. With the exception of the NAO- related SOM the spatial differences between the SOM states are subtle, with most of the uniqueness between components resulting from state intensities and transition probabilities. The regression coefficients of the LIM relating to the given SOM mapping are displayed in Figure 4.14. These inputs relate most strongly to the first, third and fourth PCs, with the latter two predictands showing almost a mirror image between the NAO+ and NAO- models and a different pattern appears for the other two weather regimes. These three PCs are similar to the patterns displayed in the SOM analysis, and so act largely to modulate the amplitude of these patterns. Distinctions between the Eastern and Western Blocking features appear in the behaviour of other PCs. Regression coefficients for other predictors (not shown) generally give more weight to the lag-1 predictors than the lag-2, and interaction terms are predom-

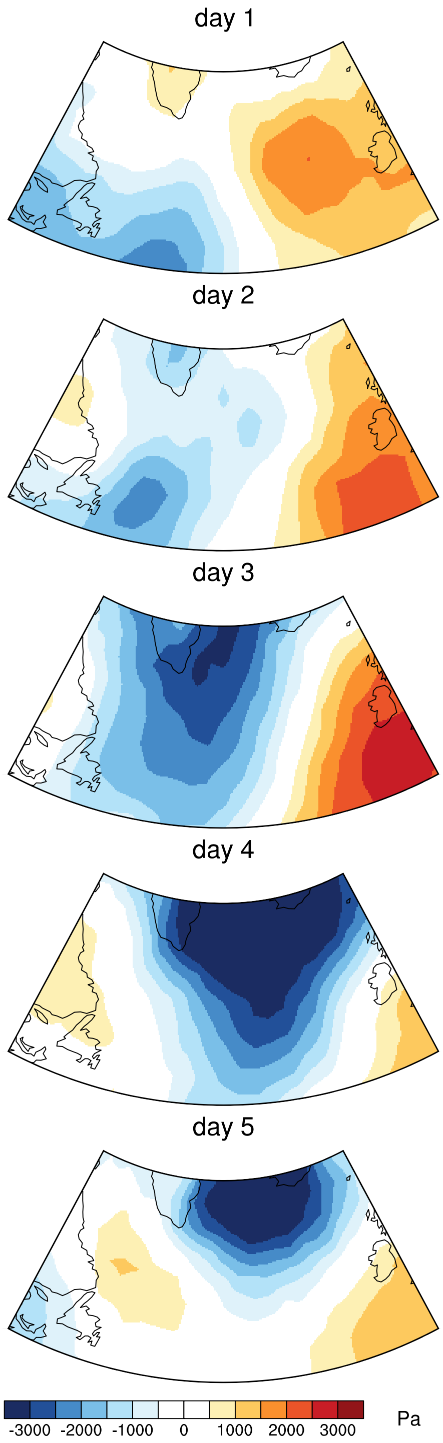


Figure 4.11: Example of low pressure system observed in the weather generator simulation following a typical "NAO+" track across the region (time proceeds from top to bottom).

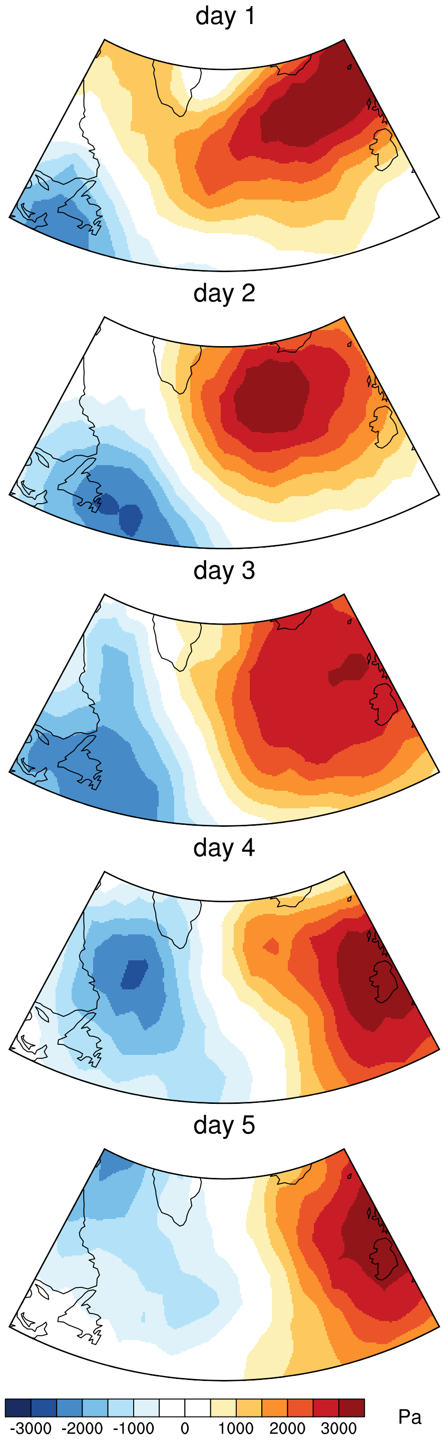


Figure 4.12: Example of low pressure system observed in the weather generator simulation durning a typical blocking event (time proceeds from top to bottom).

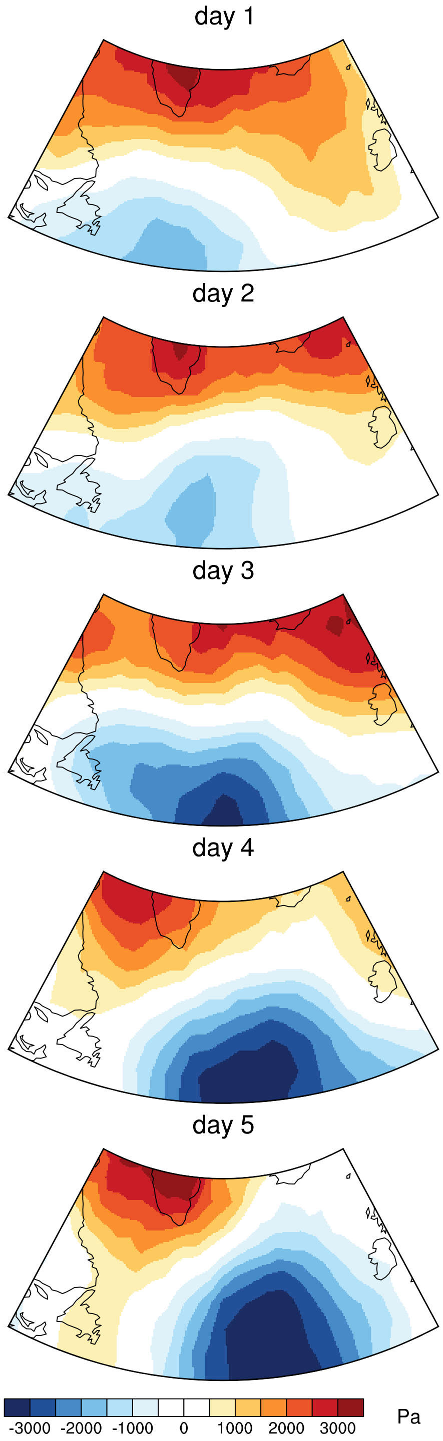


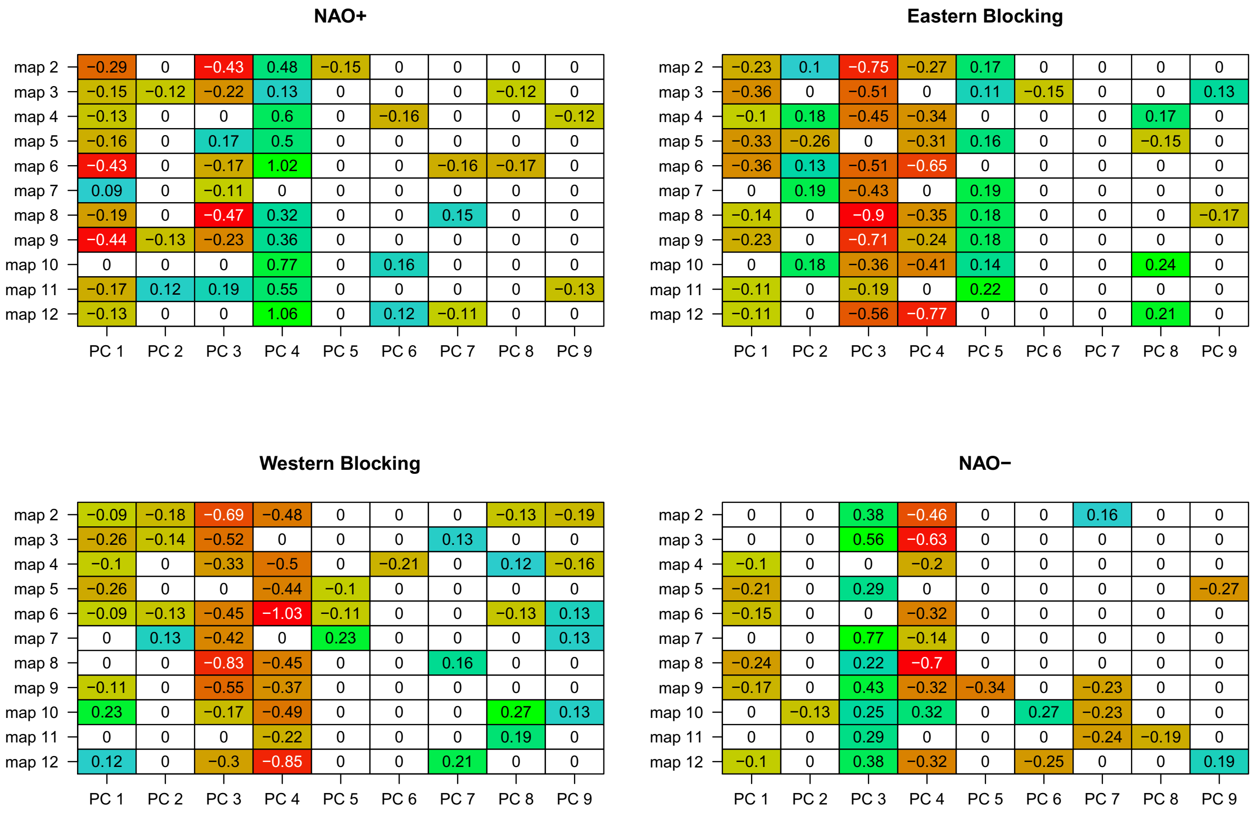
Figure 4.13: Example of low pressure system observed in the weather generator simulation following a typical “NAO-” track across the region (time proceeds from top to bottom).

inately notable only for the less dominant PCs. There is a tendency, with exceptions, for the lag-2 predictors to have higher predictive power for like elements than seen for other PCs. The structure of the regression coefficients for all components shows notable variability between weather regimes. The inputs expressing the “strength” of the relevant interannual regime were all eliminated from the model during the course of the AIC pruning. As such, we observe that for our model each of the weather regimes has a unique signature that permeates through the entire system, although more through shifts in tendencies than as a direct driver. No element of the LIM evolves independently of the others, although the system has a limited memory. Associations tend to become more indirect for the less dominant processes. We are unable to say if this is because these represent local (more “self-contained”) processes or elements driven by external considerations we have not accounted for. One of the strengths of stochastic modelling is that it allows both situations to be represented within the model framework.

4.3.7 Conclusions

In this article we describe a stochastic weather generator for a limited area of subpolar North Atlantic. The model design is multi-level and is based on the use of the dominant structure of variability at interannual, seasonal and intra-seasonal scales. Two different methods for performing the lagged regression residual modelling are tested. The differences in performance between using the empirical model reduction and BANN methods in the regression component of the model were subtle. However, the BANN forecasts had a greater tendency to underestimate the range of variability of the system.

Ensemble simulations were conducted with the stochastic weather generator. The results were compared with the original data using the metrics of Strounine et al.



NAO+

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
map 2	−0.29	0	−0.43	0.48	−0.15	0	0	0	0
map 3	−0.15	−0.12	−0.22	0.13	0	0	0	−0.12	0
map 4	−0.13	0	0	0.6	0	−0.16	0	0	−0.12
map 5	−0.16	0	0.17	0.5	0	0	0	0	0
map 6	−0.43	0	−0.17	1.02	0	0	−0.16	−0.17	0
map 7	0.09	0	−0.11	0	0	0	0	0	0
map 8	−0.19	0	−0.47	0.32	0	0	0.15	0	0
map 9	−0.44	−0.13	−0.23	0.36	0	0	0	0	0
map 10	0	0	0	0.77	0	0.16	0	0	0
map 11	−0.17	0.12	0.19	0.55	0	0	0	0	−0.13
map 12	−0.13	0	0	1.06	0	0.12	−0.11	0	0

Eastern Blocking

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
map 2	−0.23	0.1	−0.75	−0.27	0.17	0	0	0	0
map 3	−0.36	0	−0.51	0	0.11	−0.15	0	0	0.13
map 4	−0.1	0.18	−0.45	−0.34	0	0	0	0.17	0
map 5	−0.33	−0.26	0	−0.31	0.16	0	0	−0.15	0
map 6	−0.36	0.13	−0.51	−0.65	0	0	0	0	0
map 7	0	0.19	−0.43	0	0.19	0	0	0	0
map 8	−0.14	0	−0.9	−0.35	0.18	0	0	0	−0.17
map 9	−0.23	0	−0.71	−0.24	0.18	0	0	0	0
map 10	0	0.18	−0.36	−0.41	0.14	0	0	0.24	0
map 11	−0.11	0	−0.19	0	0.22	0	0	0	0
map 12	−0.11	0	−0.56	−0.77	0	0	0	0.21	0

Western Blocking

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
map 2	−0.09	−0.18	−0.69	−0.48	0	0	0	−0.13	−0.19
map 3	−0.26	−0.14	−0.52	0	0	0	0.13	0	0
map 4	−0.1	0	−0.33	−0.5	0	−0.21	0	0.12	−0.16
map 5	−0.26	0	0	−0.44	−0.1	0	0	0	0
map 6	−0.09	−0.13	−0.45	−1.03	−0.11	0	0	−0.13	0.13
map 7	0	0.13	−0.42	0	0.23	0	0	0	0.13
map 8	0	0	−0.83	−0.45	0	0	0.16	0	0
map 9	−0.11	0	−0.55	−0.37	0	0	0	0	0
map 10	0.23	0	−0.17	−0.49	0	0	0	0.27	0.13
map 11	0	0	0	−0.22	0	0	0	0.19	0
map 12	0.12	0	−0.3	−0.85	0	0	0.21	0	0

NAO−

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
map 2	0	0	0.38	−0.46	0	0	0.16	0	0
map 3	0	0	0.56	−0.63	0	0	0	0	0
map 4	−0.1	0	0	−0.2	0	0	0	0	0
map 5	−0.21	0	0.29	0	0	0	0	0	−0.27
map 6	−0.15	0	0	−0.32	0	0	0	0	0
map 7	0	0	0.77	−0.14	0	0	0	0	0
map 8	−0.24	0	0.22	−0.7	0	0	0	0	0
map 9	−0.17	0	0.43	−0.32	−0.34	0	−0.23	0	0
map 10	0	−0.13	0.25	0.32	0	0.27	−0.23	0	0
map 11	0	0	0.29	0	0	0	−0.24	−0.19	0
map 12	−0.1	0	0.38	−0.32	0	−0.25	0	0	0.19

Figure 4.14: Categorical regression coefficients (gridded values), fit to normalised values, of each input SOM map (labelled on y-axis) for forecasting each PC of the residual model (labelled on x-axis). Note that there is a different SOM grid for each map; e.g., map 2 of the NAO+ grid will not be the same feature as map 2 of the NAO- grid.

(2010). The dimensionality the simulated data is reduced through projection onto the leading PCs (95% of the variance of the data set) of the training set. The distributions of the simulated data fit with the observed distribution within the span of the ensemble for most of the PCs except for isolated areas of the first two PCs.

The further development of the method will require extension of the method to output more atmospheric variables, as well as to produce year round simulations. The present method is based on that“top down” approach, where information propagates from large to small scales. More expansive studies will require feedback effects between processes.

The results of the initial tests presented here are promising. The stochastic model is able to efficiently generate ensembles under different assumptions about the state of the large scale atmospheric variability. These simulations mimic general statistics of the observational record while producing realistic meteorological events. As well, the model components display notable shifts in short term regional dynamics associated with larger-scale/longer-term trends. As such the model can be used both as an analysis of the properties of the weather regimes it uses as predictors, and potentially to study the shifts in regime statistics associated with changing climates.

