A Mode-Based Metric for Evaluating Global Climate Models

Michael L. Kent

Supervisors: Bruce Hewitson & Christopher Jack

Department of Environmental and Geographical Science
University of Cape Town

Thesis Presented for the Degree of

Doctor of Philosophy

October 2017

Plagiarism Declaration

I know the meaning of plagiarism and declare that all of the work in this dissertation, save for that which is properly acknowledged, is my own.

Declaration of Free Licence

I, hereby:

(a) grant the University free license to reproduce the above thesis in whole or in part, for the purpose of research;

(b) declare that: (i) the above thesis is my own unaided work, both in conception and execution, and that apart from the normal guidance from my supervisor(s), I have received no assistance except as stated below; (ii) neither the substance nor any part of the thesis has been submitted in the past, or is being, or is to be submitted for a degree at this University or at any other University, except as stated below. I am now presenting the thesis for examination for the Degree of PhD.

Acknowledgements

I would like to thank my supervisors, Professor Bruce Hewitson and Dr. Christopher Jack for all the hard work that they put into me and my dissertation over the last few years. Much appreciated.

- The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.
- National Center for Atmospheric Research Staff (Eds). Last modified 20 Nov 2014. "The Climate Data Guide: ERA40." Retrieved from: https://climatedataguide.ucar.edu/climate-data/era40.
- NCEP Reanalysis Derived data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at: http://www.esrl.noaa.gov/psd/
- We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups (listed in section 5.2 of this dissertation) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

Abstract

Climate models are software tools that simulate the climate system and require evaluation to assess their skill, guide their development, and assist in selecting model simulations from among the many different ones available. There are a variety of methods and approaches that can be used to evaluate models. But there is no one best method and many possible and valid approaches exist.

Models contain inherent uncertainties which complicate their evaluation, and include limitations in the knowledge of climate process dynamics and structural errors in constructing the models. Similar to the multiplicity of methods for the evaluation of model simulations, there also exist many possible approaches to addressing these sources of uncertainty. The challenge with uncertainty, is the difficulty in disaggregating it from the underlying element of legitimate chaotic behaviour in complex systems.

In response, this dissertation is primarily one of methodological development to contribute to new ways of addressing the model evaluation challenge. The work defines and demonstrates a new evaluation method which complements the existing toolset. Specifically, the method defines a model performance metric that focuses on the extent to which a model is able to simulate global modes of climate variability (*modes*, e.g.: ENSO) evident in the observed climate data.

Modes are one aspect of the climate that can be evaluated and are fundamental to model skill. Therefore their credible simulation is a necessary (but not sufficient) condition to ensuring that models are producing the right result (appropriate variability on the range of spatial and temporal scales) for the right reason. By ranking models

by this metric of their skill in capturing fundamental global modes, poorly performing model simulations can be identified for potential exclusion (*discounted*). This metric therefore serves as a potential method to assist in the management of uncertainty when assessing multi-model data.

The method develops a novel application of Independent Component Analysis (ICA). ICA is used to find representations of modes in a record of the present day climate (represented by reanalysis data), and then their degree of manifestation in global models is assessed. Recognising the large volume of model data (highly autocorrelated in space and time) the technique includes a data reduction technique to facilitate the evaluation of multiple model simulations. The technique also includes a novel measure of variance to differentiate it from a similar technique (Principal Component Analysis), and offers an approach to improve the consistency of results (signals) when using an unmixing matrix initialized with random values.

As reanalysis data is itself a model product (constrained by observations), the performance metric is tested for its strength in discriminating modes by using two different reanalysis datasets and a dataset containing only Gaussian noise. The metric is found to perform predictably, and clearly demonstrates the ability to discriminate signal from noise when using geopotential height (GHT, 700mb and 500mb) and near surface air temperature data (TAS). The dependency of model performance on the variable measured by any metric can be a problem for model evaluation, as it introduces the choice of which variable should be measured to assess model performance. The ICA-based metric is found to be slightly less sensitive to a change in model rank between GHT (700mb) and TAS, compared to a similar novel variance metric (Fourier Distance) and a mean climate metric (bias). The ICA application is also found to produce plausible representations of modes (static maps), while a direct association to known modes is left for future work due to inherit complexities.

The plausibility, consistency, and rank sensitivity of the novel application of ICA, suggests it has value in assisting the evaluation of multi-model datasets and the ensemble members for any one model.

Glossary

AOGCM: coupled-Atmospheric Ocean Global Climate Model(s). A global model which strives to comprehensively capture processes, such as physical, chemical and biological processes. For more information see chapter 1.

Bias: The degree to which a result from a model disagrees with observational records is known as the model bias. A small bias is preferred (section 6.7.2.2).

Climate Index: Is a time series that can be derived from observation data and represents the behaviour of a mode. An example is the Niño 3.4 index (Trenberth and Stepaniak, 2001).

Component: Refers to either a part of a model with which simulates one part of the climate system (see section 1.1), or to the product of some methods for identifying modes (e.g.: Projection Pursuit, section 3.7.1).

Ensemble Member (*member*): A collection of a number of simulations are known as an ensemble, while a single result from the ensemble is known as an ensemble member. The two types of ensemble members, initial and perturbed, are discussed in section 2.3.

ESM: Earth System Model (abbreviated to *model*).

FD: Fourier Distance. An alternative metric to PCAP and ICAP that measures how closely the variance of the frequencies from a dataset match those of the reference dataset (section 6.7.2.1.

GHT: Geopotential Height (section 5.2).

ICA: Independent Component Analysis (chapter 4).

ICAP: Independent Component Analysis Performance metric (section 5.6.5).

Mode of Climate Variability (*mode*): This is an underlying space-time structure in climate data with preferred spatial patterns and temporal variations that help account for gross features in variance and in teleconnections (IPCC-2013a: Annex III: Glossary, Flato et al., 2013a).

Model: See ESM.

Pattern A pattern refers to any data structure that may be representative of the behaviour of the climate. Examples include signals, principal vectors, time series and spatial manifestations.

PC: Principal Component(s), a product of Principal Component Analysis. If principal components (P) are found by SVD then: $P_{M\times k} = X_{M\times n}V_{n\times k}$

PCA: Principal Component Analysis (section 3.4).

PCAP: Principal Component Analysis Performance metric (section 5.6.5).

PV: Principal Vector(s) (see section 4.2)

Rank Sensitivity: A simple measure of how many times datasets retained the same rank for a given metric when changing the variable used (section 6.7.2.1).

Reanalysis: Reanalysis are estimates of historical atmospheric temperature and wind or oceanographic temperature and current, and other quantities, created by processing past meteorological or oceanographic data using fixed state-of-the-art weather forecasting or ocean circulation models with data assimilation techniques. Using fixed data assimilation avoids effects from the changing analysis system that occur in operational analyses. Although continuity is improved, global reanalysis still suffer from changing coverage and biases in the observing systems (Flato et al., 2013a).

Reference Dataset: The dataset to which other datasets are compared to determine their ICAP and PCAP.

Signal: A product of Independent Component Analysis (section 3.7.2).

SST: Sea Surface Temperature (section 5.2).

SVD: Singular Value Decomposition (section 4.2).

TAS: Near Surface Air Temperature (section 5.2).

Teleconnection: A statistical association between climate variables at widely separated, geographically-fixed spatial locations. Teleconnections are caused by large spatial structures such as basin-wide coupled modes of ocean-atmosphere variability, Rossby wave-trains, mid-latitude jets and storm tracks, etc (IPCC-2013a: Annex III: Glossary, Flato et al., 2013a).

Contents

Chapter 1

Introduction

Global climate models are software tools that are used to simulate the climate. They are useful for providing information on the behaviour of the climate that theory alone cannot provide. A diverse range of climate models exist to simulate the climate, from energy balance models which describe the energy stored in the atmosphere as a function of incoming solar radiation and outgoing terrestrial radiation, to Earth System Models (*models*) which strive to comprehensively simulate the climate. This is achieved, for example, by modelling the interactions between the land, atmosphere and oceans, and by including the effects of the physical, chemical, and biological environments (IPCC-2013b, Flato et al., 2013b).

Model simulations (*simulations*) are evaluated to determine how good models are at simulating the climate. Evaluation is crucial for highlighting the shortcomings of models which can then be used to make corrections to them. This can be used to both improve the models and test current knowledge about the climate against model simulations, to see where models or theory needs improvement compared to a reference dataset. Similarly, users of simulations are dependent upon them meeting requirements before they can adopt them into their own applications. Evaluating model simulations can therefore serve as a means of ensuring that they meet the requirements of users. This interaction between models, the methods that are used to evaluate them, a reference dataset, and users of simulations can be seen in figure 1.1.

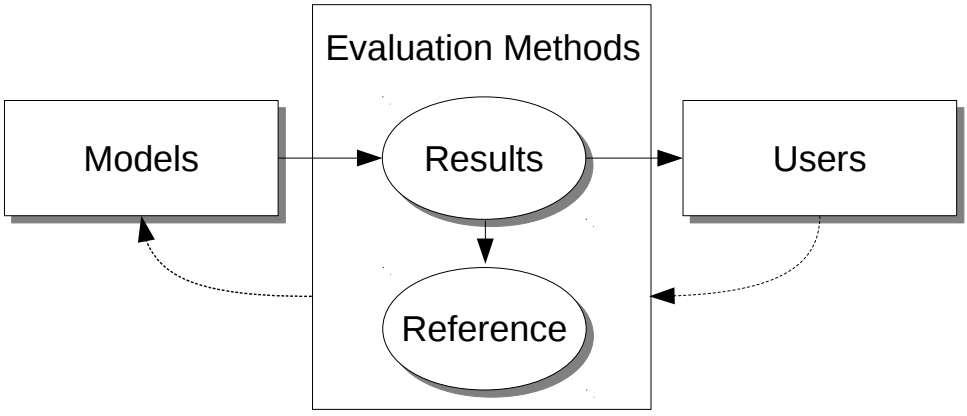

Figure 1.1: The model evaluation framework. The framework consists of the interaction between models that produce results, the evaluation methods that assess the models by comparing their results to reference data, and the users of the model results. Solid arrows show the flow of results from models to users, and the comparison of results to reference data. Dashed arrows indicate the influence of users on evaluation methods and evaluation methods on models.

1.1 Model Evaluation Framework

The evaluation of models is typically done by an evaluation method, which compares model results to reference data. The reference data serves as a standard or benchmark that the model result should ideally recreate. The better a model result compares to the reference data, the better the model that produced the result is deemed to have performed. Therefore the framework for evaluating a model consists of the model, model results, reference, evaluation methods, and users (figure 1.1). Changing any part of the framework can therefore change the measured performance of a model.

Various types of models exist, with each type designed to simulate the climate in a different manner. Generally, models offer a compromise between improved performance and an increase in computational demands. Perhaps one of the most well known model types, is the Coupled Atmospheric-Ocean Global Climate model (*AOGCM*). According to Edwards (2011) an AOGCM extends the global climate model (*GCM*), which has a "*... dynamical core, which simulates large-scale fluid motion using the primitive equations, and model physics, which*

simulates other climatologically significant physical processes such as radiative transfer, cloud formation, and convection".

An AOGCM extends a GCM by coupling it to an ocean model component, which allows interactions to take place between the atmosphere and ocean. Coupling additional components to GCM (and AOGCM) creates a model type termed the Earth Systems Model (*ESM*). The ESM allows for a more comprehensive simulation of the climate but requires more computational resources. As the availability of computational resources have increased over time, additional model components have been included in models. This can be seen in figure 1.2 from Washington et al. (2009, figure 3).

Fundamentally, models only simulate the behaviour of the climate and therefore their results can never be perfect in representing a reference. One reason for this can be seen in the inherent chaotic nature the climate (e.g.: Slingo and Palmer (2011)). Similarly, Knutti (2010) state that improving computational resources alone will be insufficient to fully address the limitations with models. Limitations with models are further discussed in chapter 2.