4.4 Bibliography

Aguilar-Martinez, S., Hsieh, W. W., 2009. Forecasts of tropical pacific sea surface temperatures by neural networks and support vector regression. International Journal of Oceanography 2009.

Baigorria, G., Jones, J., 2010. GiST, a stochastic model for generating spatially and temporally correlated daily rainfall data. Journal of Climate.

Benestad, R. E., Hanssen-Bauer, I., Chen, D., 2008. Empirical-Statistical Downcaling. World Scientific Publishing Co. Pte. Ltd.

Bersch, M., Yashayaev, I., Koltermann, K. P., May 2007. Recent changes of the thermohaline circulation in the subpolar north atlantic. Ocean Dynamics 57 (3), 223–235.

Bezdek, J. C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, New York.

Burnham, K. P., Nov. 2004. Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research 33 (2), 261–304.

Cappe, O., 2005. Inference in hidden Markow models. Springer, New York.

Cassano, E. N., Lynch, A. H., Cassano, J. J., Koslow, M. R., 2005. Classification of synoptic patterns in the western arctic associated with extreme events at barrow, alaska, USA. Climate Research 30 (2), 83.

Cassou, C., Sep. 2008. Intraseasonal interaction between the Madden–Julian oscillation and the north atlantic oscillation. Nature 455 (7212), 523–527.

Cassou, C., Terray, L., Hurrell, J. W., Deser, C., 2004. North atlantic winter climate regimes: Spatial asymmetry, stationarity with time, and oceanic forcing. Journal of Climate 17, 1055–1068.

Cattiaux, J., Vautard, R., Cassou, C., Yiou, P., Masson-Delmotte, V., Codron, F., Oct. 2010. Winter 2010 in europe: A cold extreme in a warming climate. Geophysical Research Letters 37 (20).

Cheng, X., Wallace, J., 1991. Cluster analysis of the northern hemisphere 500-hPa height field: Spatial patterns. Journal of the Atmospheric Sciences 50 (16), 2674–2696.

Chowdhury, M., Alouani, A., Hossain, F., 2010. Comparison of ordinary kriging and artificial neural network for spatial mapping of arsenic contamination of groundwater. Stochastic Environmental Research and Risk Assessment 24 (1), 1–7.

Corti, S., Molteni, F., Palmer, T., 1999. Signature of recent climate change in frequencies of natural atmospheric circulation regimes. Nature 398.

Duchon, C. E., 1979. Lanczos filtering in one and two dimensions. Journal of Applied Meteorology 18, 1016–1022.

Feldstein, S. B., Dec. 2000. The timescale, power spectra, and climate noise properties of teleconnection patterns. Journal of Climate 13 (24), 4430–4440.

Ferraris, L., Gabellani, S., Rebora, N., Provenzale, A., 2003. A comparison of stochastic models for spatial rainfall downscaling. Water Resources Research 39 (12).

Fraley, C., Raftery, A. E., 2002. Model-based clustering, discriminant analysis and density estimation. Journal of the American Statistical Association 97, 611–631.

Fraley, C., Raftery, A. E., 2006. MCLUST version 3 for r: Normal mixture modeling and model-based clustering. Tech. rep., Technical report.

Furrer, E. M., Katz, R. W., 2007. Generalized linear modeling approach to stochastic weather generators. Climate research 34 (2), 129.

Grimm, A. M., 2011. Interannual climate variability in south america: impacts on seasonal precipitation, extreme events, and possible effects of climate change. Stochastic Environmental Research and Risk Assessment 25 (4), 537–554.

Guo, J., Chen, H., Xu, C. Y., Guo, S., Guo, J., 2012. Prediction of variability of precipitation in the yangtze river basin under the climate change conditions based on automated statistical downscaling. Stochastic Environmental Research and Risk Assessment 26 (2), 157–176.

Hakkinen, S., Rhines, P. B., Worthen, D. L., Nov. 2011. Atmospheric blocking and atlantic multidecadal ocean variability. Science 334 (6056), 655–659.

Hashmi, M. Z., Shamseldin, A. Y., Melville, B. W., 2011. Comparison of SDSM and LARS-WG for simulation and downscaling of extreme precipitation events in a watershed. Stochastic Environmental Research and Risk Assessment 25 (4), 475–484.

Hauser, T., Keats, A., Tarasov, L., Sep. 2011. Artificial neural network assisted bayesian calibration of climate models. Climate Dynamics.

Heckerling, P., Gerber, B., Tape, T., Wigton, R., 2003. Entering the black box of neural networks. Methods Inf. Med 42 (3), 287–296.

Hewitson, B. C., Crane, R. G., 2002. Self-organizing maps: applications to synoptic climatology. Climate Research 22 (1), 13–26.

Hoskins, B. J., Hodges, K. I., 2010. New perspectives on the northern hemisphere winter storm tracks.

Hurrell, J. W., Kushnir, Y., Ottersen, G., Visbeck, M., 2003. An overview of the north atlantic oscillation. GEOPHYSICAL MONOGRAPH-AMERICAN GEOPHYSICAL UNION 134, 1–36.

Jones, P. D., Harpham, C., Kilsby, C., Glenis, V., Burton, A., 2009. Projections of future daily climate for the UK from the weather generator. Tech. rep., Met Office, Exeter, U.K.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gadin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., Joseph, D., 1996. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteoroligcal Society 77, 437–470.

Kaufman, L., Rousseeuw, P. J., 1990. Finding Groups in Data; An Introduction to Cluster Analysis. John Wiley & Sons.

Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J., 1996. Som pak: The self-organizing map program package. Technical A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

Kravtsov, S., Kondrashov, D., Ghil, M., 2005. Multilevel regression modeling of nonlinear processes: Derivation and applications to climatic variability. Journal of climate 18 (21), 4404–4424.

Kravtsov, S., Kondrashov, D., Ghil, M., 2010. Empirical model reduction and the modelling hierarchy in climate dynamics and the geosciences. In: Stochastic Physics and Climate Modelling, 1st Edition. Cambridge University Press, pp. 35–72.

Lee, H., 2006. Bayesian Nonparametrics via Neural Networks. ASA-SIAM.

Lohmann, K., Drange, H., Bentsen, M., Sep. 2008. Response of the north atlantic subpolar gyre to persistent north atlantic oscillation like forcing. Climate Dynamics 32 (2-3), 273–285.

Luo, D., Diao, Y., Feldstein, S. B., Mar. 2011. The variability of the atlantic storm track and the north atlantic oscillation: A link between intraseasonal and interannual variability. Journal of the Atmospheric Sciences 68 (3), 577–601.

MacKay, D., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., 2005. Cluster analysis basics and extensions.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themeßl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., Thiele-Eich, I., Sep. 2010. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. Reviews of Geophysics 48 (3).

Marsh, R., Josey, S. A., de Cuevas, B. A., Redbourn, L. J., Quartly, G. D., Apr. 2008. Mechanisms for recent warming of the north atlantic: Insights gained with an eddy-permitting model. Journal of Geophysical Research 113 (C4).

Minville, M., Brissette, F., Leconte, R., 2008. Uncertainty of the impact of climate change on the hydrology of a nordic watershed. Journal of hydrology 358 (1), 70–83.

Molteni, F., Kucharski, F., Corti, S., 2006. On the predictability of flow-regime properties on interannual to interdecadal timescales. In: Predictability of Weather and Climate. Cambridge University Press.

Neal, R., 1991. Bayesian mixture modeling by monte carlo simmulation. Tech. Rep. CRG-TR-91-2, University of Toronto.

Neal, R., 1996. Bayesian Learning for Neural Networks. No. 118 in Lecture Notes in Statistics. Springer-Verlag, New York.

Oelschlagel, B., 1995. A method for downscaling global climate model calculations by a statistical weather generator. Ecological modelling 82 (2), 199–204.

OrtizBeviá, M. J., SánchezGómez, E., Alvarez-García, F. J., Mar. 2011. North atlantic atmospheric regimes and winter extremes in the iberian peninsula. Natural Hazards and Earth System Science 11 (3), 971–980.

Palmer, T. N., 1999. A nonlinear dynamical perspective on climate prediction. Journal of Climate 12 (2), 575–591.

Reusch, D., Alley, R., Hewitson, B., 2007. North atlantic climate variability from a self-organizing map perspective. Journal of Geophysical Research 112, 1–20.

Robson, J., Sutton, R., Lohmann, K., Smith, D., Palmer, M., 2012. Causes of the rapid warming of the north atlantic ocean in the mid-1990s. Journal of Climate 25, 4116–4134.

Rust, H. W., Vrac, M., Lengaigne, M., Sultan, B., 2010. Quantifying differences in circulation patterns based on probabilistic models. Journal of Climate.

Sarafanov, A., Falina, A., Sokov, A., Demidov, A., Dec. 2008. Intense warming and salinification of intermediate waters of southern origin in the eastern subpolar north atlantic in the 1990s to mid-2000s. Journal of Geophysical Research 113 (C12).

Schaefer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology 4 (1), 32.

Semenov, M. A., Barrow, E. M., 1997. Use of a stochastic weather generator in the development of climate change scenarios. Climatic change 35 (4), 397–414.

Severijns, C. A., Hazeleger, W., 2010. The efficient global primitive equation climate model SPEEDO v2. 0. Geoscientific Model Development 3, 105–122.

Sexton, D. M. H., Murphy, J. M., Oct. 2011. Multivariate probabilistic projections using imperfect climate models. part II: robustness of methodological choices and consequences for climate sensitivity. Climate Dynamics 38 (11-12), 2543–2558.

Strounine, K., Kravtsov, S., Kondrashov, D., Ghil, M., Feb. 2010. Reduced models of atmospheric low-frequency variability: Parameter estimation and comparative performance. Physica D: Nonlinear Phenomena 239 (3-4), 145–166.

Tang, Y., Hsieh, W., Tang, B., Haines, K., 2001. A neural netowork atmospheric model for hybrid coupled modeling. Climate Dynamics 17, 445–455.

Tang, Y., Hsieh, W. W., Jul. 2002. Hybrid coupled models of the tropical pacific - II ENSO prediction. Climate Dynamics 19 (3-4), 343–353.

Thompson, D. W., Lee, S., Baldwin, M. P., 2003. Atmospheric processes governing the northern hemisphere annular mode/north atlantic oscillation. GEOPHYSICAL MONOGRAPH-AMERICAN GEOPHYSICAL UNION 134, 81–112.

Vallis, G. K., Gerber, E. P., Kushner, P. J., Cash, B. A., 2004. A mechanism and simple dynamical model of the north atlantic oscillation and annular modes. Journal of the atmospheric sciences 61 (3), 264–280.

von Hardenberg, J., Ferraris, L., Rebora, N., Provenzale, A., et al., 2007. Meteorological uncertainty and rainfall downscaling. Nonlinear Processes in Geophysics 14 (3), 193–199.

Wallace, J. M., Gutzler, D. S., 1981. Teleconnections in the geopotential height field during the northern hemisphere winter. Monthly Weather Review 109, 784–812.

Yiou, P., 2004. Extreme climatic events and weather regimes over the north atlantic: When and where? Geophysical Research Letters 31 (7).

Zhu, J., Demirov, E., Mar. 2011. On the mechanism of interannual variability of the irminger water in the labrador sea. Journal of Geophysical Research 116 (C3).

Chapter 5

Conclusions

5.1 Summary

The specific objectives of this thesis as listed in Section 1.6 were to:

1. Test the effectiveness of BANNs as climate simulator emulators given limited training data;

2. Test the estimation of posterior distributions of parametrised structural error models in the context of a climate simulator;

3. Examine current methods for defining weather regimes, looking at both the reproducibility of published results and estimating their associated classification uncertainties;

4. Describe long-term shifts in the distribution of these regimes;

5. Relate these shifts to regional processes;

6. Determine a computationally efficient approach to create realistic simulations, for the sub-polar North Atlantic, of local variables that capture the range of

observed variability;

7. Test if BANNs are needed to describe the residual between the discreet portion of the generator and the data to be simulated, or if this can be accomplished with a less structurally opaque model;

8. Use this model to investigate the daily signal of the trends described in Chapter 3.

The first two objectives are addressed in Chapter 2. The BANNs where shown to be effective emulators, locating high probability areas of the parameter space. Given the assumed computational limitations, the most effective strategy tested was using small sample sizes to allow multiple iterations of training and searching. Estimates of the emulator uncertainty also improved with repeated iterations. There were, however, examples of the BANNs underestimating their prediction error during the initial iterations of the calibration routine. This is in part due to an oversimplified description of the distribution of BANN predictions. Posterior structural error estimates were successfully obtained. The posterior distributions had evolved, where appropriate, notably from their priors. They also provided satisfyingly conservative descriptions of the simulator's limitations. The limiting effects of the simplistic error model on the calibration results were documented.

Objectives (3) - (5) are addressed in Chapter 3. For objective (3) Bayesian GMMs were used to define weather regimes from North Atlantic SLP anomalies. The resulting features match those obtained in previous studies. The ensemble of models produced by the Bayesian approach quantifies uncertainties regarding spatial structure and classification. This allows an evaluation of significance and robustness not seen in previous studies. Objective (4) is considered by fuzzy clustering interannual trends. This identifies four modes which allow a novel description of shifting distributions of commonly reported weather regimes. Fuzzy clustering is also able to address classification uncertainties, although not as formally as the GMM method. The interannual regimes show shifts in atmospheric tendencies over the last fifty years. These are correlated, through the fuzzy memberships, with ocean anomalies of ocean surface fields. These suggest previously undocumented ocean associations with long term shifts in the distribution of weather types.

The final objectives are addressed in Chapter 4. Stochastic simulations of daily SLP anomalies are created for the sub-polar North Atlantic region. These simulations are created by a novel combination of analogue and regression models, conditioned on the state of the interannual regimes. They well represent the range of observed variability and key features. However, simulation outputs are more smoothly distributed than in the observed system. Similar results are obtained from LIMs as for BANN models. The BANN models underestimate observed variability compared to the LIM, although they show potential for creating more multi-modal simulations. The weather generator components have little difference in spatial structure between different regimes. However, the regression coefficients, throughout every layer of the model, differ greatly for different regimes. This shows that these regimes do identify different behavioural states. They do not represent spatial reorganisation or the introduction of new features, but rather, shifts in tendencies and interactions within the system.

5.2 Future work

This thesis has discussed some contemporary issues for uncertainty quantification within the context of climate science. The studies presented here represent only initial investigations into ways to address these topics. There are many avenues for future work, some of which are now described.

Calibration Creating statistical emulations of climate models is an area that needs further exploration. One issue raised by Chapter 2 is: how to well represent the emulator uncertainty using the BANN ensemble? Gaussian assumptions are easy to incorporate within the likelihood function. However, in the experiments presented in Chapter 2, their appropriateness is questionable. As well, the simplification of the BANN distribution undermines the advantages of the non-parametric method. Transforms could be used to convert the distribution of emulator predictions to a Gaussian form; cf., Sexton et al. (2011). Alternately, results from this study suggest that using wider-tailed distributions would increase the accuracy of the approximation. Non-parametric estimates of the ensemble distribution could also be created. In either case the most effective form is likely to be specific to the particular study and the statistical structure of the BANN errors. Developing methods to create these estimates and incorporate them into the likelihood function would improve the effectiveness of BANNs as emulators within a rigorous calibration.

The results from Chapter 2 show that subtle differences in BANN architecture can have a significant effect on performance. ANN design is typically done using intuition and rules of thumb. The most appropriate architecture will vary depending on the simulator and calibration targets, and will need to be determined using trail and error. That said, conscientious documenting of methods when BANNs are used would benefit the field. Any general recommendations within the context of GCMs, or other earth systems models, would make the approach more accessible. Potentially there are no preferred approaches common to different studies. Evidence for this would also be beneficial, so that emulator designers will know to not limit themselves to architectures used in previous studies.

While there are many approaches to emulation in general, selecting appropriate methods for climate simulations is only beginning to be addressed. Given the complex and varied nature of earth systems models, it is doubtful if rigid selection guidelines could be established. Insights from direct comparisons of emulator performance, however, would be useful for specific problems.

Improving structural error descriptions is another direction for future work. Due to the large number of terms involved, error covariance matrices must be parametrised, if they are to be estimated using Bayesian methods. A common approach is using decorrelation lengths. Here, the correlation between variables is assumed to be a function of their distance from each other. Estimating parameters defining the correlation function allows a full matrix to be calculated. This can produce good results when modelling spatial fields, but defining 'distance' between more abstract variables, such as those used for calibration targets, is less intuitive. Subjective measures could be defined, potentially with additional adjustable parameters. Experiments would be needed to compare the potential for different schemes. This could be extended to block parametrization approach, where the matrix is subdivided and different parameters and schemes are used for different areas. As before, subjective assumptions will need to be made as to how to subdivide the matrix, as well as, if and how to connect different subgroups.

Decomposing structural error matrices through PCA may be a better way to reduce the dimensionality of the problem. This does reduce the matrix to the minimal number of independent descriptors. However, this analysis would define a fixed structure for the errors across the ensemble, based on a limited number of initial samples. These estimates would need to be repeated as part of the iterative calibration process. Again, experiments are needed to test the potential of the approach.

Long term biases are important structural error components, especially when the modelling objective is to make future projections. Estimating these terms is more problematic than others, as they are related to unobservable states. Defining biases essentially assumes correlation between current errors and those that will occur in future. One approach is to estimate biases from the period of observations and assume stationarity. As this is often not a satisfactory assumption, it may be necessary to compare how different GCMs describe the future states of the calibration targets. One possible way to test future associations between variables would be to include stochastic perturbations, representative of believed structural uncertainties, into the model and evaluate the long term deviations produced in simulator output.

Circulation Regimes Confidence in the results presented in Chapter 3 is limited by the length of the observational record. A reanalysis study that covers a longer historical period does exist (Compo et al., 2011). This reanalysis study, however, incorporates far less observational information than satellite-era reanalysis projects. It would be worthwhile to see if the low data study recreates the features recorded here, for the same time period. If so, then the study could be used to better estimate the significance of these regimes over a longer time period. As well, it will be important to check that these features can be reproduced to some degree by current GCMs, and other reanalysis products, if they are to be used as calibration targets and predictors.

Determining to what degree the interannual patterns of Chapter 3 appear in other data sets, would be assisted by more detailed descriptions of the original patterns. In this thesis, these regimes were described using a non-parametric fuzzy approach, so as to not limit the form of the solution. Initial tests show that similar patterns can be reproduced using non-spherical GMMs, fit using expectation-maximisation methods. However, to fit these with Bayesian methods will require extending the current implementation. As this requires estimates of off-diagonal terms with in a covariance matrix, similar issues to those described in the context of structural error modelling must be addressed. Use of Bayesian methods will produce more complete uncertainty estimates, and provide more straightforward comparisons between data sets then is possible with optimised GMMs; cf., Rust et al. (2010).

Further analysis of the behaviour of the sub-seasonal regimes within the interannual regimes is desirable. The shifts in distributions presented in Chapter 3 are a good first approximation. However, shifts in transition probabilities, residence, and typical sequences would further dynamical interpretation of the results. As well, more information could potentially be gained by applying nonlinear PCA, as described in Section 1.4, to the region. This method could provide a continuous extension to the SOM results shown in Appendix B, giving additional means for identifying dominant modes and describing their temporal evolution. These could be compared with typical sequences of the sub-seasonal regimes. If the patterns are similar, it suggests that the sub-seasonal regimes are good indicators of nonlinear atmospheric behaviour. If not, then it will have to be determined whether the classic regimes are oversimplifying features, or if the nonlinear PCA is not effective for the region; cf., Hsieh (2004).

Weather Generator One way to further explore the relationship between the sub-seasonal and interannual regimes would be to expand the region for weather generation, to that typically used with defining the sub-seasonal regimes. These patterns could then potentially be used as the SOM clusters are in the generator presented here. Descriptions of atmospheric multi-scale processes require the same analysis as that needed to construct such a weather generator. Framing the results in the form of a simulator allows for further testing of their effectiveness as a model of atmospheric behaviour.

One of the original motivations for constructing the presented weather generator was to create ensembles of forcing data for regional studies. For the current model to be useful for this context it will need to be augmented to produce more variables. One approach would be to use generated model states as predictors for other variables; i.e., add a new "bottom layer" to the model. This would require detailed investigation into in what way are the additional variables linked to SLP. Alternately, additional variables could be incorporated directly into current model, through multi-field PCA and SOM analysis.

Bibliography

Abdellatif, A. S., El Rouby, A. B., Abdelhalim, M. B., Khalil, A. H., 2010. Hybrid latin hypercube designs. In: Informatics and Systems (INFOS), 2010 The 7th International Conference on. p. 1–5.

Aguilar-Martinez, S., Hsieh, W. W., 2009. Forecasts of tropical pacific sea surface temperatures by neural networks and support vector regression. International Journal of Oceanography 2009.

Allen, M., Frame, D., Kettleborough, J., Stainforth, D., 2006. Model error in weather and climate forecasting. In: Predictability of Weather and Climate. Cambridge University Press.

Annan, J., Hargreaves, J., FRCGC, J., 2002. Climate prediction with imperfect models. QJR Meteorological Soc.,(Submitted, 2005).

Annan, J. D., Hargreaves, J. C., 2007a. Efficient estimation and ensemble generation in climate modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 365 (1857), 2077.

Annan, J. D., Hargreaves, J. C., 2007b. Efficient estimation and ensemble generation in climate modelling. Philos. Trans. Roy. Soc. London 365, 2077–2088.

Baigorria, G., Jones, J., 2010. GiST, a stochastic model for generating spatially and temporally correlated daily rainfall data. Journal of Climate.

Barabasi, B. Y. A. L., Bonabeau, E., 2003. Scale-free. Scientific American.

Bardossy, A., Bogardi, I., Matyasovszky, I., May 2005. Fuzzy rule-based downscaling of precipitation. Theoretical and Applied Climatology 82 (1-2), 119–129.

Benestad, R. E., Hanssen-Bauer, I., Chen, D., 2008. Empirical-Statistical Downcaling. World Scientific Publishing Co. Pte. Ltd.

Bergant, K., Kajfez-Bogataj, L., Crepinsek, Z., 2002. The use of EOF analysis for preparing the phenological and climatological data for statistical downscaling-case study: The beginning of flowering of the dandelion (taraxacum officinale) in slovenia.

Bersch, M., Yashayaev, I., Koltermann, K. P., May 2007. Recent changes of the thermohaline circulation in the subpolar north atlantic. Ocean Dynamics 57 (3), 223–235.

Beven, K., 2009. Environmental Modelling: An Uncertain Future? Routledge.

Bezdek, J. C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, New York.

Bhat, H., Kumar, N., Nov. 2010. On the derivation of the bayesian information criterion. Tech. rep., School of Natural Sciences, University of California.

Boe, J., Terray, L., Habets, F., Martin, E., 2006. A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling. J. Geophys. Res 111, D23106.

Brand, A., 2011. A downscaling method for application in wind resource assessment and wind power forecasting. In: Presented at 13th International Conference on Wind Engineering. Vol. 11. p. 15.

Burnham, K. P., Nov. 2004. Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research 33 (2), 261–304.

Campbell, E., 2006. A review of methods for statistical climate forecasting.

Cappe, O., 2005. Inference in hidden Markow models. Springer, New York.

Cassano, E. N., Lynch, A. H., Cassano, J. J., Koslow, M. R., 2005. Classification of synoptic patterns in the western arctic associated with extreme events at barrow, alaska, USA. Climate Research 30 (2), 83.

Cassou, C., Sep. 2008. Intraseasonal interaction between the Madden–Julian oscillation and the north atlantic oscillation. Nature 455 (7212), 523–527.

Cassou, C., Terray, L., Hurrell, J. W., Deser, C., 2004. North atlantic winter climate regimes: Spatial asymmetry, stationarity with time, and oceanic forcing. Journal of Climate 17, 1055–1068.

Casty, C., Handorf, D., Raible, C. C., Gonzalez-Rouco, J. F., Weisheimer, A., Xoplaki, E., Luterbacher, J., Dethloff, K., Wanner, H., Mar. 2005. Recurrent climate winter regimes in reconstructed and modelled 500 hpa geopotential height fields over the north atlantic/european sector 1659–1990. Climate Dynamics 24 (7-8), 809–822.

Cattiaux, J., Vautard, R., Cassou, C., Yiou, P., Masson-Delmotte, V., Codron, F., Oct. 2010. Winter 2010 in europe: A cold extreme in a warming climate. Geophysical Research Letters 37 (20).

Cheng, X., Wallace, J., 1991. Cluster analysis of the northern hemisphere 500-hpa height field: Spatial patterns. Journal of the Atmospheric Sciences 50 (16), 2674–2696.

Chowdhury, M., Alouani, A., Hossain, F., 2010. Comparison of ordinary kriging and artificial neural network for spatial mapping of arsenic contamination of groundwater. Stochastic Environmental Research and Risk Assessment 24 (1), 1–7.

Christiansen, B., May 2007. Atmospheric circulation regimes: Can cluster analysis provide the number? Journal of Climate 20, 2229–2250.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., et al., 2011. The twentieth century reanalysis project. Quarterly Journal of the Royal Meteorological Society 137 (654), 1–28.

Corte-Real, J., Hu, H., Qian, B., 1999a. A weather generator for obtaining daily precipitation scenarios based on circulation patterns. Climate Research 13, 61–75.

Corte-Real, J., Qian, B., Hong, X., 1998. Regional climate change in portugal: Precipitation variability associated with large-scale atmospheric circulation. International Journal of Climatology 18, 619–635.

Corte-Real, J., Quian, B., Xu, H., 1999b. Circulation patterns, daily precipitation in portugal and implications for climate change simulated by the second hadley centre GCM. Climate Dynamics 15 (12), 921–935.

Corti, S., Molteni, F., Palmer, T., 1999. Signature of recent climate change in frequencies of natural atmospheric circulation regimes. Nature 398.

Cox, D. R., Hinkley, D. V., 1974. Theoretical Statistics. Chapman & Hall, London.

Craig, P., Goldstein, M., Rougier, J., Seheult, A., 2001a. Bayesian forecasting for complex systems using computer simulators. Journal of the American Statistical Association Statistical Association 96 (454).

Craig, P., Goldstein, M., Rougier, J., Seheult, A., 2001b. Bayesian forecasting for complex systems using computer simulators. J. Amer. Stat. Assn. 96 (454), 717–729.

Dennett, D. C., 1991. Real patterns. The Journal of Philosophy 88 (1), 27–51.

Denny, M., 2001. Introduction to importance sampling in rare-event simulations. European Journal of Physics 22, 403–411.

Deque, M., Rowell, D. P., Luthi, D., Giorgi, F., Christensen, J. H., Rockel, B., Jacob, D., Kjellstrom, E., Castro, M., Hurk, B., Mar. 2007. An intercomparison of regional climate simulations for europe: assessing uncertainties in model projections. Climatic Change 81 (S1), 53–70.

Deser, C., Knutti, R., Solomon, S., Phillips, A. S., 2012. Communication of the role of natural variability in future north american climate. Nature Climate Change 2 (11), 775–779.

Doherty, J., Welter, D., 2010. A short exploration of structural noise. Water Resources Research 46 (5), W05525.

Dose, V., Menzel, A., Feb. 2004. Bayesian analysis of climate change impacts in phenology. Global Change Biology 10 (2), 259–272.

Duchon, C. E., 1979. Lanczos filtering in one and two dimensions. Journal of Applied Meteorology 18, 1016–1022.

Ebisuzaki, W., 1997. A method to estimate the statistical significance of a correlation when the data are serially correlated.

Edwards, N., Cameron, D., Rougier, J., 2010a. Precalibrating an intermediate complexity climate model. Clim. Dyn., 1–14.

Edwards, N. R., Cameron, D., Rougier, J., Oct. 2010b. Precalibrating an intermediate complexity climate model. Climate Dynamics.

Feldstein, S. B., Dec. 2000. The timescale, power spectra, and climate noise properties of teleconnection patterns. Journal of Climate 13 (24), 4430–4440.

Feldstein, S. B., 2007. The dynamics of the north atlantic oscillation during the summer season. Quarterly Journal of the Royal Meteorological Society.

Feliks, Y., Ghil, M., Robertson, A. W., Aug. 2010. Oscillatory climate modes in the eastern mediterranean and their synchronization with the north atlantic oscillation. Journal of Climate 23 (15), 4060–4079.

Ferraris, L., Gabellani, S., Rebora, N., Provenzale, A., 2003. A comparison of stochastic models for spatial rainfall downscaling. Water Resources Research 39 (12).

Fraedrich, K., McBride, J. L., Frank, W. M., Wang, R., 1997. Extended EOF analysis of tropical disturbances: TOGA COARE. Journal of the atmospheric sciences 54 (19), 2363–2372.

Fraley, C., Raftery, A. E., 2002. Model-based clustering, discriminant analysis and density estimation. Journal of the American Statistical Association 97, 611–631.

Fraley, C., Raftery, A. E., 2006. MCLUST version 3 for r: Normal mixture modeling and model-based clustering. Tech. rep., Technical report.

Frankignoul, C., Hasselmann, K., 1977. Stochastic climate models, part II application to sea-surface temperature anomalies and thermocline variability. Tellus 29 (4), 289–305.

Franzke, C., Woollings, T., Martius, O., 2011. Persistent circulation regimes and preferred regime transitions in the north atlantic. Journal of the Atmospheric Sciences 68 (12), 2809–2825.

Furrer, E. M., Katz, R. W., 2007. Generalized linear modeling approach to stochastic weather generators. Climate research 34 (2), 129.

Gaganis, P., Smith, L., 2001. A bayesian approach to the quantification of the effect of model error on the predictions of groundwater models. Water Resources Research 37 (9), 2309.

Gaganis, P., Smith, L., Apr. 2006. Evaluation of the uncertainty of groundwater model predictions associated with conceptual errors: A per-datum approach to model calibration. Advances in Water Resources 29 (4), 503–514.

Gaganis, P., Smith, L., 2008. Accounting for model error in risk assessments: Alternatives to adopting a bias towards conservative risk estimates in decision models. Advances in Water Resources 31 (8), 1074–1086.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 1995. Baysian Data Analysis. Chapman & Hall, London.

Ghosh, S., Mujumdar, P. P., 2006. Future rainfall scenario over orissa with GCM projections by statistical downscaling. Current Science 90 (3), 396–404.

Ghosh, S., Mujumdar, P. P., Jul. 2007. Nonparametric methods for modeling gcm and scenario uncertainty in drought assessment. Water Resources Research 43 (7).

Gillett, N. P., Graf, H. F., Osborn, T. J., 2003. Climate change and the north atlantic oscillation. In: Hurrell, J. W., Kushnir, Y., Ottersen, G., Visbeck, M. (Eds.), Geophysical Monograph Series. Vol. 134. American Geophysical Union, Washington, D. C., pp. 193–209.

Goldstein, M., Rougier, J., 2009. Reified bayesian modelling and inference for physical systems. Journal of Statistical Planning and Inference 139 (3), 1221–1239.

Goldstein, M., Rougier, J., 2010. Reified Bayesian modelling and inference for physical systems. J. Stat. Plan. Infer. 139.

Grimm, A. M., 2011. Interannual climate variability in south america: impacts on seasonal precipitation, extreme events, and possible effects of climate change. Stochastic Environmental Research and Risk Assessment 25 (4), 537–554.

Grosso, A., Jamali, A., Locatelli, M., 2008a. Iterated local search approaches to maximin latin hypercube designs. Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering, 52–56.

Grosso, A., Jamali, A., Locatelli, M., 2008b. Iterated local search approaches to maximin Latin hypercube designs. Innov. Advan. Tech. Sys., Comp. Sci. Softw. Eng.

Guo, J., Chen, H., Xu, C. Y., Guo, S., Guo, J., 2012. Prediction of variability of precipitation in the yangtze river basin under the climate change conditions based on automated statistical downscaling. Stochastic Environmental Research and Risk Assessment 26 (2), 157–176.

Haine, T., Boning, C., Brandt, P., Fischer, J., Funk, A., Kieke, D., Kvaleberg, E., Rhein, M., Visbeck, M., 2008. North atlantic deep water formation in the labrador sea, recirculation through the subpolar gyre, and discharge to the subtropics. Arctic-Subarctic Ocean Fluxes, RR Dickson et al., eds, 652–701.

Hakkinen, S., Rhines, P. B., Worthen, D. L., Nov. 2011. Atmospheric blocking and atlantic multidecadal ocean variability. Science 334 (6056), 655–659.