With respect to model simulations, the time period that it has been performed over has implications for both the assumptions that can be drawn from it and the corresponding references that can be used to evaluate them. Model simulations generally fall into two categories: simulations of the past climate for which there are records of the climate, and the future climate for which there are no records.

For the period of the past climate, there is the hindcast model result type. The IPCC Fifth Assessment Annex (IPCC-2013a: Annex III: Glossary, Flato et al., 2013a) defines a hindcast as "*A forecast made for a period in the past using only information available before the beginning of the forecast. A sequence of hindcasts can be used to calibrate the forecast system [e.g.: model] and/or provide a measure of the average skill that the forecast system has exhibited in the past as a guide to the skill that might be expected in the future.*" For the hindcast model result, a palaeoclimatological record (e.g. Masson-Delmotte et al. (2013)), or observational record (e.g.: Trenberth (2008)) may be appropriate depending

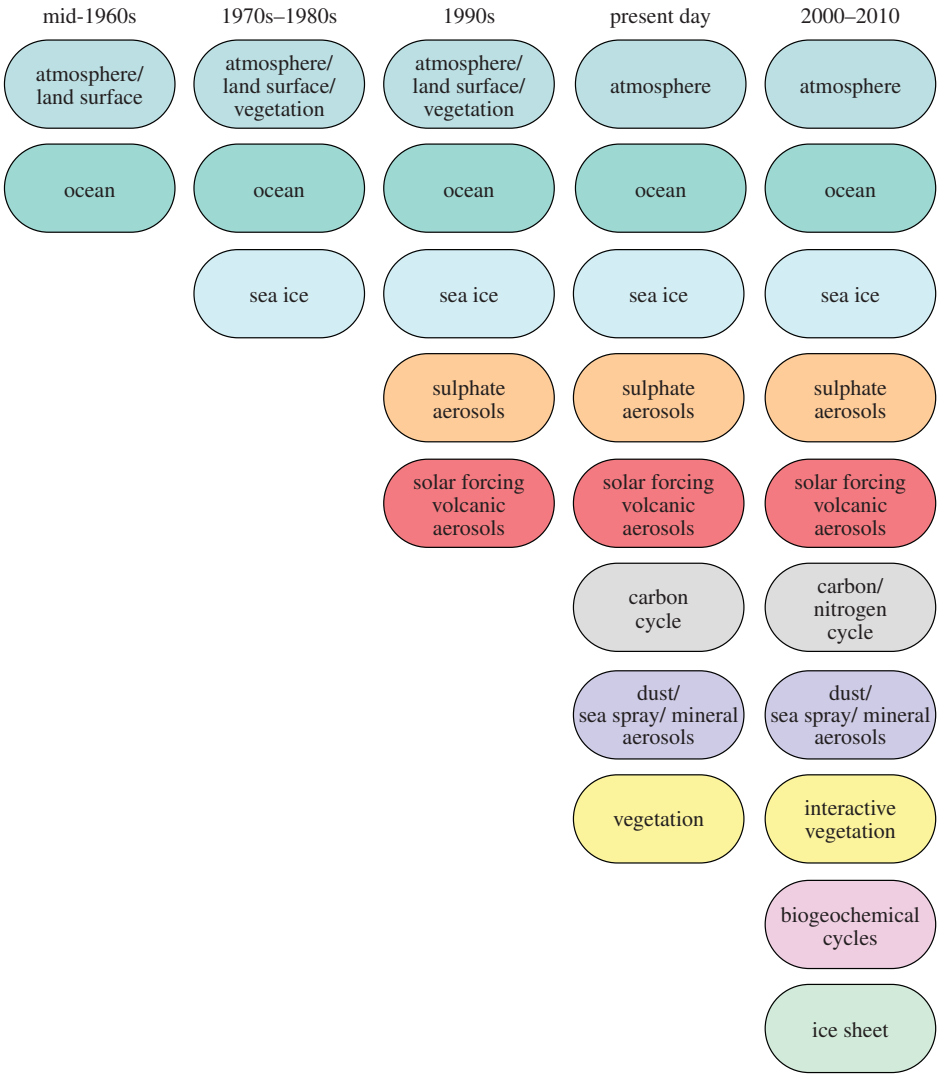

Figure 1.2: The model components in use at different times, from Washington et al. (2009, figure 3). Each colour tracks a specific component through time.

on the time period simulated. Reanalysis data may also be used as a reference, as it represents a combination of model and observations for a given time period (Overpeck et al., 2011). One benefit to using reanalysis data is given by Kalnay et al. (1996), where reanalysis data can be useful for creating a consistent dataset over the same period when observational assimilation methods changed. However, Langland et al. (2008) show that there are also limitations in reanalysis data. These include variations in the quality in satellite observations, limitations in the

methods used to assimilate data, and a sparse observations in some geographical regions which can lead to a reduction in the reliability of the reanalysis data in the corresponding regions.

For the future climate there are the climate prediction and climate projection model result types. IPCC-2013a provides the definition of a climate prediction as "*... the result of an attempt to produce (starting from a particular state of the climate system) an estimate of the actual evolution of the climate in the future, for example, at seasonal, inter-annual or decadal time scale...*". A climate projection is similar to climate prediction but uses an emission scenario to guide the simulation. According to IPCC-2013a, an emission scenario is a "*A plausible representation of the future development of emissions of substances that are potentially radiatively active (e.g., greenhouse gases, aerosols) based on a coherent and internally consistent set of assumptions about driving forces (such as demographic and socioeconomic development, technological change) and their key relationships. Concentration scenarios, derived from emission scenarios, are used as input to a climate model to compute climate projections...*". They also define a climate projection as "*... the simulated response of the climate system to a scenario of future emission or concentration of greenhouse gases and aerosols, generally derived using climate models. Climate projections are distinguished from climate predictions by their dependence on the emission/concentration/radiative forcing scenario used, which is in turn based on assumptions concerning, for example, future socio-economic and technological developments that may or may not be realized.*" As no records exist for the future climate, the results produced from other models can be used as the reference. Using another model as a reference is further discussed within chapter 2.

When evaluating a model, there are many different approaches that are available: For example, the examination of an single component (e.g.: ocean component in Gent et al. (1998)), the overall performance of a model (e.g.: mean climate in Jury et al. (2015)), or by evaluating results from multiple models in order to overcome the limitations within each individual model (see chapter 2). For details on the available evaluation methods, the reader is referred to the IPCC Fifth Assessment Report chapter on model evaluation (IPCC-2013b, Flato et al., 2013b).

1.2 User Requirements of Model Results

In addition to assisting with improving model performance, evaluation methods can also be used to assist users in determining if they should utilize a particular model result. The work of Leary et al. (2009) highlights the need to first interpret model results into user relevant information by adopting a participatory approach that involves both the producers and the users of model results. Before the information can be used, they state that it needs to be credible (the outcome portrays a realistic climate), defensible ("*in that there is a clear process-based understanding of the response of the physical and social systems to climate and other pressures*"), and actionable (the information is user relevant, robust, and is understood by the users).

Evaluation methods may therefore be used and developed to assist with answering whether model results are defensible and credible. The needs of users may then also help drive the development of evaluation methods that can better answer these two questions. Evaluation methods are less directly applicable when answering the question of whether climate information is actionable. Cash et al. (2003) further adds to the three requirements of climate information by suggesting that the institution or group responsible for creating the information also needs legitimacy. In their work, a legitimate institution is one which produces climate information in an unbiased manner and appropriately handles differences in shareholder values.

While the assessment of models cannot directly inform users on action what action is appropriate, designing an assessment with user requirement in mind, may increase its utility and ultimately its uptake by the broader community. Deficiency in any of these areas: defensible, credible, actionable information, or in the legitimacy of an institution, may ultimately hinder the uptake of climate information by users. An example of this can be seen in the work of Tang and Dessai (2012), who show some of the challenges when attempting to meet these requirements. For instance, a wide range of users may make it difficult to generalize information to a large number of users with differing needs. Customizing the information to subsets of users may be one solution, but maintaining general

consistency of action between sub-groups may then create a new challenge.

1.3 Complications with Meeting User Requirements

However, meeting the information requirements of users poses a challenge due to limitations associated with models. These limitations are known as uncertainties, and range from compromises made between model complexity and computational demands, such as cloud parametrisations, to uncertainties regarding the design of the models such as deciding which components of the climate are to be simulated by a model. Rather than creating a clearly identifiable feature in the data, the way in which uncertainties manifest in model data can be difficult to quantify. For instance, when there are multiple different but valid approaches to parametrizing clouds. Selecting one method in preference of another may change the final result, but the extent of the change in parametrization may not be known prior to running a model with the parametrization. This complicates the interpretation of data, which may in turn hamper users who utilise the interpreted climate data as part of their decision making process.

Utilising improved computational capacity has allowed for the reduction of some uncertainties in models, such as clouds which can be explicitly simulated rather than approximated (e.g.: Tomita et al. (2005)). While this does provide a way forward in reducing uncertainty, other types of more fundamental uncertainties which are linked to model design, may not easily be addressed in this manner. Working in conjunction with improving models to reduce uncertainties, are approaches that utilise data from multiple models. This is perhaps best seen in the multi-model mean of climate data, which has been shown to be able to out perform the individual datasets that contribute to it with respect to records of the observed climate (Randall et al., 2007).

Motivation for superior performance of the multi-model mean, is attributed to the cancellation of random errors present in models using the mean of the results. The multi-model mean is not the final answer to reducing uncertainties though,

as there are still concerns with regards to combining models. The combined result may not be representative of the climate, as it may contain model results that agree on the region that undergoes change, but not on the sign of that change (Knutti et al., 2010). Therefore the combined result may not represent the area of change correctly. There is no definite clear way to proceed with reducing all the uncertainties in models, and this concern is compounded when considering the future climate, where there is no climate records with which to compare model results against. Current approaches to address the uncertainties have taken an exploratory nature.

One of the recent approaches to assist in quantify the affects of model uncertainties is to combine model results according to the extent of performance that individual models exhibit. The aim of these approaches is to make the best use of models which perform well, and lessen the effects of poorly performing models on the combined result. This is in contrast to giving each model the same contribution to the combined result, as in the case of the multi-model mean. However, the difficulty with these approaches is deciding how to measure model performance, as there is no one best method for evaluating models (Gleckler et al., 2008).

A variant of this type of approach may be to use an evaluation method that assists in the interpretation of data for users. For example, Leary et al. (2009) suggest that before data can be considered as information for users, it should be credible (section 1.2). This would include knowing that the modes of climate variability (e.g.: El Niño Southern Oscillation, ENSO) are related to the present day climate. The assumption with this approach, is that models which do not simulate the present day climate correctly are also less likely to simulate the future change in climate correctly. Evaluating models in terms of their modes has also been suggested to ensure that models are simulating the correct climate for the correct reason (Tebaldi and Knutti, 2007).

The correct simulation of modes is also important to ensure that other aspects of the climate are also captured correctly by models, such as rainfall patterns. An example of this can be seen in Hart et al. (2013) who show that the Madden-Julian Oscillation can affect South African rainfall by modulating tropical temperate troughs. Another example is provided by Smith and Chandler (2010). As part

of their assessment of global models to capture Australian rainfall patterns, they evaluate how well the models capture ENSO.

Methodologies used to find representations of modes in data generally operate by using a technique, like a cluster analysis method to find patterns in data. The patterns are then associated to known modes using expert analysis. Successful association requires that a pattern (e.g.: geographic manifestation) has some property (e.g.: geographical region) in common with a more abstract understanding of a mode that is described in the literature and often supported by a corresponding data product (e.g.: Niño 3.4 index (Trenberth and Stepaniak, 2001)[1]). Upon the successful association of patterns to modes, the patterns can then be said to be representations of modes. An extension to these methodologies to multiple model datasets could serve as a means of evaluating models according to how well they have simulated modes.