Hale, J., Kocak, H., 1991. Dynamics and Bifurcations. Springer-Verlag, New York.

Hanns, J., 1887. Atlas der meteorologie.

Hargreaves, J., Annan, J., Aug. 2002. Assimilation of paleo-data in a simple earth system model. Climate Dynamics 19 (5-6), 371–381.

Hargreaves, J. C., Annan, J. D., Edwards, N. R., Marsh, R., Aug. 2004. An efficient climate forecasting method using an intermediate complexity earth system model and the ensemble kalman filter. Climate Dynamics 23 (7-8), 745–760.

Hashmi, M. Z., Shamseldin, A. Y., Melville, B. W., 2011. Comparison of SDSM and LARS-WG for simulation and downscaling of extreme precipitation events in a watershed. Stochastic Environmental Research and Risk Assessment 25 (4), 475–484.

Hauser, T., Demirov, E., Feb. 2013. Development of a stochastic weather generator for the sub-polar north atlantic. Stochastic Environmental Research and Risk Assessment.

Hauser, T., Keats, A., Tarasov, L., Sep. 2011. Artificial neural network assisted bayesian calibration of climate models. Climate Dynamics 39 (1-2), 137–154.

Hawkins, E., Robson, J., Sutton, R., Smith, D., Keenlyside, N., Mar. 2011. Evaluating the potential for statistical decadal predictions of sea surface temperatures with a perfect model approach. Climate Dynamics 37 (11-12), 2495–2509.

Heckerling, P., Gerber, B., Tape, T., Wigton, R., 2003. Entering the black box of neural networks. Methods Inf. Med 42 (3), 287–296.

Hewitson, B. C., Crane, R. G., 2002. Self-organizing maps: applications to synoptic climatology. Climate Research 22 (1), 13–26.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., Ryne, R. D., 2004. Combining field data and computer simulations for calibration and prediction. SIAM Journal on Scientific Computing 26, 448.

Holden, P., Edwards, N., Oliver, K., Lenton, T., R., W., 2010. A probabilistic calibration of climate sensitivity and terrestrial carbon change in genie-1. Cli Dyn 35 (5).

Holden, P. B., Edwards, N. R., Oliver, K. I. C., Lenton, T. M., Wilkinson, R. D., Jul. 2009. A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1. Climate Dynamics 35 (5), 785–806.

Horenko, I., Apr. 2010. On clustering of non-stationary meteorological time series. Dynamics of Atmospheres and Oceans 49 (2-3), 164–187.

Hoskins, B. J., Hodges, K. I., 2010. New perspectives on the northern hemisphere winter storm tracks.

Hsieh, W. W., 2004. Nonlinear multivariate and time series analysis by neural network methods.

Hue, C., Tremblay, M., Wallach, D., Sep. 2008. A bayesian approach to crop model calibration under unknown error covariance. Journal of Agricultural, Biological, and Environmental Statistics 13 (3), 355–365.

Hurrell, J. W., Kushnir, Y., Ottersen, G., Visbeck, M., 2003. An overview of the north atlantic oscillation. GEOPHYSICAL MONOGRAPH-AMERICAN GEOPHYSICAL UNION 134, 1–36.

Jackson, C., 2009. Use of Bayesian inference and data to improve simulations of multi-physics climate phenomena. J. Phys. 180.

Jackson, C., Sen, M., Huerta, G. Deng, Y., Bowman, K., 2008. Error reduction and convergence in climate perdiction. J. Climate 21.

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

Jaynes, E., 2003a. Probability theory : the logic of science. Cambridge University Press, Cambridge UK ;;New York NY.

Jaynes, E. T., 2003b. Probability Theory; The Logic of Science, 1st Edition. University Press, Cambridge.

Jones, P. D., Harpham, C., Kilsby, C., Glenis, V., Burton, A., 2009. Projections of future daily climate for the UK from the weather generator. Tech. rep., Met Office, Exeter, U.K.

Jung, T., Palmer, T. N., Shutts, G. J., 2005. Influence of a stochastic parameterization on the frequency of occurrence of north pacific weather regimes in the ECMWF model. Geophysical Research Letters 32 (23).

Kalnay, E., 2002. Atmospheric modeling, data assimilation and predictability. Cambridge university press.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gadin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., Joseph, D., 1996a. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteoroligcal Society 77, 437–470.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., M., C., Ebisuzaki, W., Higgins, W., Janowiak, J. Mo, K., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R. Joseph, D., 1996b. The NCEP/NCAR 40-year reanalysis project. Bull. Amer. Meteor. Soc. 77.

Kanevski, M., Pozdnoukhov, A., Timonin, V., 2009. Machine Learning for Saptial Environmental Data. EFPL.

Kannan, S., Ghosh, S., Jul. 2010. Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output. Stochastic Environmental Research and Risk Assessment 25 (4), 457–474.

Kaufman, C., Sain, S., 2010. Bayesian functional ANOVA modeling using gaussian process prior distributions. Bayesian Analysis 5 (1), 123–150.

Kaufman, L., Rousseeuw, P. J., 1990. Finding Groups in Data; An Introduction to Cluster Analysis. John Wiley & Sons.

Keats, W. A., 2009. Bayesian inference for source determination in the atmospheric environment. Ph.D. thesis, University of Waterloo.

Kennedy, M., O'Hagen, A., 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society 63 (3), 425–464.

Khu, S., 2005. Multiobjective calibration with pareto preference ordering: An application to rainfall-runoff model calibration. Water Resources Research 41, 1–14.

Khu, S. T., Micha, G., 2003. Reduction of Monte-Carlo simulation runs for uncertainty estimation in hydrological modelling. Hydro. Earth Sys. Sci. 7 (5).

Khu, S. T., Werner, M. G., 2003. Reduction of monte-carlo simulation runs for uncertainty estimation in hydrological modelling. Hydrology and Earth System Sciences 7 (5), 680–692.

Kirkpatrick, S., Gelatt, C., Vecchi, M., 1983. Optimization by simulated annealing. Science 220 (4598), 671–680.

Klawonn, F., Hoppner, F., 2003. What is fuzzy about fuzzy clustering? understanding and improving the concept of the fuzzifier. In: Advances in Intelligent Data Analysis. V. Springer, Berlin, pp. 254–264.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., Meehl, G., 2010. Challenges in combining projections from multiple climate models. Journal of Climate 23, 2739–2758.

Knutti, R., Stocker, T., Plattner, G., 2003a. Probabilistic climate change projections using neural networks. Climate Dyn. 21.

Knutti, R., Stocker, T. F., Joos, F., Plattner, G.-K., Sep. 2003b. Probabilistic climate change projections using neural networks. Climate Dynamics 21 (3-4), 257–272.

Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J., 1996. Som pak: The self-organizing map program package. Technical A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

Kosko, B., 1990. Fuzziness vs. probability 17 (1), 211–240.

Kravtsov, S., Kondrashov, D., Ghil, M., 2005. Multilevel regression modeling of nonlinear processes: Derivation and applications to climatic variability. Journal of climate 18 (21), 4404–4424.

Kravtsov, S., Kondrashov, D., Ghil, M., 2010. Empirical model reduction and the modelling hierarchy in climate dynamics and the geosciences. In: Stochastic Physics and Climate Modelling, 1st Edition. Cambridge University Press, pp. 35–72.

Lee, H., 2004. Bayesian nonparametrics via Neural Networks. ASA and SIAM.

Lee, H., 2006. Bayesian Nonparametrics via Neural Networks. ASA-SIAM.

Lohmann, K., Drange, H., Bentsen, M., Sep. 2008. Response of the north atlantic subpolar gyre to persistent north atlantic oscillation like forcing. Climate Dynamics 32 (2-3), 273–285.

Lorenz, E., 1963. Deterministic nonperiodic flow. Journal of the atmospheric sciences 20, 130–141.

Lorenz, E. N., Aug. 2006. Regimes in simple systems. Journal of the Atmospheric Sciences 63 (8), 2056–2073.

Lunkeit, F., Böttinger, M., Fredrich, K., Jansen, H., Kirk, E., Kleidon, A., Luksch, U., 2007a. Planet Simulator Reference Manual. University of Hamburg, 15th Edition.

Lunkeit, F. Blessing, S., Fraerich, K., Jansen, H., Kirk, E., Luksch, U., Sielmann, F., 2007b. Planet Simulator User's Guide. University of Hamburg, 15th Edition.

Luo, D., Diao, Y., Feldstein, S. B., Mar. 2011. The variability of the atlantic storm track and the north atlantic oscillation: A link between intraseasonal and interannual variability. Journal of the Atmospheric Sciences 68 (3), 577–601.

MacKay, D., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

MacKay, D., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themessl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., Thiele-Eich, I., 2010. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. Reviews of Geophysics 48 (3).

Marsh, R., Josey, S. A., de Cuevas, B. A., Redbourn, L. J., Quartly, G. D., Apr. 2008. Mechanisms for recent warming of the north atlantic: Insights gained with an eddy-permitting model. Journal of Geophysical Research 113 (C4).

Matthies, H. G., 2007. Quantifying uncertainty: modern computational representation of probability and applications. In: Extreme Man-Made and Natural Hazards in Dynamics of Structures. Springer, p. 105–135.

McKay, M. D., Beckman, R. J., Conover, W. J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 239–245.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 23.

Michel, C., Riviere, G., Terray, L., Joly, B., May 2012. The dynamical link between surface cyclones, upper-tropospheric rossby wave breaking and the life cycle of the scandinavian blocking. Geophysical Research Letters 39 (10).

Milliff, R. F., Bonazzi, A., Wikle, C. K., Pinardi, N., Berliner, L. M., Apr. 2011. Ocean ensemble forecasting. part i: Ensemble mediterranean winds from a bayesian hierarchical model. Quarterly Journal of the Royal Meteorological Society 137 (657), 858–878.

Minville, M., Brissette, F., Leconte, R., 2008. Uncertainty of the impact of climate change on the hydrology of a nordic watershed. Journal of hydrology 358 (1), 70–83.

Molteni, F., Kucharski, F., Corti, S., 2006. On the predictability of flow-regime properties on interannual to interdecadal timescales. In: Predictability of Weather and Climate. Cambridge University Press.

Monahan, A. H., 2001. Nonlinear principal component analysis: Tropical indo-pacific sea surface temperature and sea level pressure. Journal of climate 14 (2), 219–233.

Monahan, A. H., 2002. Stabilization of climate regimes by noise in a simple model of the thermohaline circulation. Journal of physical oceanography 32 (7), 2072–2085.

Monahan, A. H., Alexander, J., Weaver, A. L., 2010. Stochastic models of the meridional overturning circulation: time scales and patterns of variability. In: Stochastic Physics and Climate Modelling, 1st Edition. Cambridge University Press, pp. 266–286.

Monahan, A. H., Pandolfo, L., Fyfe, J. C., 2001. The preferred structure of variability of the northern hemisphere atmospheric circulation. Geophysical Research Letters 28 (6), 1019–1022.

Morris, M., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics, 161–174.

Mosegaard, K., Rygaard-Hjalsted, C., 1999. Probabilistic analysis of implicit inverse problems. Inverse problems 15, 573.

Mosegaard, K., Sambridge, M., 2002. Monte Carlo analysis of inverse problems. Inv. Prob. 18.

Mukherjee, K., Dutta, R., 2011. India struggles to perfect art of monsoon forecasting. Reuters.

Muller, P., Storch, H. v., 2004. Computer modelling in atmospheric and oceanic sciences : building knowledge. Springer, Berlin ;;New York.

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

Murphy, J., 1999. An evaluation of statistical and dynamical techniques for downscaling local climate. Journal of Climate 12 (8), 2256–2284.

Murphy, J., Booth, B., Collins, M., Harris, G., Sexton, D., Webb, M., Aug. 2007a. A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 365 (1857), 1993–2028.

Murphy, J., Booth, B., Collins, M., Harris, G., Sexton, D., Webb, M., 2007b. A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. Philos. Trans. Roy. Soc. London 365, 1993–2028.

Nakicenovic, N., Alcamo, J., Davis, G., 2000a. IPCC Special Report on Emissions Scenarios. Working Group III of the Intergovernmental Panel on Climate Change IPCC. Cambridge University Press.

Nakicenovic, N., Alcamo, J., Davis, G., de Vries, B., Fenhann, J., Gaffin, S., Gregory, K., Grubler, A., Jung, T., Kram, T., La Rovere, E., Michaelis, L., Mori, S., Morita, T., Pepper, W., Pitcher, H., Price, L., Riahi, K., Roehrl, A., Rogner, H., Sankovski, A., Schlesinger, M., Shukla, P., Smith, S., Swart, R., van Rooijen, S., Victor, N., Z., D., 2000b. IPCC Special Report on Emissions Scenarios. Cambridge University Press, Cambridge, United Kingdom and New York.

Neal, R., 1991. Bayesian mixture modeling by monte carlo simmulation. Tech. Rep. CRG-TR-91-2, University of Toronto.

Neal, R., 1996a. Bayesian Learning for Neural Networks. No. 118 in Lecture Notes in Statistics. Springer-Verlag, New York.

Neal, R., 1996b. Bayesian Learning for Neural Networks. Springer-Verlag, New York.

Neal, R., 2001. Annealed importance sampling. Statistics and Computing 11, 129–139.

Neal, R., 2003a. Slice sampling. Ann. Stat. 31 (3), 705–767.

Neal, R. M., 2003b. Slice sampling. Annals of Statistics, 705–741.

New, M., Hulme, M., 2000. Representing uncertainty in climate change scenarios: a monte-carlo approach. Integrated Assessment 1 (3), 203–213.

Oelschlagel, B., 1995. A method for downscaling global climate model calculations by a statistical weather generator. Ecological modelling 82 (2), 199–204.

OrtizBevia, M. J., SanchezGomez, E., Alvarez-Garcia, F. J., Mar. 2011. North atlantic atmospheric regimes and winter extremes in the iberian peninsula. Natural Hazards and Earth System Science 11 (3), 971–980.

Palmer, T., 2006. Predictability of weather and climate: from theory to practice. In: Predictability of Weather and Climate. Cambridge University Press.

Palmer, T., Shutts, G., Hagedorn, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., May 2005. REPRESENTING MODEL UNCERTAINTY IN WEATHER AND CLIMATE PREDICTION. Annual Review of Earth and Planetary Sciences 33 (1), 163–193.

Palmer, T. N., 1999. A nonlinear dynamical perspective on climate prediction. Journal of Climate 12 (2), 575–591.

Palmer, T. N., Apr. 2012. Towards the probabilistic earth-system simulator: a vision for the future of climate and weather prediction. Quarterly Journal of the Royal Meteorological Society 138 (665), 841–861.

Prange, M., Jongma, J., Schulz, M., 2010. Centennial-to-millennial-scale holocene climate variability in the north atlantic region induced by noise. In: Stochastic Physics and Climate Modelling. Cambridge University Press, pp. 307–326.

Preisendorfer, R., 1988a. Principal Component Analysis in Meteorology and Oceanography. Elsevier.

Preisendorfer, R., 1988b. Principle Component Analysis in Meteorology and Oceography. Elservier.

R Development Core Team, 2011. R: A language and environment for statistical computing. Manual, R Foundation for Statistical Computing, Vienna, Austria.

Rabiner, L., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77 (2), 257—286.

Rasmussen, P. F., Akintug, B., Jun. 2004. Drought frequency analysis with a hidden state markov model. American Society of Civil Engineers, pp. 1–10.

Reusch, D., Alley, R., Hewitson, B., 2007. North atlantic climate variability from a self-organizing map perspective. Journal of Geophysical Research 112, 1–20.

Reusch, D. B., Alley, R. B., 2007. Antarctic sea ice: a self-organizing map-based perspective. Annals of Glaciology 46 (1), 391–396.

Richardson, D., 2006. Predictability and economic value. In: Predictability of Weather and Climate. Cambridge University Press.

Roberts, J., Pryse-Phillips, A., Snelgrove, K., Sep. 2012. Modeling the potential impacts of climate change on a small watershed in labrador, canada. Canadian Water Resources Journal 37 (3), 231–251.

Robson, J., Sutton, R., Lohmann, K., Smith, D., Palmer, M., 2012. Causes of the rapid warming of the north atlantic ocean in the mid-1990s. Journal of Climate 25, 4116–4134.