This dissertation defines and demonstrates an original performance metric which aims to help create credible information for users by evaluating multiple model results according to how well they have simulated modes of climate variability.

1.4 Thesis Aims and Objectives

The aim of this dissertation is:

To design and demonstrate a methodology which finds representations of modes in a reanalysis dataset, and then uses a model performance metric to measure the extent to which model results contain those representations.

The aim is to be realised by developing a model performance metric which:

1. Ranks models according to how each model result contains the representations of modes that are found in reanalysis data.
2. Is applicable to multiple datasets, a task which is currently not feasible

[1]Niño 3.4 Index: http:/www.cgd.ucar.edu/cas/catalog/climind/TNI_N34/index.html#Sec5

using current approaches that evaluate results according to modes of climate variability.

3. Ranks artificial datasets appropriately: noise should be ranked poorly, while an alternate reanalysis dataset should be ranked highly.

Chapter 2 discusses the uncertainties associated with models which helps drive the need to have different ways of evaluating models. To potentially address some of the uncertainties, different clustering techniques are presented in chapter 3. In chapter 4, the details of one such clustering technique, Independent Component Analysis (*ICA*) are examined. Chapter 5 outlines the methodology for finding representations of modes using ICA and presents the performance metric. The performance metric is applied to various types of global data and the results compared to those from similar metrics to determine how the performance metric behaves relative to expectations in Chapter 6. The focus of this chapter and the dissertation as a whole, is on global climate model data as further regional application is complex. Chapter 7 concludes the work and contains recommendations for further research.

Chapter 2

Addressing Model Uncertainty

2.1 Introduction

While an improved understanding of the climate system and an increase in computational capacity has enabled more comprehensive climate models to be run (Edwards, 2011), limitations still exist with models and their usage. These limitations are broadly known as *uncertainties*, and can complicate the analysis of model results. This has consequences for assessing model performance and for translating model results into information with users.

The topic of uncertainty is broad and it has many varieties. Enserink et al. (2013) provide an in-depth view on the topic and how different users perceive it. As this dissertation is concerned with demonstrating a new model performance metric, only the uncertainties associated with models are discussed.

The types of uncertainties associated with models are discussed within section 2.2. Following which, section 2.3 reviews pioneering methods for addressing some of the uncertainties. Lastly, section 2.4 discusses some of the remaining limitations and challenges associated with evaluating models in light of the uncertainties present in them.

2.2 Types of Model Uncertainty

The types of uncertainties associated with models are as follows:

- **Initial Condition Uncertainty** is the range of possible initial conditions that are valid for use in a model prior to simulating the climate. Although different values may result in a different climates, this effect is seen to be less important for longer time scales compared to other types of uncertainty (Tebaldi and Knutti, 2007).

- **Boundary Condition Uncertainty** "*... is introduced if datasets are used to replace what in reality is an interactive part of the system, e.g. if sea surface temperature and sea ice cover are prescribed in an atmosphere-only model, or if radiative forcing (e.g. changes in solar insolation, changes in atmospheric concentrations of greenhouse gases) is prescribed over time.*" Tebaldi and Knutti (2007)

- **Parameter Uncertainty** is introduced during the development of a model. During this phase there are a number of compromises that have to be made with regards to the degree to which a model will be able to fully simulate the climate. These compromises are largely dependent upon computational limits. To work around these, local spatial scale processes are approximated at a larger global scale using parameters. An example of this is discussed in Randall et al. (2003), where the effect of small cloud related processes are represented at a more global scale using parameterizations.

- **Structural Uncertainty** represents the choices that are made with regards to the overall design of a model (Tebaldi and Knutti, 2007). This includes which components and grid resolutions will be used by a model. Importantly, Tebaldi and Knutti (2007) state that this type of uncertainty is unlikely to be overcome by perturbing model parameters. Stocker et al. (2010) add that processes that are not fully understood also contribute to this uncertainty.

Interestingly, the term uncertainty can also be used to describe the distribution of multiple model results. When the term is used to describe the distribution of

multiple model results, the term distribution can also be substituted with *range* or *spread*. According to Stocker et al. (2001), the distribution of model results is not a direct measure of one or more types of model uncertainty, but can rather be used to better classify model uncertainties. The analysis of multiple model results is further discussed in section 2.3.

2.3 Combining Multiple Results

Due to the uncertainties associated with the results taken from a single model, various approaches have been developed to address them. These approaches aim to explore and quantify some types of uncertainties by combining multiple results from either one model or multiple models.

In order to measure the magnitude of the effect that the uncertainties can have on the results, models can be run while changing parameters. A collection of a number of runs are known as an *ensemble*, while a single result from the ensemble is known as a *member*. While the term *member* is adopted from statistics, it does not necessarily adhere to the same sampling assumptions. Ensembles generally fall into two groups, perturbed physics ensembles and multi-model ensembles.

Perturbed physics ensembles can be used to sample parameter uncertainty. This is where a parameter is changed (perturbed) in a single model to better explore the set of possible results (Stocker et al., 2010). An example of this can be seen in Murphy et al. (2007). Multi-model ensembles can be used to sample initial condition uncertainty. According to Stocker et al. (2010), this is where multiple models are run with many different but valid initial conditions, as only initial condition uncertainty can be captured from one model. While a multi-model ensemble may capture the effects of using some different parameters, they argue that it does not systematically sample parameter uncertainty.

Structural and boundary uncertainties can also be sampled to some degree using the multi-model ensemble approach. An example can be seen in the set of inter-model comparison projects, like CMIP5 (Taylor et al., 2012). The degree of the

sampling will ultimately depend on the models, observational datasets, and components that are incorporated within the project. A more detailed discussion on the interpretation of an ensemble distribution is presented in section 2.4.2.

Perhaps the best known method of combining the results from multiple models is the multi-model mean. The multi-model mean is the mean result taken from an ensemble, and the ensemble includes different models using different initial conditions. The value of using the multi-model mean, is that it has been found to sometimes outperform individual members when it comes to its agreement with the observed climate due to the cancellation of random model biases (Randall et al., 2007). Knutti et al. (2010) discuss two challenges with regards to using the multi-model mean. The first is that the averaged result may not be physically valid, as the complex relationships between different variables may be lost by averaging their results. Secondly, they discuss the instance where model results could differ in sign for a particular region. In the case of precipitation, the average of the results from models with the same region of change, but different sign, could result in the area of change in the multi-model mean being under represented. Despite these challenges, the improved performance of the multi-model mean has spurred the creation of alternative weighting methods. These methods no longer assign an equal weighting to each model, rather they determine the weighting of a model based on a measure of its performance in an attempt to improve upon the multi-model mean. Generally, there are three schools of pioneering thought exploring the weighting of models and the combination of their results.

2.3.1 Weight by Performance and Convergence

The first school of thought proposes that models can be weighted according to their agreement with both the observational record (model performance), and the degree to which their change (future less present climate) differs from the group tendency of change (model convergence). This approach is formalized by Giorgi and Mearns (2002) and is known as the Reliability Ensemble Averaging method (REA). The rational for the performance aspect of the weighting, is that models that simulate the present day climate better than other models (small biases),

should also have more reliable future climates, and therefore should be more favourably weighted. The convergence aspect of the weighting seeks to compare model results to the behaviour of the group in the future as defined by an iterative solution. Each model result is then weighted according to its performance and convergence value.

The REA method was shown to be able to reduce the spread in the future climate over many of the investigated regions. The reason for the increase in spread in some regions is generally attributed to the models having poor performance (large biases) in those regions. Interestingly, they find that the present day biases are much greater than the spread in the simulated future climate change. This is interesting as the models are stated by them to potentially be tuned for the present day climate, and so in theory they should be less reliable when simulating the future climate. As the overall reliability of the models depends partially on the biases of the models, they recommend further research into improving the simulation of present day regional climates.

Bayesian Networks provide another method for combining the output from multiple models. A Bayesian network captures the relationships between variables (not fixed to be temperature, precipitation, etc.), and uses their initial (prior) distributions and inter-relationships to predict their future (posterior) distributions. The process of prediction is referred to as inference, and it is conducted according to Bayes' theorem. For more detail on Bayesian Networks see Korb and Nicholson (2004). Tebaldi et al. (2005) employ Bayesian Networks to quantify the uncertainty in the present and future climate. Their approach adopts the same bias and convergence criteria as the REA approach but it also allows for the construction of a distribution around the uncertainty in the change in temperature. They state that a benefit of their approach is that it may reveal multi-modal characteristics and tails in the probability distribution of the future climate. Their approach could then offer additional insights into the agreement amongst models.

Giorgi and Coppola (2010) assesses the validity of the dependency assumption between the bias and convergence criteria made in the REA method. The importance of this work lies in the results. For temperature data, they find that

the behaviour of future climate change is not dependent upon present day model biases. They propose that this may limit the application of the REA method and its Bayesian variant, which assume that there is relationship between the present and future climate. For precipitation they find that only around thirty percent of their regional simulated future climate results are dependent upon their regional biases. The implication for their work is that at the regional scale, improving the mean simulation of the present day climate appears to be of little use to predict the future climate. They suggest further investigation into understanding the behaviour of processes that can give rise to potentially similar biases while still producing different future climates.

The process of weighting models by observational datasets may be complicated by substantial differences that can exist between them. For example Sylla et al. (2012). These differences would require additional research to fully understand their impact on model weightings, and how to choose between the datasets despite their differences.

2.3.2 Weight Models by Inter-model Similarities

Räisänen (2007) outlines a potential concern with the REA method. As the future climate may fall outside the range of model projections, down-weighting an outlier model may not be justifiable. Rather an alternative exploratory approach is proposed by Räisänen et al. (2010). This school of thought proposes that models can be weighted based on inter-model similarity in both the observed climate and future climate periods.

The inter-model similarity method is based on finding a relationship between models simulations of the present and future climate. The weighting is based on the strength of the relationship found, and the stronger it is, the greater the difference in weights between the better and worse performing models. Specifically, the weighting is established by using linear regression. For example, the relationships in present climate between models and observations, and the inter-model differences in future and present day climate.

To test their weighting method, a procedure of model cross-validation was used. The cross-validation functions by removing a model from the set, and determining how well the results from the remainder of the models compare to those of the removed model. This procedure was repeated for all the models using weighted and unweighed results. Generally however, only a small decrease was seen in cross-validation error between the weighted and unweighed results, and the establishment of any inter-model relationship depends on a justifiably large correlation between the predictor and predictand of the linear regression.

Räisänen and Ylhäisi (2011) apply this method in a probabilistic context in order to improve on the deterministic weighting method. Distributions were constructed from the weighted and unweighted projections, and were compared using cross-validation. The results though, show little improvement over the original deterministic approach.

2.3.3 Discounting Model Results

Instead of using non-uniform weights to combine model results, the uncertainty (e.g.: spread) in the future climate may be reduced using a method known as model discounting. This method is analogous to the weighting methods, except that the goal is to determine the worst performing models using the weights. The worst performing models can then be removed from further use, and any reduction in spread can therefore be attributed to the removal of a poorly performing model. This process of removing poorly performing models is known as *discounting* models.

In the case of Kirono and Kent (2011), they discount models based on their simulation of present day regional drought intensity using rainfall and potential evapotranspiration. They calculate the mean climatology (spatial patterns, model bias), inter-annual variability, and long-term trends of the models and compare them to observations. They find that by discounting the poorly performing models, more substantial changes in the mean area affected within some of their regions can be seen. They also show that in some regions, the spread can be reduced when only using the top few ranked models.

Similarly, Smith and Chandler (2010) discount models according to their inability to capture both rainfall patterns and the El Niño Southern Oscillation in the Murray Darling Basin within south east Australia. The retained model results are not combined in this work, but the mean and standard deviations of them are compared to those from the original set of models. They show that by only using the first few models they can reduce the amount of spread in their results. They only use present day climate to discount models, because if poorly performing models end up performing well in simulating the future climate, then it would imply that there is a hidden measure of model performance. They argue that the consequence of this is that in theory a model functioning as a random number generator could not have its results discounted despite its poor performance with the present day climate.

The examples presented here do not focus on recombining the subset of retained model results, rather they focus on finding and providing a subset of retained models for further use by users. Also, the reductions in spread are tested and found not to be due to the smaller number of models used.

Knutti (2010) state that discounting models may be one of the lesser disputable ways of reducing the spread of model results when compared to seeking agreement amongst models. They state that agreement amongst models should be viewed with caution, as there is potential for agreement due to tuning (section 2.4.2) or peer pressure rather than on a greater understanding of the climate system.

2.4 Limitations and Challenges

General limitations to weighting methods are discussed along with the properties of multi-model datasets upon which they are dependent.

2.4.1 Weighting Methods

One of the first decisions that has to be made before models can be weighted, is to determine which weighting method should be used. This is a difficult question, as

there is no one method which is able to comprehensibly evaluate the performance of models (Gleckler et al., 2008).

The weighted results may also be sensitive to the selected weighting method and variable (Chandler and Bates, 2011). So just by changing the weighting method or variable, the combined set of results may change as well. The consequence of this is that there has to be strong justification for any method and variable chosen, which may be difficult to do as there is no one comprehensive method to evaluate models yet.

Even if a method of weighting is chosen and a combined set of results obtained, there is no guarantee that the set of better performing model results will necessarily outperform a randomly selected set of results. This is shown by Pierce et al. (2009), who find no large differences in their study when selecting skilful models, and when selecting models at random. Similarly, Weigel et al. (2010) reason that any measure of model performance should further be accompanied by an understanding of model error and noise. Without taking these into account, the weighted result may end up having less skill than the unweighted result.

2.4.2 Multiple Model Results

There are a number of limitations involved in the construction and interpretation of a multi-model ensemble. These are best captured by the phrase "Ensemble of Opportunity" by Tebaldi and Knutti (2007). They stress that ensembles are generally only contributed to by interested modelling communities which are themselves limited by available funding and computational resources. This results in an ensemble which is limited in which models it contains, and therefore what uncertainties the ensemble is able to adequately sample.

Tuning of Models

Model tuning is one such example of a limitation in model construction. When models have been constructed, the parameters of the models can be modified in

such a way as to improve the performance of the model with respect to the observed climate. The improvement in performance itself is not the problem, rather as Tebaldi and Knutti (2007) discuss, it is the reasoning behind the improved performance that may be of concern.

An observational dataset is normally used as the reference or goal to which the performance of a model is tuned towards. However if that same dataset is also used to later validate the model, then the resulting performance of the model may be artificially good.

A consequence of this is that the adjusted parameters may not be strongly related to the underlying problem causing the poor performance of a model. To elaborate on this, Tebaldi and Knutti (2007) pose a hypothetical example where a model which is poorly simulating the temperature of a region can be improved by modifying the albedo of the dominant vegetation type in the region. Even so, the actual underlying cause of the poor performance of the model may not be in the albedo, but rather in circulation patterns. The tuning of the albedo parameter therefore improves the performance of the model but for the wrong reason.

Common Structural Error and Ensemble Distribution Interpretations

Knutti (2008) define structural error as the existence of an irreducible difference between a model simulation and the observational climate, and state that it cannot be reduced by changing the values of parameters in parametrizations or by tuning a model. Structural uncertainty may also be a source of model error, as the exclusion of a processes from a model may result in it obtaining a poor fit with observations (section 2.2). A consequence of structural error, is that models in an ensemble may have common errors which may result in common biases Knutti et al. (2010).

Common biases can arise in models which share code, as they are likely to produce more similar results than had they not shared the code. Cases of shared code include those models which are newer versions of older models, and models

which share components (Knutti et al., 2010). Yussouf et al. (2004) demonstrate structural error when they show that ensemble members can have a bias towards the models which were used to create them rather than being equally likely. Figure 2.1 also shows how the development of models can include other models, where Edwards (2011) state that most modelling groups have begun their models based on a previous model from a different modelling group. The challenge with common structural error though, is to determine the magnitude of the bias and how to take it into account when combining model results and interpreting their distributions.

The combination of model results in the context of common structural error has definite implications for the first two schools of weighting methods reviewed in section 2.3. Structural error could potentially cause the multi-model mean in the REA method to be closer to those models which share components, making those models appear to be performing better but for the wrong reason. Likewise in the case of cross-validation during the inter-model similarity approach, those models with similar structural error could potentially predict each other better than models without the shared components. This again could cause the results of the cross-validation to favour those models with common structural error. It is unclear how directly structural error could effect the model discounting approach, though it may play a more implicit role in determining which models are retained and which ones are rejected.

The challenge of how to account for structural error manifests itself in the interpretation of an ensemble distribution. The distribution can no longer be seen as being produced by independent models and also has to account for various uncertainties (section 2.2). This therefore requires a detailed understanding of an ensemble to ensure that any interpretations that are draw from its distribution, reflect the structural errors and uncertainties associated with it.

Taking up this challenge is Annan and Hargreaves (2010), who discuss two different views on the interpretation of members. They use the term *truth* to refer to results from an ideal and perfect simulation of the climate. In their first view, the

[2]Image also online: http://pne.people.si.umich.edu/vastmachine/agcm.html

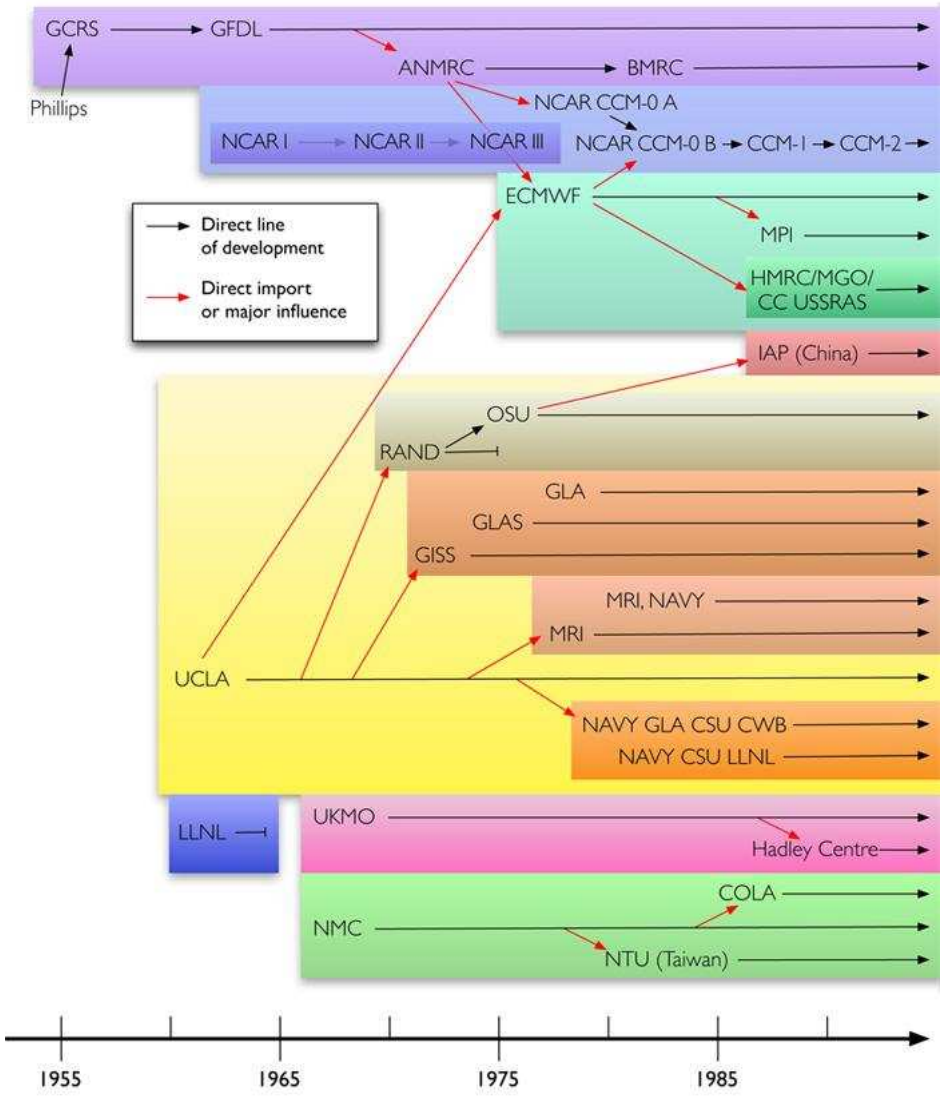

Figure 2.1: Shows the relationships between some models, which may lead to structural errors. Original figure from Edwards (2010)[2].

members are seen as being able to simulate the climate, but with some (random and structural) error. So the ensemble distribution is centred around the truth, and members that contain more error are situated further from the truth.

In their second view, the members are seen as being *indistinguishable* from the truth. Each member represents an equally likely simulation of the climate, but with a different sampling of an initial condition or parametrisation. Therefore if

a result was randomly taken from this distribution, it would be equivalent to the truth, or in other words, indistinguishable from it.

Sanderson and Knutti (2012) expand on the discussion by examining when each interpretation could be applicable. In the case of the first view of members centered around truth with some error, additional members should reduce the uncertainty in a consensus of results (e.g.: multi-model mean) through the cancellation of random model errors. Therefore the first view may be applicable for simulations of the present day climate, where members can be combined into a multi-model mean which better agrees with observations the more members are included. However, they also state that common structural error will prevent perfect precision from occurring with respect to observations.

Both views could also be relevant for the future climate. This fundamentally depends on whether there is a relationship between present day simulations of the climate and the future climate. A non-negligible relationship between them would indicate that models could still be constrained by the present day climate to some degree, and so the truth and error interpretation could be a valid choice. If the relationship is not strong, then the indistinguishable interpretation may be more valid. They see the latter view as more plausible, as the current lack of ability to further reduce model uncertainty for future climate using observations suggests that the models should currently be viewed as indistinguishable from each other for the future climate.

2.5 Summary

Limitations exist with models that can complicate the analysis of their results, which can also affect how the performance of models is determined. Different combinations of results from individual and multiple models have been explored in an attempt to better categorize some of the types of model uncertainties. The multi-model mean is one approach which has shown to be highly successful in reducing some of the types of model uncertainties. However, alternative weighting methods have been proposed to improve on it and explore other model and inter-

model relationships.

A concern seen from the application of some of these approaches, is that model biases do not strongly relate to the ability of a model to simulate the future climate change. So as Knutti et al. (2010) state, it is not known how much the models need to improve at simulating the present day climate in order to ensure that they will also better simulate the future climate change change. The field of combining model results is also complicated by the presence of common structural error, which limits the interpretations that can be drawn from an ensemble of models.

The next chapter looks at how modes of variability (e.g.: ENSO) can be used to assess model performance, which can be used to discount models which perform poorly. This could ultimately be used to help reduce model uncertainties in future work.

Chapter 3

Methods for Identifying Modes

3.1 Introduction

In the previous chapter, methods for addressing global model uncertainties were discussed. One of the approaches for reducing the effect of model uncertainties on the spread of results was by discounting models which performed poorly (section 2.3.3). This chapter looks at how modes of climate variability (e.g.: ENSO) can be identified in model results. The degree to which modes of variability are manifested in model results can then be used as a measure of model performance, and ultimately as a means of discounting models.