Rougier, J., Jan. 2007a. Probabilistic inference for future climate using an ensemble of climate model evaluations. Climatic Change 81 (3-4), 247–264.

Rougier, J., 2007b. Probabilistic inference for future climate using an ensemble of climate model evaluations. Climate Change 81, 247–264.

Rougier, J., 2008. Efficient emulators for multivariate deterministic functions. J. Comp. Graph. Stat. 17 (4), 827–843.

Rougier, J., Guillas, S., Maute, A., Richmund, A., 2007a. Emulating the thermosphere-ionosphere electrodynamics general circulation model. Tech. rep., Statistical and Applied Mathematical Sciences Instituted, Research Triangle Park, NC, USA.

Rougier, J. C., Guillas, S., Maute, A., Richmond, A., 2007b. Emulating the thermosphere-ionosphere electrodynamics general circulation model (TIE-GCM). submission, currently available at http://www. maths. bris. ac. uk/\ mazjcr/EmulateTIEGCM. pdf 48.

Rust, H. W., Vrac, M., Lengaigne, M., Sultan, B., 2010. Quantifying differences in circulation patterns based on probabilistic models. Journal of Climate.

Sain, S., Nychka, D., Mearns, L., 2010. Functional ANOVA and regional climate experiments: a statistical analysis of dynamic downscaling. Environmetrics.

Sambridge, M., 1999. Geophysical inversion with a neighbourhood algorithm—I. searching a parameter space. Geophysical Journal International 138 (2), 479–494.

Sambridge, M., Mosegaard, K., 2002a. Monte carlo methods in geophysical inverse problems. Rev. Geophys, 40 (3) 1009.

Sambridge, M., Mosegaard, K., 2002b. Monte carlo methods in geophysical inverse problems. Rev. Geophys. 40 (3), 1–29.

Sanderson, B., Piani, C., Ingram, W., Stone, D., Allen, M., 2008. Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics GCM simulations. Climate Dyn. 30, 175–190.

Sanso, B., Forest, C., 2009. Statistical calibration of climate system properties. Journal of the Royal Statistical Society Series C-Applied Statistics 58, 485–503.

Sanso, B., Forest, C., Zantedeschi, D., 2008. Inferring climate system properties using a computer model. Bayesian Analysis 3 (1), 1–38.

Sarafanov, A., Falina, A., Sokov, A., Demidov, A., Dec. 2008. Intense warming and salinification of intermediate waters of southern origin in the eastern subpolar north atlantic in the 1990s to mid-2000s. Journal of Geophysical Research 113 (C12).

Sato, T., 2004. The earth simulator: roles and impacts. Nuclear Physics B-Proceedings Supplements 129, 102–108.

Schaefer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology 4 (1), 32.

Schaefli, B., Hingray, B., Niggli, M., Musy, A., 2005. A conceptual glacio-hydrological model for high mountainous catchments. Hydrology and Earth System Sciences Discussions 2 (1), 73–117.

Schlesinger, M. E., Ramankutty, N., 1994. An oscillation in the global climate system of period 65-70 years. Nature 367, 723–726.

Semenov, M., Stratonovitch, P., Jan. 2010. Use of multi-model ensembles from global climate models for assessment of climate change impacts. Climate Research 41, 1–14.

Severijns, C. A., Hazeleger, W., 2010. The efficient global primitive equation climate model SPEEDO v2. 0. Geoscientific Model Development 3, 105–122.

Sexton, D. M. H., Murphy, J. M., Oct. 2011. Multivariate probabilistic projections using imperfect climate models. part II: robustness of methodological choices and consequences for climate sensitivity. Climate Dynamics 38 (11-12), 2543–2558.

Sexton, D. M. H., Murphy, J. M., Collins, M., Webb, M. J., Nov. 2011. Multivariate probabilistic projections using imperfect climate models part i: outline of methodology. Climate Dynamics 38 (11-12), 2513–2542.

Sivia, D., Skilling, J., 2006a. Data Analysis A Bayesian Tutorial, Second Edition. Oxford University Press Inc.

Sivia, D. S., Skilling, J., 2006b. Data Analysis: A Baysian Tutorial. Oxford University Press.

Snyder, C. W., Mastrandrea, M. D., Schneider, S. H., 2011. The complex dynamics of the climate system: constraints on our knowledge, policy implications and the necessity of systems thinking. In: Philosphy of Complex Systems. Vol. 10 of Handbook of the Philosophy of Science. Elsevier BV, pp. 467–505.

Solomon, S., Qin, D., Chen, Z., Marquis, M., Averyt, K., Tignor, M., Miller, H., 2007a. The physical science basis : contribution of Working Group I to the fourth assessment report of the Intergovernmental Panel on Climate Change, 1st Edition. Cambridge University Press, Cambridge [etc.].

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., Tignor, M., Miller, H., 2007b. 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.

Sparrow, C., 1982. The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors. Vol. 64. Berlin-Heidelberg-New York.

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., Kettleborough, J. A., Knight, S., Martin, A., Murphy, J. M., et al., 2005. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature 433 (7024), 403–406.

Steinhaeuser, K., Chawla, N. V., Apr. 2010. Identifying and evaluating community structure in complex networks. Pattern Recognition Letters 31 (5), 413–421.

Steinhaeuser, K., Chawla, N. V., Ganguly, A. R., 2010. An exploration of climate data using complex networks. ACM SIGKDD Explorations Newsletter 12 (1), 25–32.

Stephenson, D. B., Hannachi, A., O'Neill, A., Jan. 2004. On the existence of multiple climate regimes. Quarterly Journal of the Royal Meteorological Society 130 (597), 583–605.

Strounine, K., Kravtsov, S., Kondrashov, D., Ghil, M., Feb. 2010. Reduced models of atmospheric low-frequency variability: Parameter estimation and comparative performance. Physica D: Nonlinear Phenomena 239 (3-4), 145–166.

Struik, D. J., 1987. A Concise History of Mathematics, 4th Edition. Dover Publications Inc., New York.

Sura, P., Newman, M., Penland, C., Sardeshmukh, P., 2005. Multiplicative noise and non-gaussianity: A paradigm for atmospheric regimes? Journal of the atmospheric sciences 62 (5), 1391–1409.

Tang, B., 1998. Selecting latin hypercubes using correlation criteria. Statistica Sinica 8, 965–977.

Tang, Y., Hsieh, W., Tang, B., Haines, K., 2001. A neural netowork atmospheric model for hybrid coupled modeling. Climate Dynamics 17, 445–455.

Tang, Y., Hsieh, W. W., Jul. 2002. Hybrid coupled models of the tropical pacific - II ENSO prediction. Climate Dynamics 19 (3-4), 343–353.

Tangang, F., Tang, B., Monahan, A., Hsieh, W., 1998. Forecasting ENSO events: A neural network–extended EOF approach. Journal of Climate (11).

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

Tarasov, L., Dyke, A. S., Neal, R. M., Peltier, W., Jan. 2012. A data-calibrated distribution of deglacial chronologies for the north american ice complex from glaciological modeling. Earth and Planetary Science Letters 315-316, 30–40.

Tarasov, L., Peltier, W., Jun. 2005a. Arctic freshwater forcing of the younger dryas cold reversal. Nature 435, 662–665.

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

Teng, Q., Monahan, A. H., Fyfe, J. C., 2004. Effects of time averaging on climate regimes. Geophysical Research Letters 31 (22).

Terray, L., Demory, M. E., Deque, M., de Coetlogon, G., Maisonnave, E., 2004. Simulation of late-twenty-first-century changes in wintertime atmospheric circulation over europe due to anthropogenic causes. Journal of climate 17 (24), 4630–4635.

Tibaldi, S., Molteni, F., 1990. On the operational predictability of blocking. Tellus A 42 (3), 343–365.

Tsonis, A. A., Swanson, K. L., Oct. 2012. Review article "On the origins of decadal climate variability: a network perspective". Nonlinear Processes in Geophysics 19 (5), 559–568.

Urban, N., Fricker, T., 2010a. A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of a climate model. Comp. Geos. 36, 746.

Urban, N. M., Fricker, T. E., 2010b. A comparison of latin hypercube and grid ensemble designs for the multivariate emulation of an earth system model. Computers & Geosciences 36 (6), 746–755.

Vallis, G. K., Gerber, E. P., Kushner, P. J., Cash, B. A., 2004. A mechanism and simple dynamical model of the north atlantic oscillation and annular modes. Journal of the atmospheric sciences 61 (3), 264–280.

van der Veen, C. J., 2001. The two-mile time machine: Ice cores, abrupt climate change and our future. Eos, Transactions American Geophysical Union 82 (4), 44–44.

van Oijen, M., Cameron, D., Butterbach-Bahl, K., Farahbakhshazad, N., Jansson, P.-E., Kiese, R., Rahn, K.-H., Werner, C., Yeluripati, J., Jul. 2011. A bayesian framework for model calibration, comparison and analysis: Application to four models for the biogeochemistry of a norway spruce forest. Agricultural and Forest Meteorology.

Villagraon, A., Huerta, G., Jackson, C., Sen, M., 2008. Computational methods for parameter estimation in climate models. Bayes. Analy. 3 (4), 823–850.

von Hardenberg, J., Ferraris, L., Rebora, N., Provenzale, A., et al., 2007. Meteorological uncertainty and rainfall downscaling. Nonlinear Processes in Geophysics 14 (3), 193–199.

von Storch, H., 1999. On the use of “inflation” in statistical downscaling. Journal of Climate 12 (12), 3505–3506.

von Storch, H., Zwiers, F. W., 1984. Statistical Analysis in Climate Research. Cambridge University Press, Cambridge.

Wagener, T., Boyle, D., Less, M., Wheater, H., Gupta, H., Sorooshian, S., 2001a. A framework for development and application of hydrological models. Hydro. Earth-Syst. Sci. 5 (1), 13–26.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., Sorooshian, S., 2001b. A framework for development and application of hydrological models. Hydrology and Earth System Sciences 5 (1), 13–26.

Wallace, J. M., Gutzler, D. S., 1981. Teleconnections in the geopotential height field during the northern hemisphere winter. Monthly Weather Review 109, 784–812.

Webster, P. J., Hopson, T., Hoyos, C., Subbiah, A., Chang, H.-R., Grossman, R., 2006. A three-tier overlapping prediction scheme: tools for strategic and tactical decisions in the developing world. In: Palmer, T., Hagedorn, R., Palmer, T., Hagedorn, R. (Eds.), Predictability of Weather and Climate. Cambridge University Press, Cambridge, pp. 645–673.

Wilby, R. L., Charles, S. P., Zorita, E., Timbal, B., Whetton, P., 2004. Guidelines for use of climate scenarios developed from statistical downscaling methods. Tech. rep.

Wilby, R. L., Harris, I., 2006. A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the river thames, UK. Water Resources Research 42 (2), W02419.

Wilks, D. S., Jul. 2008. Effects of stochastic parametrization on conceptual climate models. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 366 (1875), 2475–2488.

Wilks, D. S., 2011. Statistical Methods in the Atmospheric Sciences, 3rd Edition. Elsevier.

Willet, H. C., Sanders, F., 1952. Descriptive Meteorology, 3rd Edition. Academic Press Inc.

Yashayaev, I., Clarke, A., 2006. Recent warming of the labrador sea. Gulf Region/Région du Golfe Joël Chassé1, Doug Swain6 Marine Environmental Data Service/Service des données sur le milieu marin Bob Keeley3, Mathieu Ouellet3, 7, 8, Don Spear3, 12.

Yiou, P., 2004. Extreme climatic events and weather regimes over the north atlantic: When and where? Geophysical Research Letters 31 (7).

Zadeh, L. A., 1965. Fuzzy sets. Information and Control 8 (3), 338–353.

Zhang, C., 2005. Madden-julian oscillation. Reviews of Geophysics 43 (2).

Zhu, J., Demirov, E., Mar. 2011. On the mechanism of interannual variability of the irminger water in the labrador sea. Journal of Geophysical Research 116 (C3).

Zhu, J., Demirov, E., Dupont, F., Wright, D., Aug. 2010. Eddy-permitting simulations of the sub-polar north atlantic: impact of the model bias on water mass properties and circulation. Ocean Dynamics 60 (5), 1177–1192.

Zwiers, F. W., 1990. The effect of serial correlation on statistical inferences made with resampling procedures. Journal of Climate 3, 1452–1461.

Appendix A

Artificial neural network assisted Bayesian calibration of climate models

A.1 Analysis of error models

The following is the supplementary material provided for the publication of Chapter 2

Here we present further analysis of the error models used for the emulators and the model discrepancy. Figure A.1 shows a normal Q-Q plot of the results of scaling the differences between the mean BANN prediction and actual model output by the uncertainties estimated for each prediction using the 1σ range of the BANN posterior for that target and parameter set. These are the emulators which were used to select the final sample of Ensemble A for the perfect model experiment; i.e., they were trained on the first 80 model runs of the ensemble, which are tested against the final model runs produced for this ensemble. The fit to a Gaussian is reasonable for the most part, except for at the tails of the distribution of emulator errors, which are quite exaggerated. Our model of emulator error does not account for correlations, and so we do not expect a close fit, but it is clear that here the errors are too long tailed to be normally distributed. As such it would seem a more appropriate choice of error model would be a thicker tailed distribution. Figure A.2 shows similar behaviour for the emulators which were used to select the final sample of Ensemble A for the calibration to reanalysis data. The behaviour seen here is more muted however, except for an extreme outlier.

To give an impression of the general evolution of the central tendencies of BANN performance we include Figures A.3 and A.4. These show the spread of RMS errors (using their mean and standard deviation) between model output and the emulator predicted values for each iteration of the calibration experiments (plots are for BANNs used to predict temperature values, and are representative of the overall BANN behaviour). Also included is the mean predicted emulator uncertainty at each iteration.

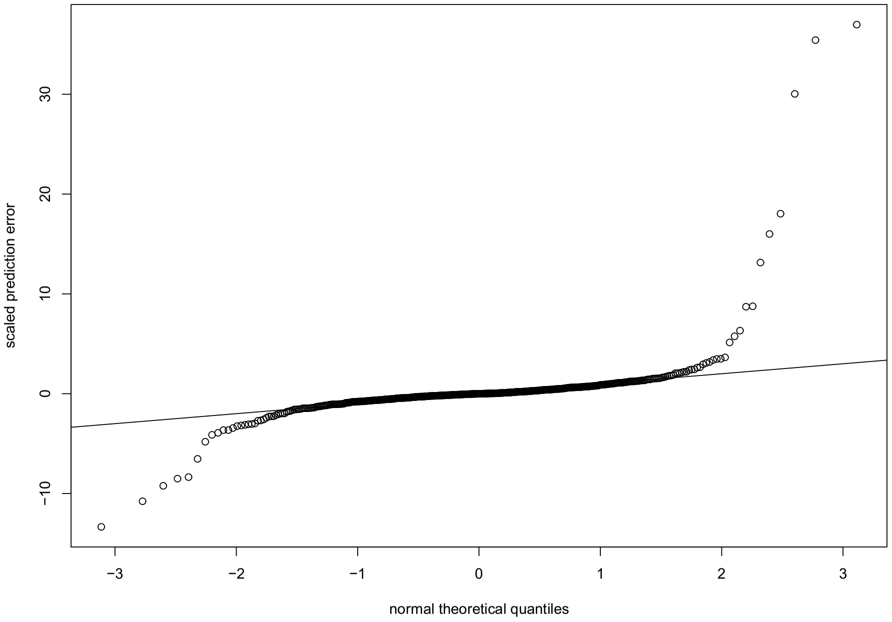


Figure A.1: Normal Q-Q plot of differences between emulator prediction and actual model output, with each error scaled by its associated emulator-predicted uncertainty. Also plotted is the line of unit slope. These are the emulators which where used to select the final sample of Ensemble A for the perfect model experiment.