Specifically, a mode of climate variability (*mode*) is an underlying space-time structure in data with preferred spatial patterns and temporal variations that help account for gross features in variance and in associations between climate variables in widely separated but geographically-fixed spatial locations (IPCC-2013b: Annex III: Glossary, Flato et al., 2013a). An example of a mode is the El Niño Southern Oscillation (ENSO), the details of which can be seen in Burroughs (2003, p141-158).

This dissertation creates a measure of model performance according to how well models have simulated modes. The correct simulation of modes is important in two areas:.

1. One of the requirements that users have of model results, is that the results are credible: the results portrays a realistic climate (section 1.2). A performance metric which defines how well models have simulated modes may assist users in deciding whether model results are credible or not.

2. Modes can be responsible for generating regional biases (Giorgi and Coppola, 2010) along with structural error. So ensuring that models are simulating modes correctly may help in reducing model bias. This is similar to the work by Tebaldi and Knutti (2007) who state that models may be getting the correct result (e.g.: mean climate) due to the wrong reason (e.g.: tuning, section 2.4.2) rather than by correctly simulating the climate (e.g.: modes).

The task of identifying modes in climate data involves using a method to find patterns of potential modes in data and then successfully associating the patterns to known modes. *Patterns* are defined in this dissertation to be any structure within data, with often spatial or temporal aspects to them that may or may not be a mode. There are numerous methods for uncovering different patterns in data, the details of which can be found in Xu and Ii (2008). As this dissertation is concerned with patterns which may potentially represent modes, methods with prior application to climate data are reviewed in this chapter. The potential for one pattern to be associated to multiple modes (i.e.: represent a mode interrelationship) is discussed in section 3.7.2 but falls outside the scope of this work. The complexities of associating patterns to modes in general is discussed later in section 4.4, and so this dissertation adopts an alternative approach which is presented in chapter 5.

3.2 Correlation Maps

Wallace and Gutzler (1981) provide a pioneering approach using to explore potential relationship between geographically separate locations. These relationships are found using correlations maps and if successful, they are known as teleconnections. The maps are constructed by taking each grid point of the data, termed

the basis point, and correlating it with all the remaining grid points. This creates a series of correlation maps, one for each basis point of the data. To find the most representative patterns, the grid points with the highest negative correlation in a correlation map are viewed as the most distinct and therefore the most likely to represent modes. Figure 3.1 shows the strongest negative correlation for each correlation map plotted as a single image. Teleconnections describe the geographical distribution of spatial patterns found within gridded data.

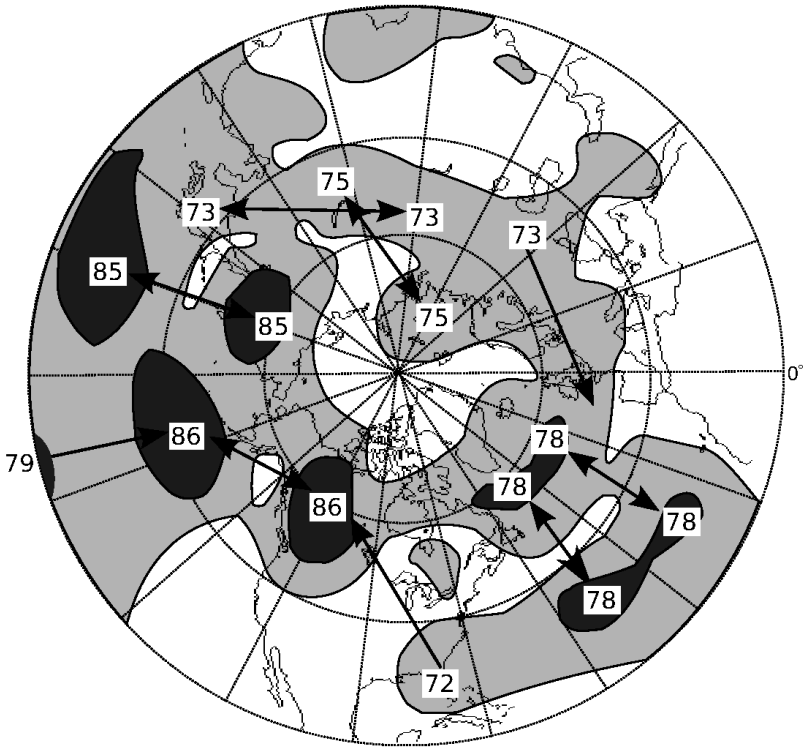

Figure 3.1: The strongest negative correlations from each of the correlation maps for the 500mb data, indicating teleconnectivity. The values are unsigned and multiplied by 100. Regions where the correlation coefficient (ρ_i) is less than 60 are unshaded, where $60 \leq \rho_i \leq 75$ are shaded lightly, and $75 \leq \rho_i$ are shaded heavily. Arrows link the center of a region of negative correlation, with the grid point that is most uncorrelated to it in its corresponding correlation map. Figure and caption recreated from Wallace and Gutzler (1981, figure 7b)

They find that the teleconnection patterns from 500mb geopotential height data are also present as some spatial distributions in both the North Atlantic Oscillation and the Pacific-North American Pattern. In doing so they identify these

modes within the same spatial region and time period as their data.

3.3 Cluster Analysis

A clustering technique groups data with complex interrelationships into a few simpler parts, which are known as clusters. Clusters can represent the data as a whole and can make it easier to interpret the data by uncovering previously unknown relationships within it (Xu and Ii, 2008). As clusters can represent relationships over time or geographical locations, they may be useful patterns for identifying modes in data.

Steinbach et al. (2003) are interested in using cluster analysis to find the time series of known climate indices, such as Niño 3.4. They wish to use the methodology they establish to find new climate indices. They employ the Shared Nearest Neighbor clustering (SNN) algorithm to cluster grid points from data (e.g.: SST). The SNN first calculates the similarity between a pair of grid points using a metric (e.g.: Euclidean distance (Ertöz et al., 2003)). If the pair are found to be similar, and share grid points that are also similar, then the pair are placed in the same cluster (Jarvis and Patrick, 1973). Within each cluster a representative time series, the centroid, is constructed using the mean of the grid points within the cluster. The centroids of the clusters represent possible climate indices. Existing climate indices are used to construct a threshold of minimum similarity.

They applied the algorithm primarily to SST data. After eliminating clusters that fell below the threshold, they found that the remaining centroids of the clusters were correlated with the known climate indices. When comparing one of the centroids with El Niño indices in terms of their correlation to land based vegetation growth, they found that the centroid highlighted different geographical regions that were related to the index. Their approach may therefore provide additional information on the spatial extent of phenomena such as vegetation growth.

Viewing similarity in terms of the absolute correlations with multiple variables, are Steinhaeuser et al. (2009) who implement the WalkTrap clustering algorithm.

Multidimensional nodes have time series from each of the variables associated with them. Following which, the distances between the nodes are adjusted according to the magnitude of the absolute correlation between their different time series. To reduce the total number of nodes, only the top few nodes with the greatest similarities are kept. The results produced show four globally distributed patterns, indicating processes generally related to Monsoons and Tropical Wet-Dry climate zones (figure 3.2). While the clusters are not explicitly linked to modes in their work, the communities may be seen as indicative of the behaviour of modes.

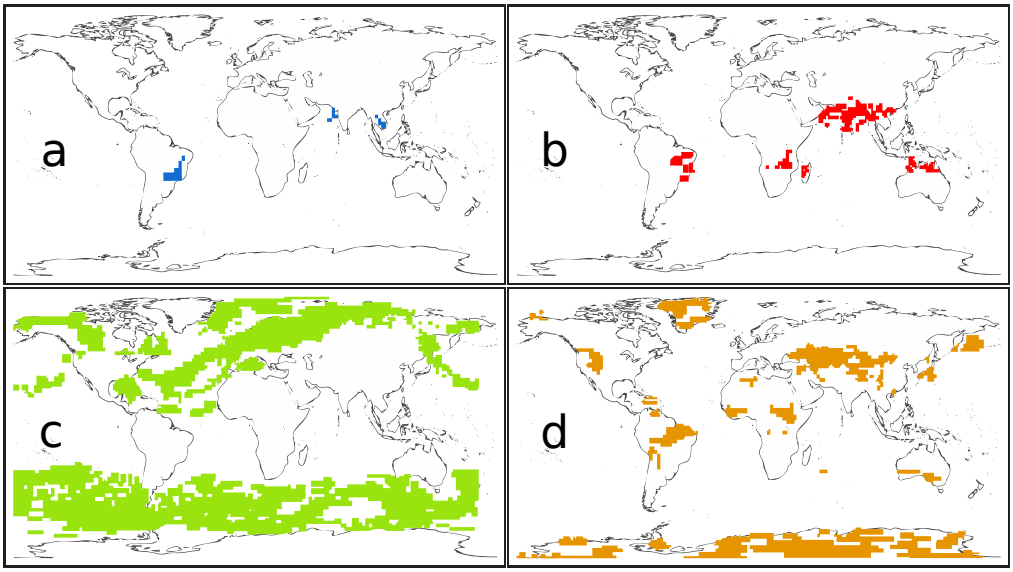

Figure 3.2: Select years from the four globally distributed communities of clusters, images and caption adapted from Steinhaeuser et al. (2009, figure 7). (a) Indicative of Tropical Wet-Dry (South America) and South East Asia Monsoon climate zones (1963). (b) Tropical Wet-Dry and Monsoon regions, especially in Northern India (1948). (c) Continental Sub-Arctic climate zone (1983). (d) Suggested to be relationships between precipitation and temperature which are effected by relief and deserts represented in some of the individual clusters (1963).

Christiansen (2007) highlights some problems seen with clustering algorithms in general. As the resulting clusters are dependent upon the number of clusters seen to exist within the data, deciding on the correct number is necessary to ensure that the clusters found are valid. The number however may be sensitive to the period of the data or threshold used. They also caution the use of clustering

algorithms which have poorly defined statistical properties as these methods may incorrectly determine the number of clusters present in the data.

3.4 Principal Component Analysis

Jolliffe (1986, p1) describes Principal Component Analysis (*PCA*) as a method which decorrelates a number of inter-related variables. This leads to a new set of uncorrelated and orthogonal variables. The new variables are constructed in such a manner as to maximize the amount of variance they represent, with each subsequent variable representing less of the total variance.

Compagnucci and Richman (2008) evaluate different applications of PCA to determine how well the results represent a set of artificial modes. Amongst other aspects, they examine S- and T- mode PCA. S-mode PCA groups spatial points (e.g.: grid cells) which have similar variability over time, while T-mode PCA groups the spacial elements at each point in time. They find that T-mode is best for discovering spatial clusters or teleconnections, which could be indicative of modes.

Richman (1986) opt to rotate the results produced from PCA to better capture individual modes. Compared to PCA, RPCA strives to associate a single rotated principal component to a subset of the input variables. In doing so relations may be found which may be more representative of individual modes. In addition to potential theoretical differences, they also discuss the limits of unrotated results to consistently identify the same modes when changing the size of the domain, and sampling errors that may be mitigated with RPCA when the components are similar and therefore difficult to computationally differentiate between.

Barnston and Livezey (1987) state that one of the practical advantages of using RPCA (and PCA) is that it creates a measure of importance for each of the representations of modes. With teleconnections this is much more difficult to do, as they have to be ordered more subjectively by the strength of the negative correlations associated with the basis points.

3.5 Self Organizing Maps

A Self Organizing Map (SOM) is a method of organizing a dataset into categories based upon a predefined number of categories. One way that SOM have been used is to find synoptic scale processes. An example of this can be seen in Hewitson and Crane (2002), where they categorize circulation patterns of sea level pressure reanalysis data.