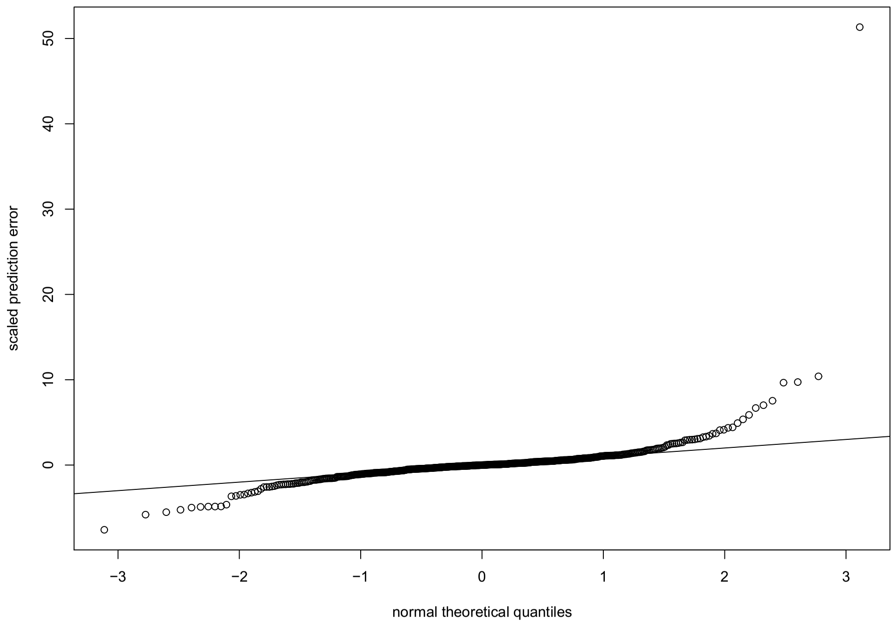


Figure A.2: Normal Q-Q plot of differences between emulator prediction and actual model output, with each error scaled by its associated emulator-predicted uncertainty. Also plotted is the line of unit slope. These are the emulators which where used to select the final sample of Ensemble A for the calibration to reanalysis experiment.

It can be seen that for much of the calibration routine this value is comparable to the mean error.

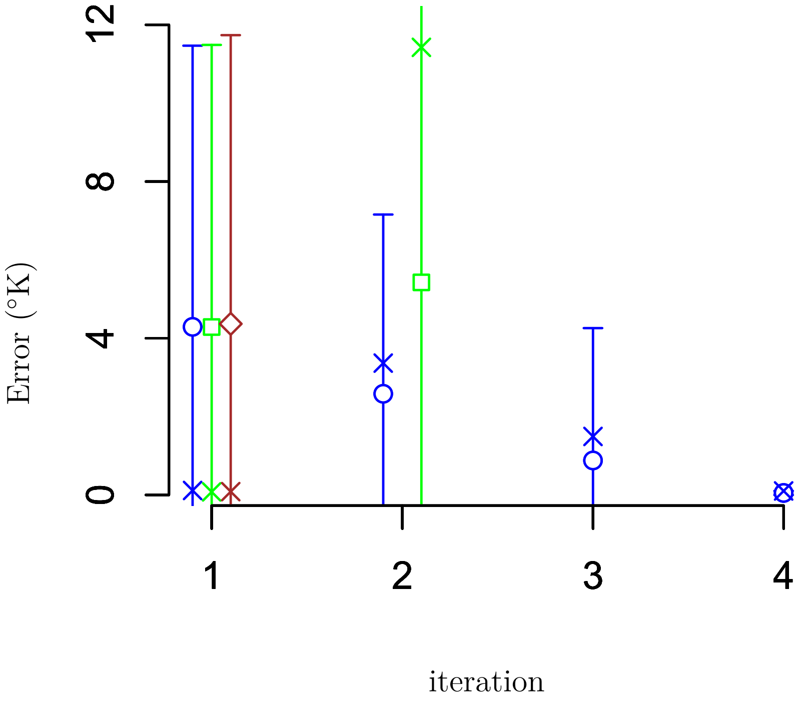


Figure A.3: Spread of RMS errors (y-axis) between actual model responses and those predicted by the emulator for each iteration (x-axis, points are offset for clarity) of the perfect model experiment, are displayed with the mean bracketed by the standard deviation. Ensembles A, B, and C are represented by the colours blue (circle), green (square), and brown (diamond), respectively. Crosses represent the mean predicted emulator error as estimated by the emulator.

To check the validity of the our model discrepancy estimates, given our poor description of the distribution of emulator predictions, we again construct Q-Q plots, this time for the differences between the outputs of the members of the final model ensemble (for the calibration to reanalysis data experiment) and the calibration targets. Each error is scaled by the associated model discrepancy and (much smaller) observa-

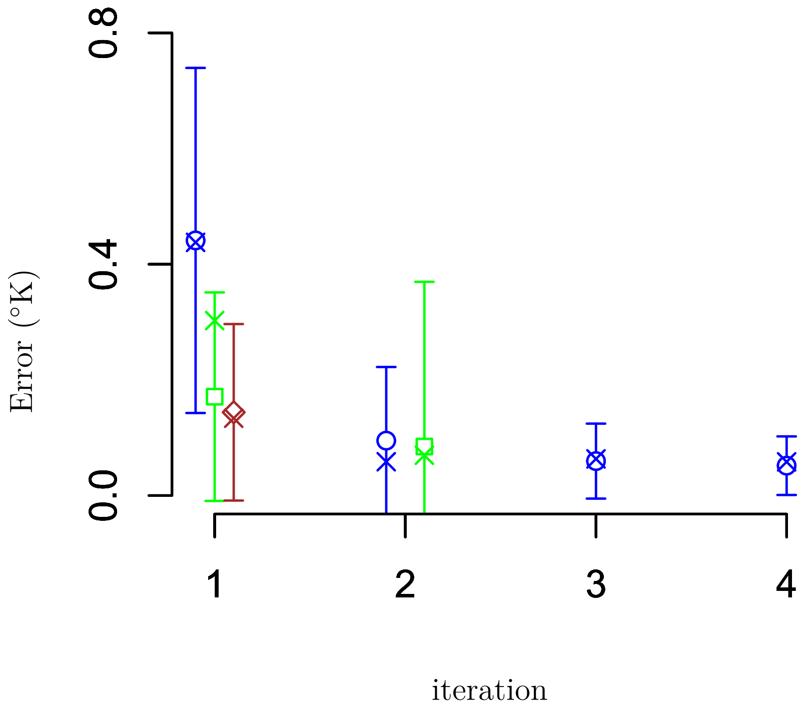


Figure A.4: Spread of RMS errors (y-axis) between actual model responses and those predicted by the emulator for each iteration (x-axis, points are offset for clarity) of the calibration to NCEP/NCAR data, are displayed with the mean bracketed by the standard deviation. Ensembles *A*, *B*, and *C* are represented by the colours blue (circle), green (square), and brown (diamond), respectively. Crosses represent the mean predicted emulator error as estimated by the emulator.

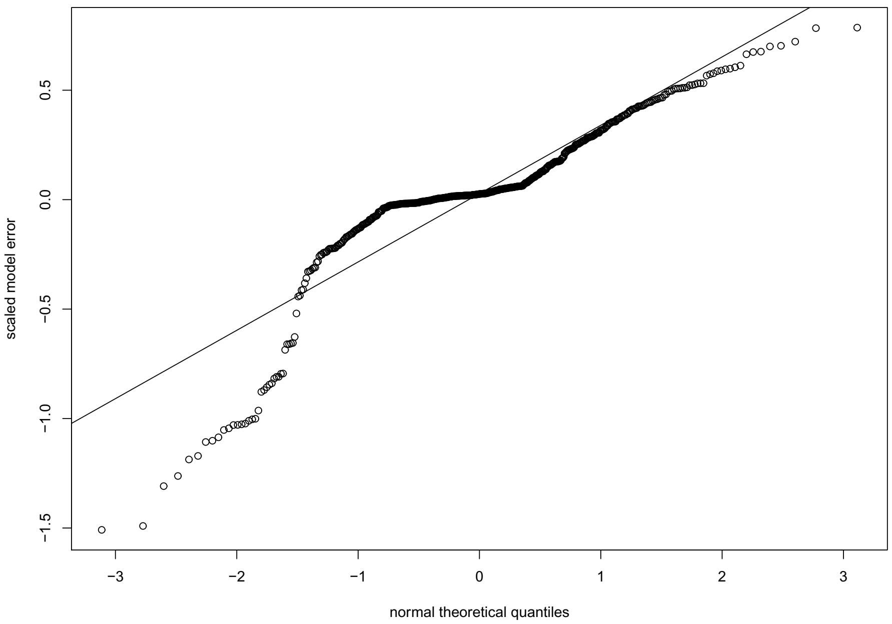


Figure A.5: Normal Q-Q plot of differences between model output and calibration targets, with each error scaled by its associated estimated model and observational uncertainty. The line of best fit to the presented points is also plotted.

tional uncertainty estimate and this distribution is compared to a standard Gaussian in Figure A.5. The distribution is skewed, and somewhat biased, as expected, as no attempt was made to account for bias or correlation in the model output. However, the range of the scaled errors compared to the standard Gaussian shows that the estimated model discrepancy is quite conservative. Without more sophisticated error models, we can't know what effect the over simplified emulator error model has had on the results of the model discrepancy estimates, although given the large difference in scale between the emulator and model errors it is unlikely to have been significant.

A.2 Examiner Discussion

The following documents additional discussion of the topics of Chapter 2

1. *Within the Chapter 2 paper, the candidate developed an emulator based on a BANN methodology to calibrate parameters of a climate model. It is clear that employing an emulator to increase the number of parameter combination tests is a positive development, it is less clear that these tests will be successful in producing optimal model parameters for the GCM used. In addition, the implications of the data summarization and filtering on the parameter selection is not clear.*

   - The goals of calibration are different than those of optimisation. However, the "perfect model experiment" does suggest that, while caution has to be used (eg ensemble B), the method is capable of locating 'ideal' areas of the parameter space. Studying the effects of calibration targets was beyond the scope of the presented study. Due to the simplicity of the GCM, the described experiment represents more of a 'toy problem' for exploring the methodology, rather than an exhaustive study designed to result in a fully calibrated model for use in climate investigations.

2. *It would be useful to see a flow chart, similar to the one used in Chapter 4 to assist in focusing the discussion of data preparation and information flow in for the parameter selection study.*

   - Such a chart is given as Figure A.6. This is a conceptual diagram only which ignores certain steps in the procedure. A description of the full iterative procedure is in Section 2.3.4.

3. *Why are the calibration targets based on the EOF fields split into positive and negative domains? This choice seems rather arbitrary. What is the justification for this choice? Were any alternatives considered? This point needs further elaboration in the thesis.*

    - As is stated in Chapter 2, the calibration targets are not the focus of the study being performed. As such, generic features, essentially continental scale averages, were calculated and used to test the methodology. As observed in the article these generic targets do not appear to be ideal should someone want to perform a full calibration of the model. As the EMIC used here serves only as a test case for the method further discussion of observational features is reserved for Chapter 4, although in a different context.

4. *I don't understand why amplitude vectors with weights less than* 10^{-5} *are set to zero. Surely the definition of what is small is more dependent on what maximum value of the particular vector is. Are these vectors normalised?*

    - This is an arbitrary cutoff to simplify the calculation. There is no notable difference in the result whether such small values are included in the averaging or not.

5. *I flagged this text on reading but then when I went back it seemed a bit clearer. But, what is the "default" value here, is it the mean of the proposed distribution? I'm not sure I know what "multiplicatively expanding" means here. Aren't you just assigning a distribution with a large variance?*

- The term "default", which is undesirable but lacks a concise alternative, is described in Footnote 10.

6. *Am I to understand that large variances are being selected arbitrarily to accommodate unknown variability? Does this approach in fact capture the outer limits of variance supported by inputs absent covariance?*

    - As discussed in the article in reference to figures 2.8 and Figure A.5 the structural error model is if anything an overly conservative description of model error.

7. *You say "order of magnitude higher" comparing targets between fields, but what is this relative to. What if you're comparing temperature to sea level?*

    - This is a reference to the ratio between the target and its standard deviation, indicated by the term relative uncertainty.

8. *"the overall error is reduced ..." How is this reduction quantified? By how much is the error reduced?*

    - This is quantified in Table 2.2

9. *On the 3rd line of Equation 2.10, how does f suddenly appear among the variables being conditioned upon? Its presence would suggest that f is perfectly known given theta and z - this doesn't seem consistent with the formalism developed above, in which there is an error in f associated with the emulator.*

    - Equation 2.10 describes making forecasts with an ensemble of calibrated GCMs, and so the emulator is not used to generate simulator outputs. If the emulators were used to construct such an ensemble their uncertainties would have to be taken into account.

10. *What is it about θ_3 that results in its particularly poor estimation in Ensemble B? Is this poor estimation related to the bimodality of theta1?*

    - This illustrates one of advantages of using an iterative technique when a small number of GCM runs can be sampled. This gives the emulator 'a chance to learn from it's mistakes'. It is possible for a limited sample of training data to suggest erroneous locations of high probability parameter space, which can then be explored in the next iteration. For the B experiment where the training data limit was reached by iteration 2, there was no opportunity for this. Note, as can be seen in Figures 2.3 and A.3 the BANN used for iteration 2 of experiment B appears to have performed comparatively poorly at emulating the model response, suggesting the emulator's extrapolations about the portion of the parameter space shown in Figure 2.2 turned out to be inaccurate. This could have been resolved by retraining the emulator with the newly generated data. Also, simpler emulator architectures were used in the 'perfect model experiment'. The multi-layer BANNs used in the second experiment do not have these sorts of errors, suggesting they are able to better able to detect features of the parameter space.

11. *Did you investigate how the parameter calibration changed with changes in the calibration targets? For example, if you calibrate on just temperature, how different are the results from calibrating on all of temperature, pressure, and humidity? Based on conversations I have had with colleagues who tune such models "by hand", it's common that instead of parameter estimates becoming sharper as more target fields are included, they become broader (because of systematic errors).*

- Sharper parameter estimates are expected when using more specific targets. This is the motivation for using multiple fields as targets in the presented experiment, so as to better simulate a multi-objective calibration. These are typical in GCM calibration where it's necessary to be wary that the model is not being over fit to one variable/field at the expense of having 'realistic' physical mechanisms.

12. *It would be good to show similar distributions based on the model parameter priors, as well as distributions of the change in predicted temperature change by the mid 21st century for the calibrated and uncalibrated models (which would indicate if the model calibration has any effect on climate sensitivity).*

    - Figure 2.10 was included in the article only as an example of the potential of the probablistic method, since future projects don't give any way to evaluate the accuracy of the calibration. However, this does appear to be a good way to compare the effect of different calibration methods and targets when such experiments are performed.

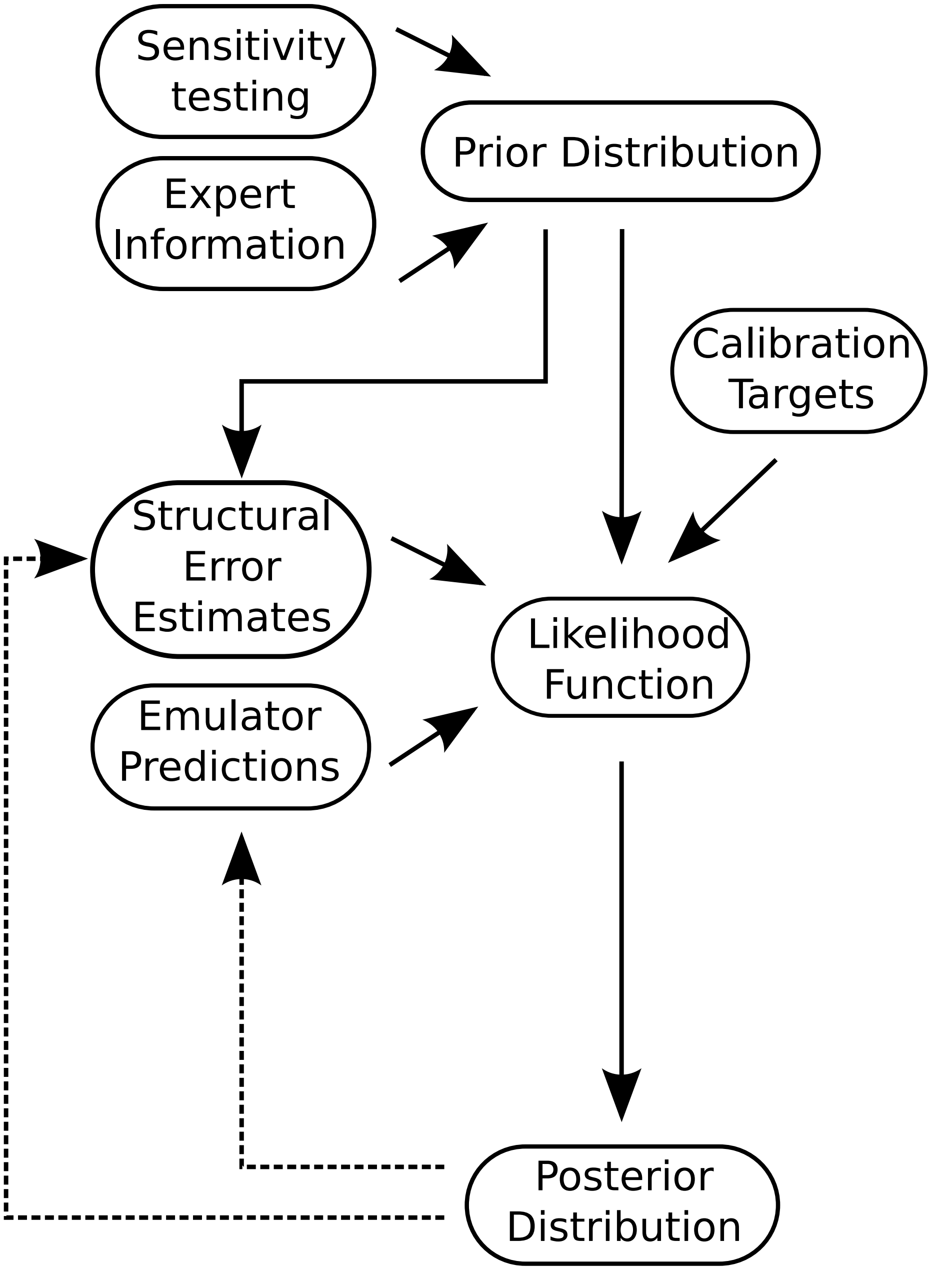


Figure A.6: Conceptual diagram of information flow during model calibration procedure.