More formally, they state that a SOM provides an unsupervised method for finding archetypal points that depict the multi-dimensional distribution function of the input data. To construct a SOM, nodes are placed randomly within the input data space. The distance between an individual input data element and the nodes is calculated, and the closest node is adjusted to decrease its distance to the element. This is repeated for all elements. Nodes surrounding the closest node also have their weightings modified. After a series of large modifications are made to the node weightings, a set of smaller weighting modifications are made to refine them. The refinement is conducted until no further changes to the node locations can be made. Following which, the frequency and variance of the identified categories can be calculated.

Having constructed the SOM, it can be viewed as a two dimensional map to show the spatially different categories (or patterns) over the region of interest. These patterns can then be associated to physical circulation patterns using an expert knowledge of the regional climate. Over Pennsylvania Hewitson and Crane (2002) find representations of strong central high pressure systems and transitional synoptic states.

Liu et al. (2006) contrast SOM patterns with PCA patterns using an artificial dataset that includes noise. Noise in data can complicate the process of finding patterns. By including noise in the artificial data, the performance results of the two methods may be useful for predicting their performance when actual data which may contain noise is used. They find that both methods produced similar results despite the addition of noise to the input data. However, their results differed when the input data contained asymmetrical components, as the SOM method was able to find them while PCA was unsuccessful.

Reusch et al. (2007) evaluate the differences between SOM patterns and PCs using reanalysis mean sea level pressure data over the North Atlantic. When they compare the results of PCA and SOM to the original data, they find that the techniques do uncover different spatial patterns, with the RMSE of the SOM patterns being generally lower than that of the PC.

3.6 Denoising Signal Separation

Denoising Signal Separation (DSS) (Särelä and Valpola, 2005) is a framework for uncovering components (e.g., time series) from data containing Gaussian noise. This is achieved through the use of a specified filter (linear or non-linear) which removes noise from the data. Prior knowledge about the components can also be incorporated into the filter to facilitate the extraction of the components.

The details of the linear DSS procedure are described in Ilin et al. (2006) using surface temperature, sea level pressure, and precipitation and are as follows:

1. The multivariate data is whitened: mean removed and the result decorrelated (e.g.: through time). This creates a set of noisy sources which are uncorrelated and have unit variance.

2. The Denoising / Filtering step: A linear filter is applied to the noisy components, changing their unit variances according to the prior knowledge. The variances of the components no longer serve to represent their variances, but are rather modified by the filter to represent the prior knowledge about the components. For example, as Ilin et al. (2005) want to capture inter-annual variability, they apply a filter which reduces the variance of components with frequencies of less than 12 months. This filter assumes that the remaining components will therefore have inter-annual variability.

3. New components are then found in directions within the filtered data which maximise the prior knowledge.

In the case of Ilin et al. (2005), they use a linear DSS procedure which identifies components with inter-annual variability. Using the filtered components from re-

analysis surface temperature, sea level pressure, precipitation, and a combination of them, they find representations of the El Niño Southern Oscillation.

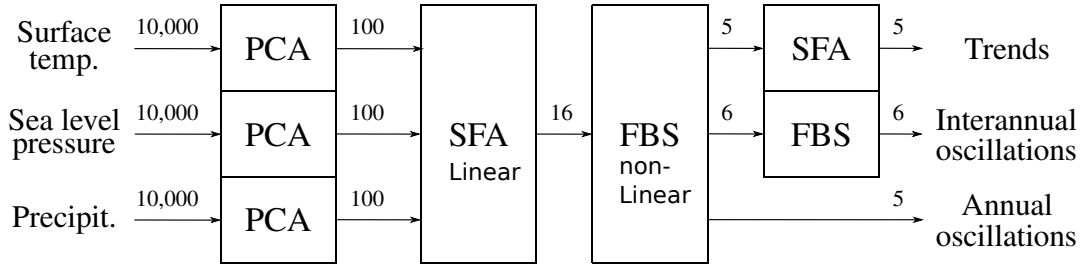

Figure 3.3: The methodology for finding the components using the denoising procedure, figure and caption adapted from Ilin et al. (2006, figure 7). The numbers above the arrows indicate the spatial dimensionality of the data (number of components). SFA is slow feature analysis and FBS corresponds to frequency-based separation.

Extending the procedure to a non-linear DSS approach are Ilin et al. (2006). The non-linear procedure performs the same first step of the linear procedure, but repeats (iterates) the last two steps until the separated components no longer change (figure 3.3). They first perform a linear DSS procedure on a combination of reanalysis surface temperature, sea level temperature, and precipitation datasets. Following which they then apply the non-linear approach to the results from the linear procedure. The filter they use is also based on the frequency of the components, but instead of assuming a fixed frequency, they rather use an adaptive one for finding components with variable frequencies.

In their results ENSO is represented by several components, which each differ in their spatial and frequency manifestations. One component is found to represent patterns of precipitation over the Chaco plain located in South America and the Sahel in Africa, the spatial manifestations of which can be seen in figure 3.4.

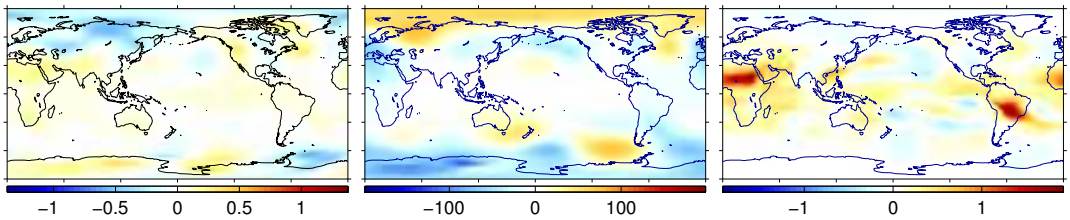

Figure 3.4: The spatial patterns of component 9 from Ilin et al. (2006, figure 10) for (*left*) Surface Temperature ($°C$), (*middle*) Sea Level Pressure (Pa), (*right*) Precipitation (kg/m^2).

3.7 Blind Source Separation

Cardoso (1998) states that Blind Source Separation techniques (BSS) uncover patterns from mixtures of them using a predefined model. The use of a predefined model enables the methods to directly incorporate assumptions about the data and patterns when no prior additional information (e.g.: frequencies) is available on the mixtures.

There are a variety of techniques available, and the interested reader is referred to Pedersen et al. (2007). This dissertation focuses on two such methods that have techniques that have been applied to the area of climatology, namely Projection Pursuit and Independent Component Analysis.

3.7.1 Projection Pursuit

Projection Pursuit (PP) seeks to reduce the dimensions of the data by maximizing a defined measure of *usefulness* (Friedman and Tukey, 1974). Xu and Ii (2008) state that the measure can be varied, and can be equivalent to other methods under certain conditions. For example, PP can be equivalent to PCA when the measure maximized is variance.

Christiansen (2009) explores the use of linear PP with different non-Gaussian measures, and apply them to stratospheric and tropospheric (20mb & 500mb) reanalysis geopotential height datasets. They explore using non-Gaussian measures as they argue that the results from some clustering algorithms (e.g.: k-means

clustering) may be due to skewness of data rather than being indicative of mode behaviour. In the troposphere using Kurtosis as a measure of non-Gaussianity, the projection is shown to represent a combination of patterns over Europe and the Pacific-North American pattern. When neg-Entropy (Hyvärinen, 1999a) is used as the measure of non-Gaussianity, the projection is seen to represent the Arctic Oscillation. When increasing the number of inputs for PP, the projection using Kurtosis only represents patterns over Europe. The projection using neg-Entropy is "identical" to the Pacific-North American pattern.

PP is generally found to be sensitive to the number of dimensions that are used in the input dataset, and the non-Gaussian measures that are used. They therefore recommend and apply a Monte Carlo statistical significance test to their projections to ensure their robustness. As for an overall best non-Gaussian measure, they rather discuss the results from the different measures, time periods, and geopotential heights.

3.7.2 Independent Component Analysis

Independent Component Analysis (*ICA*) is a method for identifying independent and non-Gaussian signals (e.g: time series) that are distinct from Gaussian mixtures of them. The problem that ICA aims to solve, is analogous to the "Cocktail Party Problem" given by Cherry (1953), where at a cocktail party, guests talking to each other are interested in hearing what each other have to say despite background noise, such as a piano playing.

A key motivation for applying ICA, is to provide a means of solving the *mixing problem* that can be seen with PCA. Aires et al. (2002) describe the problem where a single mode is represented by multiple principal components and propose ICA as a potential solution. The mixing problem therefore complicates the association of results to modes. They demonstrate that ICA can separate artificial signals that include noise, when PCA is unable to fully separate them.

The adopted approaches for finding signals are generally exploratory though, with works applying different ICA methodologies or using different data. For

example, different variables (surface temperature (Fodor and Kamath, 2003) and sea surface temperature (Westra et al., 2010)), data duration (10 years (Basak et al., 2004), 59 years (Hannachi et al., 2009)), and different ways of handling noise (noise free (Aires et al., 2000) and noise inclusive (Mori et al., 2006)). Therefore it may be difficult to compare the results when the application of ICA and data differ on more than one aspect.

One commonality that does exist between some of the works is both the geographical area of interest and the comparison of results from ICA and PCA. Using the same area allows for the different approaches to explore how ICA can be used to describe the modes of the region. As PCA is an already established technique in data analysis (section 3.4) it can often serve as a benchmark to compare the results of ICA against. The ICA applications according to geographical region are as follows:

Northern Hemisphere

The variability of the Northern Hemisphere climate is one such common region. The focus on this region is due to the interest in determining the true modes from the apparent ones. ICA is primarily introduced in the methods as an alternative to existing ones (such as PCA), where the aim is to differentiate between the real and apparent modes using the independence criteria of ICA. The complexities of associating signals to modes of climate variability are discussed in section 4.4.

Basak et al. (2004) use ICA to examine the variability of the region, with a focus on understanding the North Atlantic Oscillation in terms of spatio-temporal independent signals. They implement this approach by using ICA to separate temporally independent signals, where each mixture consists of a random spatial sampling of the input data at each time step. However it is not clear how the combination of temporal ICA and spatial sampling ensures both spatial and temporal independence between the signals. When using PCA to reduce the dimensions of the data, they find that the first signal matches one NAO dipole, while the second represents an average of the two dipoles. However, when they

used the original dataset without the PCA reduction step, they find that each of the two signals represents a different dipole related to NAO.

Mori et al. (2006) also examine the variability of the region, but look more to demonstrate ICA as a solution to the mixing problem using mainly reanalysis data. They find that the first PC best represents the Arctic Oscillation, while the first and second signal further divides the oscillation into representations of the Aleutian and Icelandic lows. So rather than finding one signal to represent one mode, their results show that one mode has been split into multiple representations in the signals. This is the opposite to the mixing problem, with multiple signals representing one mode.

Extending the analysis to uncover the relationships between additional modes are Itoh et al. (2007). They promote the use of data with long time periods in order to ensure the statistical significance of the results. To this end, they compare two 53 year reanalysis datasets of sea level pressure and geopotential height data (500mb), to the same variables from a set of longer climate model runs of 350 years. From both variables in the reanalysis dataset they conclude that the Arctic Oscillation is an apparent mode compared to the North Atlantic Oscillation and Pacific-North American Pattern. The mode is apparent because it is viewed as product of linear combinations of independent components, and therefore it does not represent an existing mode. Using the longer model data they are also able to show that the Arctic Oscillation is not an independent mode. As for the negative correlation mode between the Atlantic and Pacific, statistical significance could only be obtained using the model data. The results indicate that it too, is not independent of the other modes.