Appendix B

North Atlantic atmospheric and ocean interannual variability over the past fifty years - spatial patterns and decadal shifts

B.1 Alternate approaches to clustering North Atlantic daily SLP anomalies

As stated in Chapters 1 and 3, the classical daily-data regimes for North Atlantic SLP anomalies of Chapter 3 can be reproduced using a variety of methods. The results from different cluster analysis methods, applied to the same data set, is presented here. As expected most of the methods produce similar features. However, some interesting variations occur when the number of clusters is allowed to vary.

B.1.0.1 K-Means

Centres obtained using the k-means algorithm (the method most commonly applied in the literature) are shown in Figure B.1. These reproduce the standard results as reported by the sources cited in Section 1.4. The regime assigned for each day of the winter of 2012 is shown in Figure B.2 The percentage of occurrence of each regime for each winter is shown in Figure B.3.

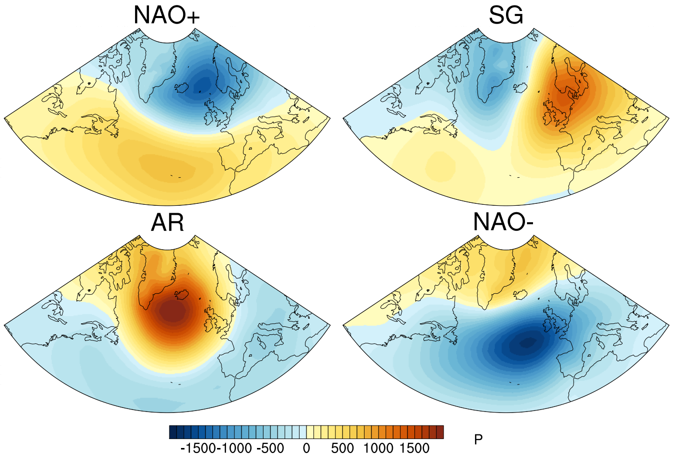


Figure B.1: Centres calculated using K-Means algorithm on winter (DJF) daily SLP anomalies.

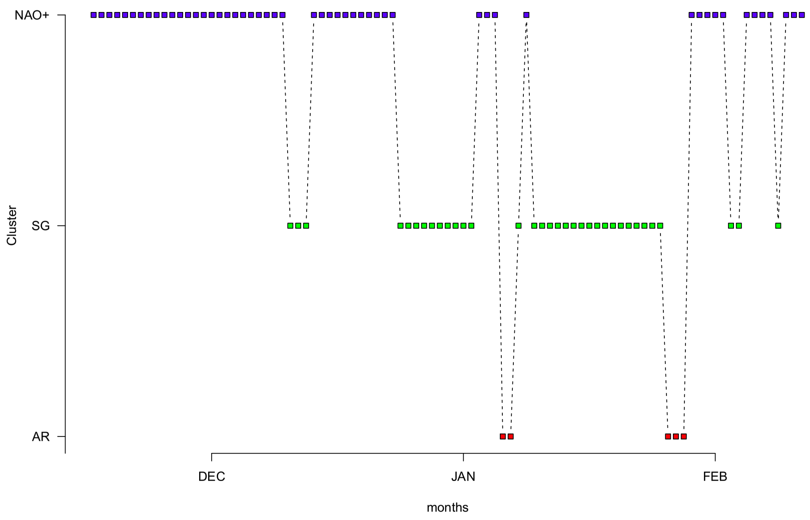


Figure B.2: Classification of each day of the 2012 winter (DJF) season, as given by the K-Means clustering of SLP anomalies.

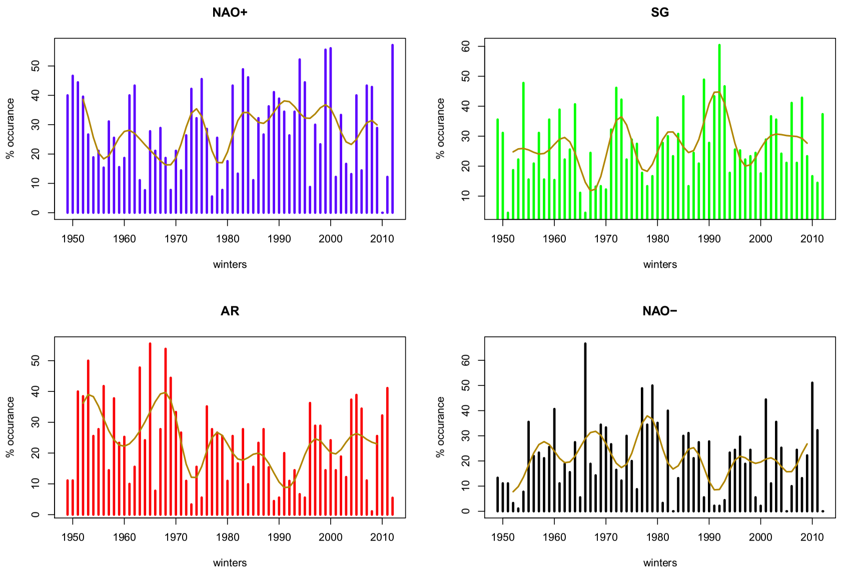


Figure B.3: Winter (DJF) occurrence frequencies for the K-Means SLP centres.

B.1.0.2 Fuzzy Clusters

Centres obtained using the fuzzy clustering algorithm (with fuzziness parameter set to 1.2) are shown in Figure B.4. Note that these are very similar to the k-means results. The degree of membership for each day of the winter of 2012 is shown in Figure B.5. The mean cluster memberships for each winter is shown in Figure B.6. These time lines match the k-means results but allow for a continuous measure of the system evolution. This is particularly notable when comparing Figure B.2 and Figure B.5.

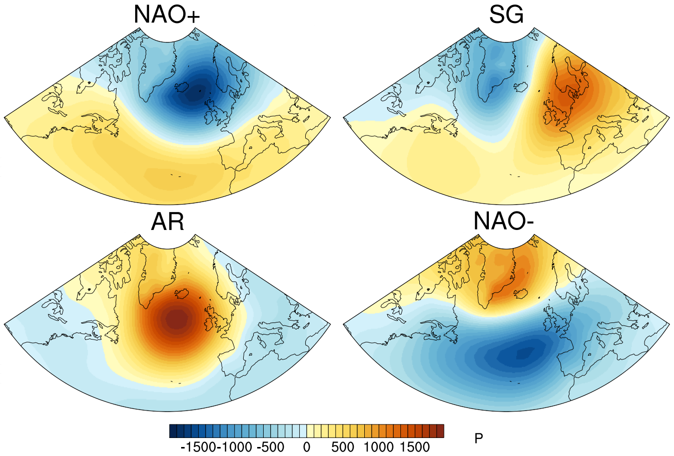


Figure B.4: Centres calculated using Fuzzy Clustering on winter (DJF) daily SLP anomalies.

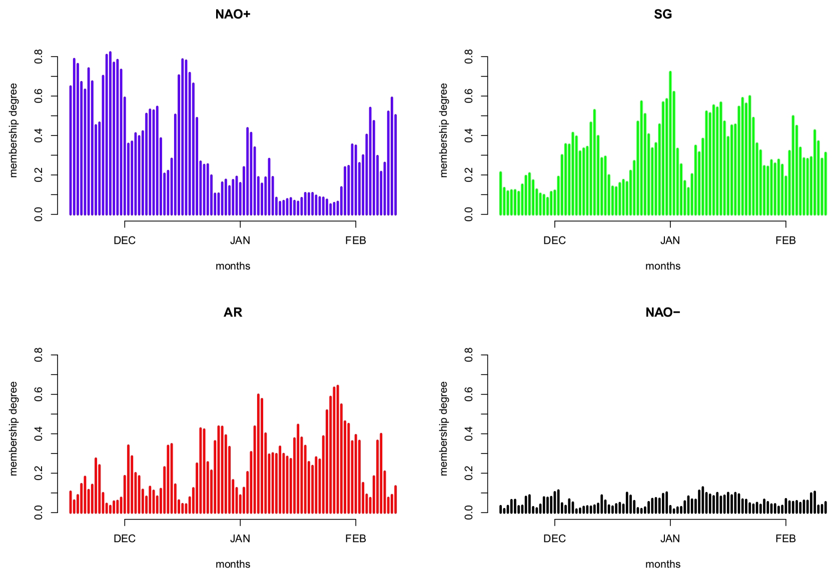


Figure B.5: Daily fuzzy memberships for the 2012 winter (DJF) season.

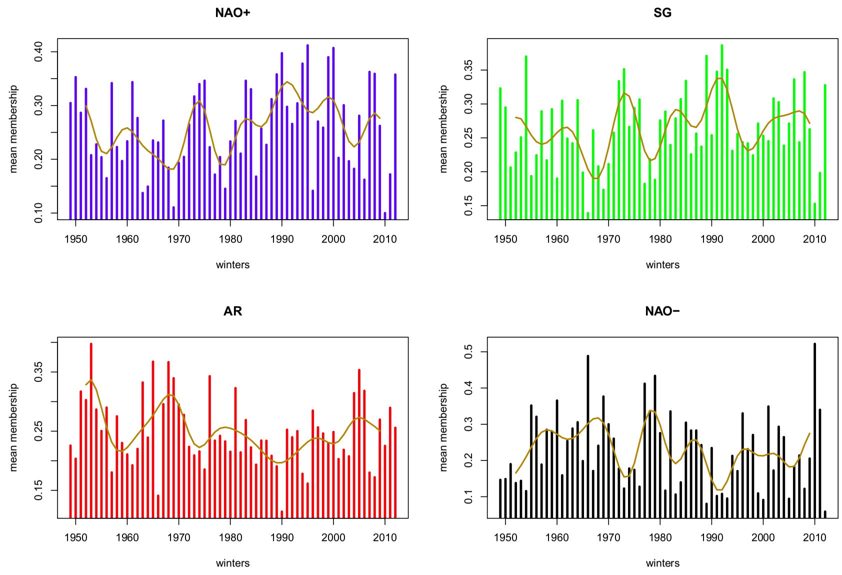


Figure B.6: Mean winter (DJF) memberships for the Fuzzy Clustering SLP centres.

B.1.0.3 GMM (Expectation-Maximisation)

Centres obtained using a four component spherical GMM,fit using an Expectation-Maximisation (E-M) routine, are given in Figure B.7. The probability of membership to each cluster for each day of the winter of 2012 is shown in Figure B.8. The mean probability of membership for each winter is shown in Figure B.9. Results are similar to the first two methods. Note that the distribution of membership probabilities assigned by the GMM are more sharply defined modal than those produced by the fuzzy method.

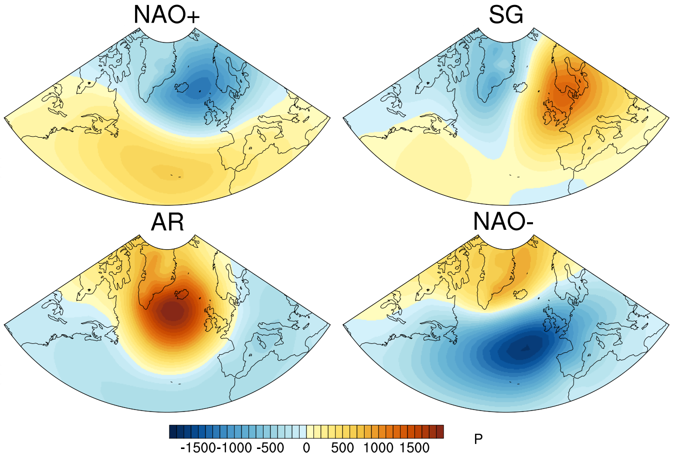


Figure B.7: Centres calculated using E-M optimised spherical Gaussian Mixture Models using winter (DJF) daily SLP anomalies.

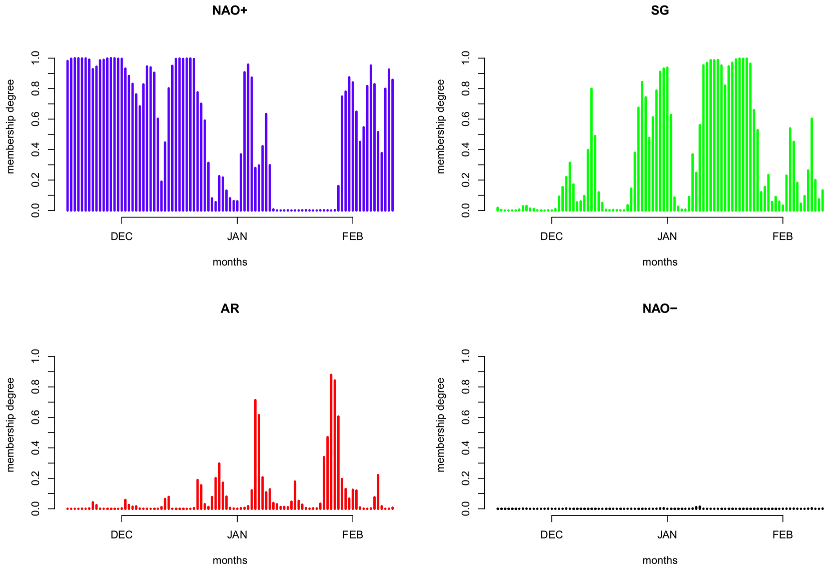


Figure B.8: Daily probability of membership for the 2012 winter (DJF) season for the E-M optimised spherical GMMs.

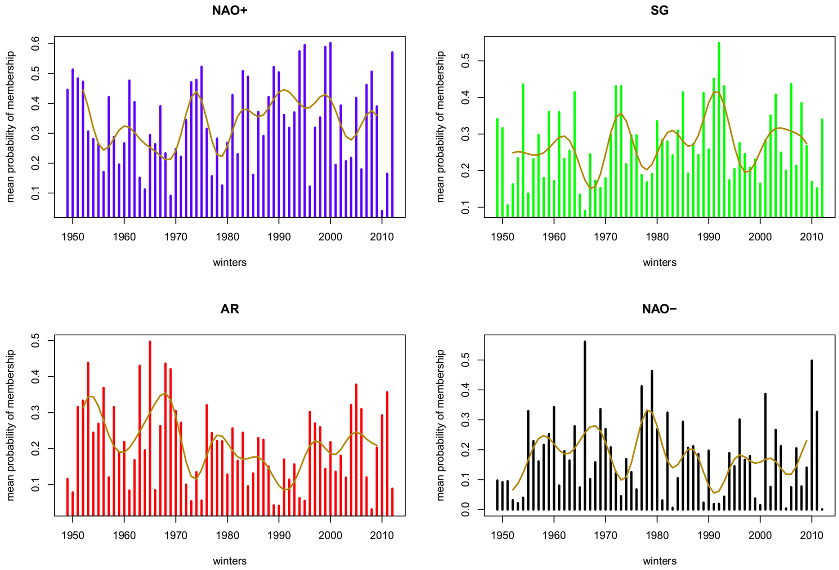


Figure B.9: Mean winter (DJF) probabilities of membership for the E-M optimised Gaussian Mixture Model SLP centres.

B.1.0.4 GMM (BIC)

As stated, all results presented so far are produced by fixing the number and shape of the clusters estimated. Here E-M optimised spherical GMMs are created for a wide variety of clusters and then the best model is determined by their BIC values, as described in Chapter 1. The result is a mixture of 27 clusters with centres shown in Figure B.10. Note that the previously identified modes all appear within this larger set. The AR appears as Mode-10, the SG-Dipole as Mode 13, NAO+ as Mode 20, and NAO- as Modes 26. This high number of clusters most likely does not represent the number of regimes in the data, but rather is a result of the model attempting to describe a highly non-linear data set; i.e., it requires a number of spherical clusters to describe a single non-spherical mode. The test is expanded by allowing for GMMs with non-spherical covariance matrices. Selecting the GMM with the best BIC value of various fitted models gives a GMM with three ellipsoidal clusters, shown in Figure B.11. The resulting centres appear as combinations of the standard regimes from previous examples; i.e, a merger of the NAO+ and SG Dipole and merger of the NAO- and Atlantic Ridge, as well as a new feature, depicting a zonally extended low over the middle of the region. The mean winter probability of membership for these clusters is shown in Figure B.12. The probability of membership to each cluster for each day of the winter of 2012 is shown in Figure B.13. Trends for the first two clusters are similar to observed NAO index behaviour, while the third cluster does not show any clear trend over the time period.

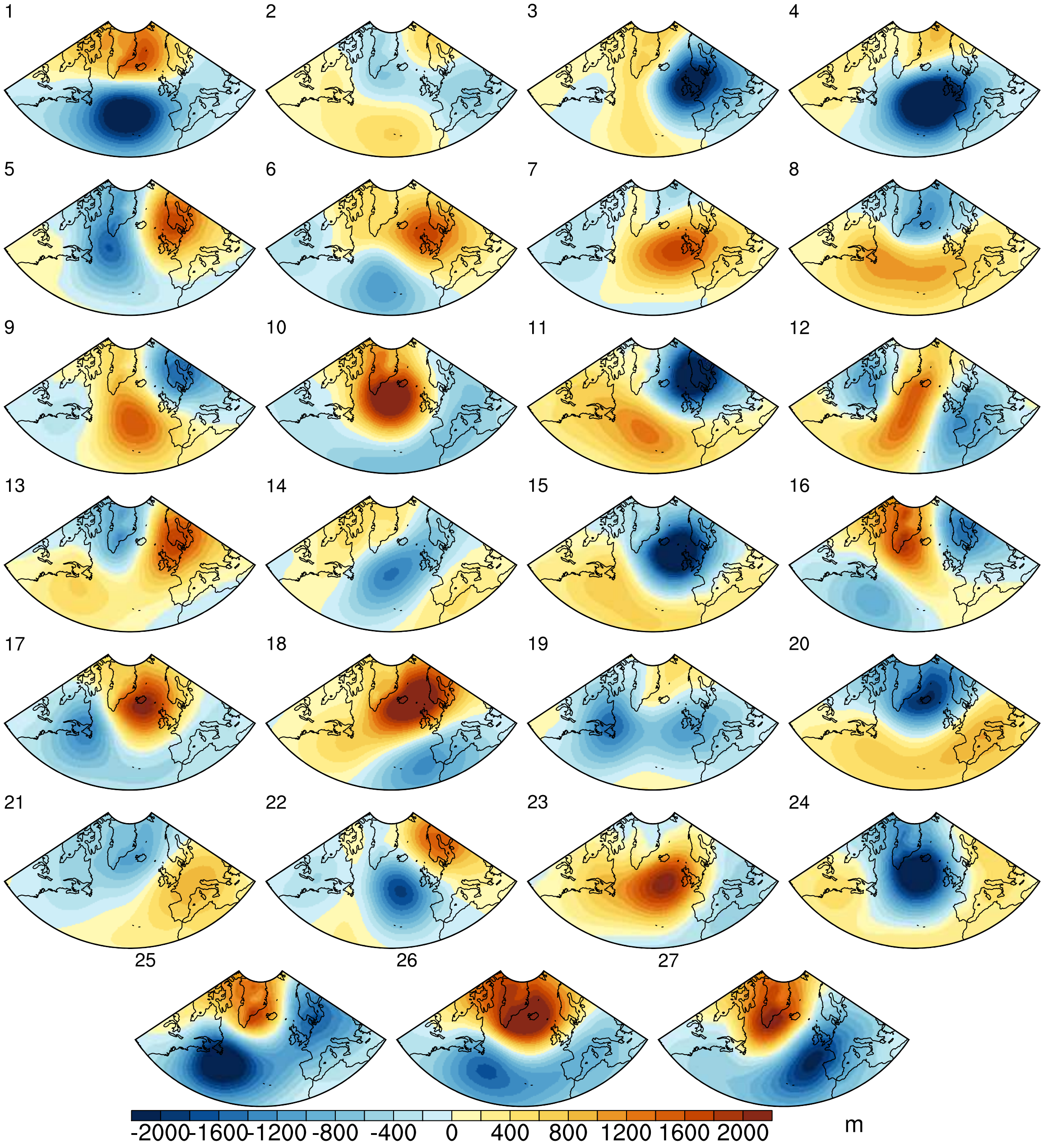


Figure B.10: Centres estimated using the BIC optimal number of spherical GMMs

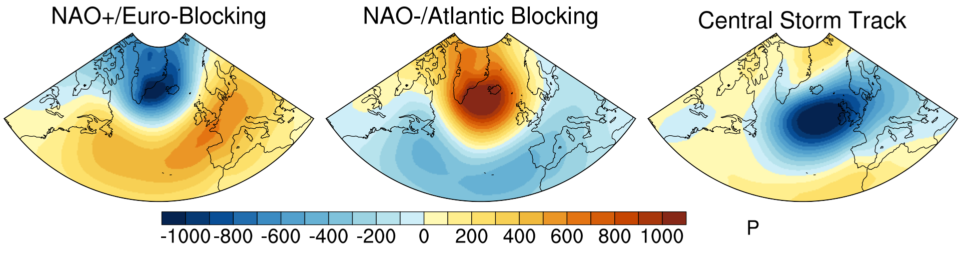


Figure B.11: Centres calculated using BIC selected E-M optimised GMMs using winter (DJF) daily SLP anomalies.

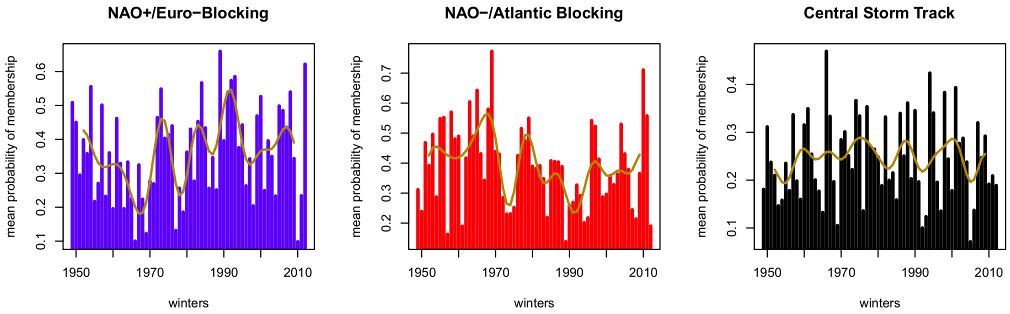


Figure B.12: Mean winter (DJF) probabilities of membership for the BIC selected E-M optimised spherical GMM SLP centres.

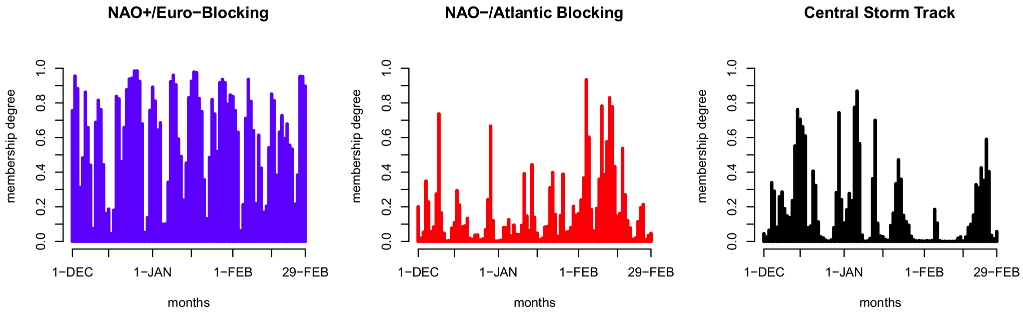


Figure B.13: Daily probability of membership for the 2012 winter (DJF) season for the BIC selected E-M optimised GMMs.

B.1.0.5 SOM

An analysis of the data field using SOMs is presented in Figure B.14. The results show the four typical patterns, with the modes with the highest frequency of occurrence giving the closest matches. Note that by this measure the NAO- regime more closely resembles the “central storm track” regime from the three cluster elliptical-GMM than that of the k-means NAO- regime.

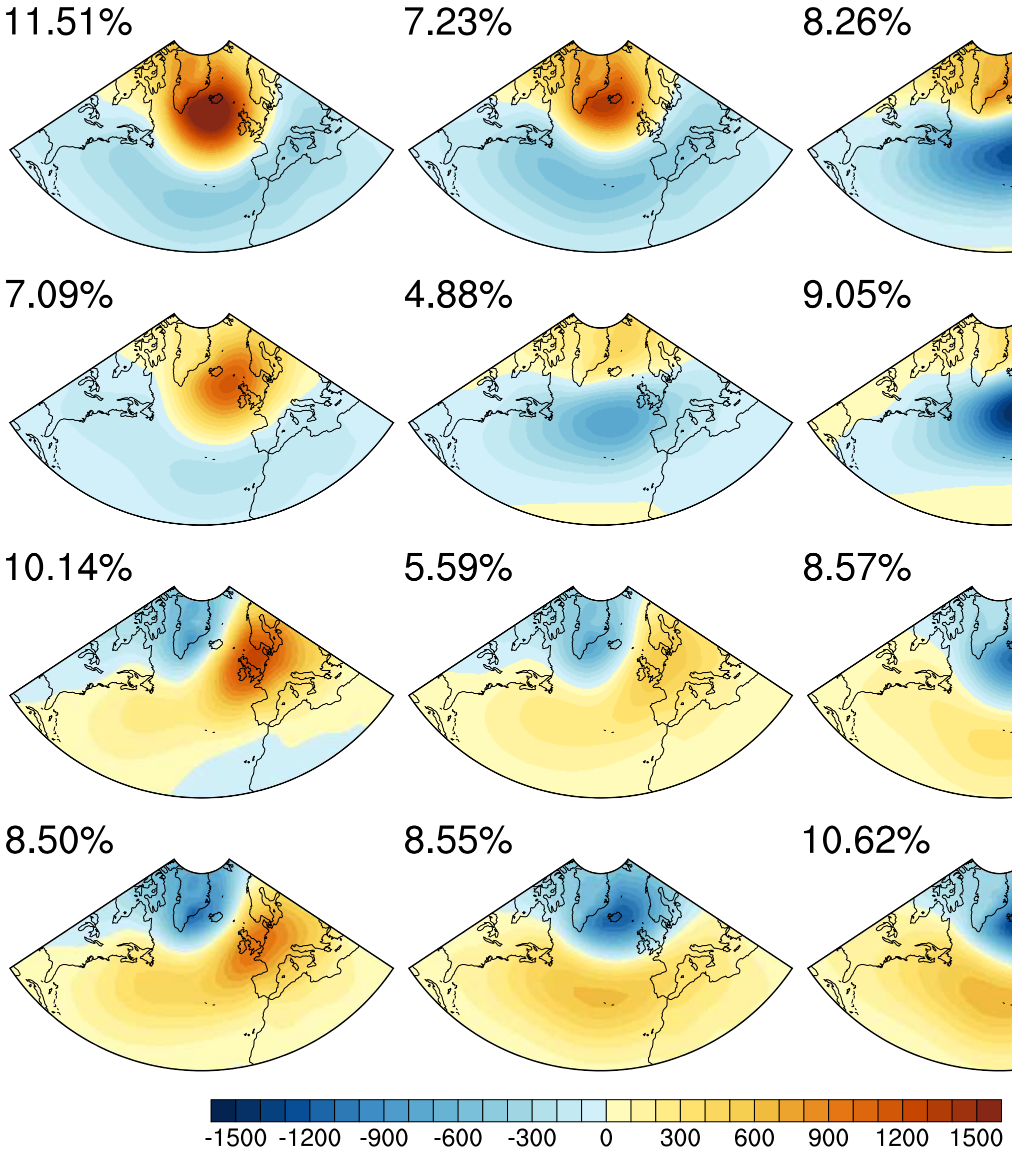


Figure B.14: SOM analysis of the daily field. Percentages give the frequency of occurrence of each mode within the data set.

Appendix C

Development of a Stochastic Weather Generator for the Sub-polar North Atlantic

C.1 Examiner Discussion

The following documents additional discussion of the topics of Chapter 4

1. *A discussion is given that is essentially a visual comparison between autocorrelation functions. That is okay, but a quantitative evaluations would have been better especially given the overall philosophy of the thesis. Also, when comparing functions, especially when the differences are not so large, it is easier to look at the difference (which I believe you do elsewhere in the thesis).*

    - As stated in Section 4.3.6 more quantitative analysis were performed and further are possible. However, since there is limited physical interpretation that can be made of the components of the PCA decomposition it was decided that long tables of lagged auto-regression coefficient ranges would not provide much of interest to the reader. Figure 4.14 is an example of where the quantitative information available is presented and interpreted. The decision to not use difference plots is in part to show the form of the observation distributions (which have some interesting subtle deviations from Gaussivity) and to what degree this is captured by the ensemble spread. This information is mostly lost when showing difference plots.

2. *In the conclusion you talk of subtle differences but can't this be quantified? How subtle are the differences and are you sure they are significant?*

    - This question refers to is a summary line in the concluding paragraph of the article, the differences referenced are described in the results section using quantitative methods. For some measures a visual representation was considered the most informative way to express the information (see previous comment), while others (such as the mixture model analysis) are described numerically. Significance is in part described by the ensemble spread of model realisations for both methods. There are many metrics where these do not overlap (e.g., PCs 4 and 6 in figures 4.7 and 4.8). Whether these differences 'matter' is hard to define in absolute terms for this project since the aims are descriptive rather than predictive. There does not exist a consensus on how exact the representation of observations must be for them to be 'good enough' to provide effective forcing fields for ocean models. A long term aim of the project is to test whether these more nuanced descriptions create a notable difference in the output of simulations where they are applied. In that context it would be possible to talk about the significance of differences in representation in a more concrete way. For the study presented, however, the question investigated and answered is: do the different methods give different representations, and if so, which representation is most similar to the target?

3. *How do the stochastic terms generate instabilities? The occurrence of instabilities would suggest a problem with the fit, because the actual residuals have bounded dynamics.*

    - Because the stochastic terms are drawn from Gaussian distributions they will occasionally produce comparatively large values. This is rarely a problem, but there are isolated occurrences where feed backs with the interaction terms expand these variations into physically unrealistic values (which continue to expand until the model crashes). As stated, the actual residuals should have bounded dynamics and so a truncated Gaussian distributions are more appropriate. This is effectively what is implemented in the Weather Generator.

4. *The Kravtsov and BANN models fit in this section are quite complicated. How many statistical parameters are fit in each of these models? This is needed to be able to estimate the robustness of the fit.*

   - The number of parameters for the Kravtsov models range from 22 to 41. The number of parameters for the BANN models are an order of magnitude more with values from 251 to 580. This difference is in part because the AIC pruning used for the Kravtsov models actively removes regression coefficients, thus decreasing the number of terms to be fitted. The ARD for the BANN models, however, reduces the influence (associated weights) of uninformative predictors, without actually removing model parameters. As stated in the article the all models are designed such that the initial number of free parameters is much lower than the available training data. Also, the experiment is to see whether advantage can be gained from the more complex model, rather to produce technical insights from an 'even' compassion between two operations.