Hannachi et al. (2009) investigate the variability over the region using a novel ICA algorithm. The results from the application of the algorithm to reanalysis sea level pressure data indicate that the North Atlantic Oscillation, Arctic Oscillation, and a Scandinavian Pattern are independent of each other. The independence of the Arctic Oscillation is in direct contrast with the findings of Itoh et al. (2007), and further research would have to be conducted to determine if this only a product of the algorithm, or another aspect of their methodology (e.g.: preprocessing).

The Northern Hemisphere has been studied using several different ICA approaches which do not always agree on the independence of the identified modes.

Other areas of geographic application

Only a few other areas have been studied using ICA. Focusing on the tropics are Aires et al. (2000). Using their results they are able to interpret more signals as representing modes than they are able to with principal components, indicating the potential value of ICA in this area. Multiple signals are also shown to represent the El Niño Southern Oscillation, indicating the case were a mode is represented by multiple signals (see also Mori et al. (2006)).

Applying ICA on a global scale are Fodor and Kamath (2003). The basis of their research is to uncover signals which represent an the effect of either a volcano or the El Niño Southern Oscillation. As the signals are independent of each other, the effect could be removed without causing any additional changes in the rest of the data. They apply ICA to zonally averaged temperature data, and successfully associate one signal to the El Niño Southern Oscillation. They do not find any evidence of volcanic activity within their climate data.

Continuing with the application of ICA on a global scale are Westra et al. (2010), who apply it to observational sea surface temperature data. They find that the representation of the El Niño Southern Oscillation is split over several signals. As for the North Atlantic Oscillation, they do not find any signals which strongly represent it. Comparing the results of ICA to PCA and Varimax rotations, they conclude that ICA results are generally less easily associated to modes and may not add any new interpretations to them.

ICA Summary

The application of ICA to climate data has been focused on exploring the limitations and interpretations that can be drawn using different approaches. An interesting aspect of the research is, that on one hand ICA can provide a solution to the mixing problem, while on the other hand it can also create a reverse

scenario with one mode represented by multiple signals. This posses a challenge when trying to associate signals to modes. This challenge is further discussed in section 4.4.

3.8 Summary

There are a variety of methods with potential for revealing modes of climate variability (*modes*) within data. This dissertation has focused on some of the methods that already have applications in the field of climatology.

Selecting a method for identifying representations of modes in data is complex. Firstly, each method differs in its definition of a pattern and it is not always clear if a change in definition will result in a better representation of a mode. Secondly, all the methods discussed require some degree of expert knowledge to determined if a pattern does indeed represent a known mode. Concerns regarding reliance on expert analysis are further discussed in section 4.4. The result of these complexities is that the selection of method is ultimately subjective and will differ on the context of application.

Independent Component Analysis (*ICA*) is selected in this dissertation due to its enforcement of independence between the signals. Although strictly speaking, modes are not independent of each other in the truest sense, it may be assumed that they are independent in an attempt to find their unique behaviour. It is this property of ICA that may make it more able to avoid capturing mixtures of modes in a single pattern. However the solution is not straight forward, as ICA can split the representation of a modes over multiple signals. Therefore although the mixing problem may be avoided using ICA, the problem of finding individual modes using ICA is still complex.

Chapter 4 examines ICA, including the theory and details specific to its application. Subsequently, in chapter 5, a novel performance metric is defined which uses ICA but implements an alternative to the reliance on expert assessment.

Chapter 4

Independent Component Analysis and its Application

4.1 Introduction

In order to evaluate models according to how well they have represented modes of variability, modes have to be found within the data. The process of finding modes is a two part process. First patterns in the data have to be found using a technique, such as one of the techniques discussed in chapter 3. In the case of this work, Independent Component Analysis (*ICA*) is selected as it is maximises independence between the patterns that it finds (section 3.7.2). This property is believed to be useful in the second part of finding modes in data.

The second part of the process is associating the patterns found by a technique to known modes, and is performed using expert analysis. As ICA maximises independence between the patterns it finds, the patterns are less likely to represent multiple modes, which can complicate the association of patterns to modes.

This chapter outlines the background theory of ICA (section 4.2), how to determine the number of patterns within the data (section 4.3), and explores the challenges of manually associating patterns to modes (section 4.4). The technique is then grafted into the performance metric within chapter 5, to enable the

evaluation of models according to how well they have simulated modes.

4.2 Linear Noise-Free ICA Model

Hyvärinen and Oja (2000) describe ICA as a method for separating non-Gaussian signals from Gaussian mixtures of signals, where the signals are constructed to be independent of each other by maximising a non-Gaussian measure. The mixtures are assumed to obey the Central Limit Theorem, which states that the sum of non-Gaussian signals tends towards a Gaussian distribution. So the mixtures are assumed to be Gaussian distributions of independent signals which can be unmixed using a non-Gaussianity measure. The linear noise-free ICA model is selected due to its frequent use in the literature (section 5.4.1):

$$X_{m\times n} = A_{m\times m}S_{m\times n} \tag{4.1}$$

In the model, X is the input data matrix which contains the mixtures of the signals in its rows. It has m mixtures, and it is assumed that each mixture is observed over n time intervals. The rows of S are the original *signals* that were mixed within X, but can be recovered from it. The columns of A provide us with the degree to which the source signals were mixed within X, and therefore A is known as the *mixing matrix*. Unlike the recovered signals, the columns of A are not mutually independent of each other.

A closer look the at the ICA model shows us that it performs its separation from the signal mixtures in X. So the number of recovered signals (m) is equal to the number of rows in X. Due to ICA estimating both the A and S matrices, there are also a total of three ambiguities associated with the ICA model that Hyvärinen and Oja (2000) outline:

1. **Signs of S are unknown**
   A sign change in a signal has the equivalent effect of a sign change in the corresponding column of A. ($A(-S) \equiv (-A)S$). This can be remedied by manually setting their signs.

2. **Variances of $\boldsymbol{S}$ are unknown**
   A signal multiplied by a constant, has the equivalent effect if the corresponding column of A was divided by the constant. $(A(kS) \equiv (A/k)S)$

3. **Order of $\boldsymbol{S}$ are unknown**
   The results can appear in an arbitrary permutation, and therefore a custom ordering method can be imposed.

Hyvärinen and Oja (2000) also advise that preprocessing steps be conducted. They state that the mixtures should first be centred by having their means removed, and secondly the mixtures should be whitened. The whitening of the mixtures reduces the number of parameters to estimate in the ICA model, by decorrelating the mixtures and scaling them to have unit variances. One way that whitening can be achieved is through the use of Singular Value Decomposition (SVD). The SVD of X can be seen in equation 4.2

$$X_{m\times n} = U_{m\times n} D_{n\times n} V^T_{n\times n} \tag{4.2}$$

The columns of the U and V matrices are both orthogonal (uncorrelated) and have unit length. These columns are known as the left and right singular vectors of their respective matrices. The D matrix contains the singular values of the decomposition within its diagonal, with each singular value corresponding to a common singular vector in both of its neighbouring matrices. The singular values are the square root of the eigenvalues from the covariance matrix of X, and represent the standard deviation of the singular vectors. In addition to this, they are arranged in decreasing order of variance for convenience, along with their corresponding singular vectors. The remainder of the D matrix however, contains only zeros. The result is a decomposition of uncorrelated vectors, which are arranged in decreasing order of the variance they explain.

For performing ICA, the right singular vectors (rows of V^T) can be used as the whitened mixtures (Stone, 2004, p179-181). These are also known as the Principal Vectors (PV). The underlying assumption when using the PV as mixtures, is that the mixed signals are independent through time (as opposed to space). A

benefit of using SVD to whiten the mixtures, is that it can be used to reduce the number of dimensions of the data as well. Fodor and Kamath (2003) show that it is unlikely that every grid cell within the gridded data will represent a unique signal. To solve this problem, they recommend using PCA to reduce the dimensions of the data prior to using ICA.

SVD can be used to implement PCA (section 3.4), reducing the dimensions of the data for ICA by only retaining k PV. k is much smaller than $\min(m, n)$ but at least two, as this is the minimum number of mixtures that are needed for the ICA model. To apply this SVD approach to three dimensional data, the latitude and longitude spatial dimensions are reshaped into one large spatial dimension (M), with the *time* dimension forming the second dimension. The input data is now *space* $\times$ *time* ($M \times n$) which be seen in equation 4.3. By retaining only the first k PV associated with a large variance, small variance noise may be eliminated from the data.

$$X_{M\times n} = U_{M\times k} D_{k\times k} V^T_{k\times n} \tag{4.3}$$

Equation 4.4 shows the combination of applying PCA via SVD followed by ICA from equations 4.3 and 4.1 respectively. Note that the row means of the matrix ($J_{k\times n}$) that were removed during the centring preprocessing step, are added back to balance the equation. The utility of retaining variance, is further discussed within section 5.6.3.

$$V^T_{k\times n} = (A_{k\times k} S_{k\times n}) + J_{k\times n} \tag{4.4}$$

The ICA model (equation 4.1) assumes that there is no noise within the mixtures. So without removing low variance noise prior to performing ICA, highly non-Gaussian signals may be recovered, when they are actually low variance noise. SVD may be used to solve this problem by removing potentially noisy components. Due to the arbitrary ordering of the signals, the PV and Signals are ordered from least like noise to most like noise using their absolute Kurtosis. The appropriate modifications of the remaining matrices in equations 4.3 and

4.4 are performed to ensure that only the order of the results change (see section 5.4.2). Other possible orderings are by uncertainty (Westad and Kermit, 2003) or by a non-Gaussian measure (Hyvärinen, 1999b). Determining how many components to retain is a context specific question, that is discussed further in section 4.3.

4.3 Number of Signals to Retain

In the area of climatology there are many different ways of determining the number of signals to retain from the data (denoted by k in e.g.: equation 4.3). A common method, is to separate the same number of independent components as principal vectors. One method for determining the number of PV to retain for ICA is based on the proportion of variance that the PC explained of the data (Fodor and Kamath (2003), Lotsch et al. (2003)), while another was to chose the number based on computational performance (Basak et al. (2004)). Aires et al. (2000) choose to adopt the methodology offered by Nadal et al. (2000)), which suggested retaining only a few strong PC which would allow an adequate number of signals to be separated.

In Scholz et al. (2004), they were interested in extracting leptokurtic signals for use in metabolic fingerprinting. Their work considered selecting the best number of signals to use, by calculating the kurtosis of the signals per each set of PC. They then plotted the number of signals with leptokurtosis against the current number of PC being used. Each time they extracted the same number of independent components as they had PC.

When viewing the plot, the number of leptokurtic signals was seen to be at a maximum when 6 PC were used, while after 8 PC the minimum number of leptokurtic signals was seen to persist. Therefore by calculating the kurtosis for each signal per set of PC, they were able to decide on the best number of PC and subsequently, the best number of signals to extract from their data.

Koch and Naito (2007) propose a combined method grafting PCA and ICA together. The method offers a trade off between lower dimensional data and the

information that it contains. They propose the use of a non-parametric method which uses kurtosis and skewness to determine the number of signals to retain. The advantage of their method is that it requires no prior information on the data to determine the number of components to retain.

The Rule-N method by Overland and Preisendorfer (1982), ensures that the retained PV, and therefore signals, are above the level of noise. This is carried out by performing the same decomposition analysis on a dataset containing only Gaussian noise. Only PV that have the same or more variance than their equivalently ranked components from the noise dataset are retained for further analysis. This method provides an objective approach to determining the number of components to retain and so it is chosen as the method to use in this dissertation.

4.4 Associating Signals to Modes

To show that data contains representations of modes, potential representations are found in the data and associated to modes. With ICA, the signals become the potential representations of modes which have to be associated to modes. If the signals are successfully associated to modes, then the data can be said to contain representations of the modes. A *signal* refers to a time series that has been found using ICA. Signals are separated from data as a set, and the statistical independence between the signals is maximised during their separation from the data. To assist in the association of a signal to a mode, a *climate index* or *index* is often used. A climate index is a time series that can be derived from observation data and represents the behaviour of a mode. It is useful in associating signals to modes as it often used as the definition for the temporal behaviour of a mode. An example is the Niño 3.4 index (Trenberth and Stepaniak, 2001), which is used to describe some of the behaviour of the El Niño Southern Oscillation.

The identification of representations of modes in data requires both the separation of signals and the association of the signals to modes. While the former may be more mathematically defined, the latter often takes on a more subjective approach. In fact, the success of a pattern recognition technique lies in the

ability of the researcher to associate the signals to modes or other meaningful instances (e.g.: trends, inhomogeneities, etc.). While there are many methods for assisting the researcher in the this context (e.g: spectral analysis, spatial distributions), there is still no one equivalent method for associating the signals to modes. Westra et al. (2010) echo this difficulty with associating signals to modes.

Westra et al. (2010) further discuss another association problem, namely the difficulty of associating non-Gaussian signals to modes, when the modes are themselves defined by other pattern recognition techniques. They depict the case where a signal (as a time series) is associated to the Pacific Decadal Oscillation (PDO) by correlating the signal to the PDO Index. However, the PDO Index is defined as the leading principal component. As the signal and index are derived using two different methods (ICA and PCA respectively), they may have different distributions which may make it difficult to associate the signal with the PDO. Similar cases involving indices generated from other techniques can be seen in (Jones et al., 2007, p.287), where the Northern Annular Mode Index is derived using PCA and the Southern Oscillation Index is created using station differences.

Additionally, there exists the potential for the illusion of non-Gaussian modes to occur if the data is not sampled correctly, which is discussed by Itoh et al. (2007). They state that an ICA mixture which contains samples (e.g.: months) that have Gaussian distributions, but differ in mean and variance, may create the illusion of a non-Gaussian mode being present in the mixture.

Contrary to the interpretation of non-Gaussian signals as modes, is the work by Sura et al. (2005), who suggest that non-Gaussian distributions could be caused by multiplicative noise. They show that some linear systems, when combined with multiplicative noise, can produce non-Gaussian distributions. This type of analysis highlights a difficulty in correctly associating the signals with modes or other instances.

Identifying modes using non-Gaussian signals is a complex task, given the above mentioned limitations. The task at hand is to therefore use the available methods

and to also validate the results as far as possible. However it remains a subjective approach, and so its success ultimately depends upon the skill of the researcher to validate any associations between signals and modes.

Due to the challenges of the association procedure, this dissertation rather develops an alternative approach that allows the association procedure to be side stepped. This approach is described in chapter 5.

4.5 Assumptions and Limitations of ICA

There are a number of assumptions and limitations with the application of ICA that may be present depending on the data and method of performing ICA. The assumptions and limitations are that are relevant to this dissertation are outlined in this section:

- Determining the number of signals present that may be present in the data has to be known prior to their separation from data. Examples of methods for determining the number signals can be seen in section 4.3.
- The presence of noise in data can complicate the separation of signals. This may effect what approach is taken to handle noise, such as removing noise using PCA as preprocessing step or having an ICA algorithm isolate the Gaussian noise. Handling noise is further discussed in section 4.4. This work assumes that any noise present in the data is noise with low variance that is removed during the preprocessing of the data (see section 5.3).
- Itoh et al. (2007) discuss an artificial case where one signal is incorrectly found to be more independent of the data than another signal. In the case of the first signal, it is really the sum of two Gaussian distributions with different means and variances which when combined give the false impression of a signal. To address this they recommend caution in selecting the data period to which ICA is applied to ensure that the variance between samples does not differ greatly.

- Due to current applications of ICA generally being of space by time (e.g.: Westra et al. (2010)), the spatial manifestation of a signal is implicitly assumed to be constant over the period of study. While this may not be an issue when examining standing modes like the North Atlantic Oscillation that generally remain spatially static in their manifestation, another mode such as the Madden-Julian oscillation (MJO) (Madden and Julian, 1971) may be more difficult to classify when examining their spatial patterns because it does not remain spatially static. The consequence of not remaining static over a period, is that while a signal may be associated to the mode, its corresponding spatial pattern may show the mode spread out over the region where the mode has started and ended its movement.

- Hyvärinen et al. (1999) point out a potential problem that may occur when the period of data used is too short. In this case ICA produces signals which are almost entirely zero except for a single large spike. This problem is known as overfitting of the data. They recommend using PCA to remove noise or increase the time duration of the data used. Using PCA is shown by Fodor and Kamath (2003) to successfully solve the overfitting problem in their work. Increasing the duration of the data used is shown by Itoh et al. (2007) to assist with obtaining statistically significant correlations of the signals with climate indices, thereby providing an additional benefit to solving the overfitting problem. Nevertheless, using a longer period of data may not be a panacea. Hannachi et al. (2009) state that for longer periods of time there is potential for non-Gaussian behaviour that exists on short time scales which may be reduced if the data is averaged over time. This may make it more difficult to find non-Gaussian signals in the data.

- Richman (1986) discuss the dependency of the results from PCA to the region over which it was applied, namely that if the region changes slightly then the identified modes can also change. This is not ideal, as the identified components should remain the same over similar region. The degree of this effect in ICA has not explored yet, but if PCA is used as a preprocessing step, then it follows that the application of ICA would also inherit this sensitivity to the geographical region.

Chapter 5

Performance Metric Design

5.1 Introduction

This chapter presents the design of the performance metric. Specifically, the performance metric uses Independent Component Analysis (*ICA*, chapter 4) to find representations of modes of variability (*modes*, chapter 3) from within reanalysis data. Here the reanalysis dataset serves as the reference or standard to compare model results against (section 1.1). The degree to which initial condition ensemble members (*members*, section 2.3) contain the representations is used as the measure of how well the models have simulated the modes found in reanalysis data. The design also presents a solution to the association problem seen in section 4.4.

The datasets (section 5.2) preprocessing steps (section 5.3) are presented in this chapter. The method for finding the PV and signals in reanalysis data is shown in section 5.4, while the method for finding the reanalysis PV and signals in other datasets is shown in section 5.5. The measure of reanalysis PV and signals in non-reference datasets is presented in section 5.6 along with the ICA based performance metric in section 5.6.5. Lastly, the limitations of the performance metrics and data are reviewed in section 5.8

5.2 Datasets

Ten datasets are used: one reference (reanalysis), one alternative reanalysis, six climate model ensemble members (members), one Gaussian dataset to represent noise, and one multi-model mean dataset constructed from the ensemble members. For each dataset, global gridded monthly geopotential height data at 700 hPa is used as it is assumed to contain mixtures of modes.

Using geopotential height data has proven useful for finding modes in similar works (e.g.: Wallace and Gutzler (1981); Itoh et al. (2007)). This may be because it can be used to measure weather systems in the lower troposphere within the extratropics, but to a lesser degree the weather systems in the tropics where the geostrophic approximation does not hold. The period of January 1961 to December 1990 (30 years, 360 months) was selected due to data availability.

NCEP *reanalysis* data (Kalnay et al., 1996)[1] is used as the reference dataset, as it is a data product that similar works have used (e.g: Basak et al. (2004); Mori et al. (2006))(see also section 1.1). ERA-40 reanalysis (Uppala et al., 2005) is used as the alternative reanalysis dataset due to data availability. Initial condition ensemble members are taken from the hindcast simulations of the core CMIP5 near-term experiment number 3.2 (Taylor et al., 2012). Due to data availability, members from models BCC-CSM1.1 (Xiaoge et al., 2012), CNRM-CM5 (Voldoire et al., 2013)[2], and MPI-ESM-LR (Giorgetta et al., 2012) are used. Each model contributes two different realisations: r1i1p1 and r2i1p1 which are initialised at the end of 1959 or during 1960.

As the spatial resolution of the reference and ERA-40 datasets are 2.5 degrees (144 longitude and 73 latitude), the ensemble members are bilinearly interpolated from 1.4 degrees (256 longitudes and 128 latitudes) to the spatial resolution of the reference dataset. A dataset containing only noise is also constructed from a Gaussian distribution with the same variance and mean as the reference dataset. This is to ensure that the only difference between it and the reference dataset is that it contains noise. This will ensure that the mean and variance of the Gaussian

[1]The data include a shift to using full satellite data in 1979.
[2]See also: http://www.cnrm.meteo.fr/cmip5/

dataset is not the cause behind the performance of the Gaussian dataset. The multi-model mean is constructed from the mean of all the members.

Table 5.1 shows the shortened nomenclature for the datasets used in this work.

Simplified Name	**Full Dataset Name**
R	Reference (NCEP)
E4	ERA-40
B1	BCC-CSM1.1 r1i1p1
B2	BCC-CSM1.1 r2i1p1
C1	CNRM-CM5 r1i1p1
C2	CNRM-CM5 r2i1p1
M1	MPI-ESM-LR r1i1p1
M2	MPI-ESM-LR r2i1p1
MM	Multi-model Mean
G	Gaussian Noise

Table 5.1: The nomenclature for all the datasets.

5.3 Preprocessing Steps

The performance metric design requires that the datasets be preprocessed before ICA can be applied to it. The steps are as follows, with a summary presented in section 5.7

1. **Bilinear Interpolation**
   Bilinearly interpolate each member to the same spatial resolution as the reference dataset. This method was selected as it did not introduce artefacts into the data.

2. **Subset the Data Temporally**
   The data from 1961 to 1990 is retained due to the availability of the ensemble member data (section 5.2).

3. **Remove Linear Trend**
   The linear trend is removed to prevent its detection by ICA. It also removes

the mean for the data over the period. Trend removal is not a requirement of ICA, but is rather done to better enable the detection of signals that are not dominated by a trend. This ensures that the PV and signals are more likely to represent modes rather than a trend, the complexities of which are discussed in section 4.4.

4. **Remove Seasonal Cycle**
   This is also not a requirement of ICA, as can be seen in the work of Kent (2011) where the seasonal cycle was not removed and was still detected in the signals. Its removal however, does not hinder the detection of non-obvious representations, as subsequent PV are no longer constrained to be orthogonal to the seasonal cycle. To remove the seasonal cycle, the seasonal mean is removed from the time series of each grid cell. The seasonal mean is constructed from only the corresponding months in the period. For example, the mean for March is calculated based on the mean of all the month of March over the period. The time series are then normalised over the corresponding months to prevent high latitude regions, which have large variability, from dominating the variability of the data. This method is known as the Monthly Z score method (Tan et al., 2001). Tan et al. (2001) also present some alternative methods for determining the magnitude of the seasonal cycle. In their work the Discrete Fourier Transform shows similar results to the Monthly Z score method and to a lesser extent also compares to the Singular Value Decomposition method. They also found that the number of singular vectors to retain has to be subjectively determined.

5. **Weight Cells By Latitude**
   Adjust the values of cells with relatively small latitudes (Baldwin et al., 2009). Failure to account for this problem will result in geographical regions at higher latitudes dominating the variance of the results due to the spatial size of the grid cells. This may hinder patterns from being properly detected in lower latitude regions.

5.4 Finding PV and Signals in Reanalysis Data

In section 4.4, the complexities of associating signals to modes of climate variability using expert analysis are discussed. The challenges makes it difficult to apply ICA to multiple datasets as a performance metric, and so this section outlines a solution to the problem. To address the problem, the PV and signals are found in reanalysis data. The degree to which they are found in other datasets is then used to assess the performance of the datasets. By taking the PV and signals from reanalysis data, they are automatically deemed to represent modes (see figure 5.1). This is based on the presumption that the reanalysis data represents the climate and therefore any patterns from it will also represent valid modes. The removal of the expert assessment from the application of ICA to data, has the benefit of making its application to multiple datasets more computationally feasible.

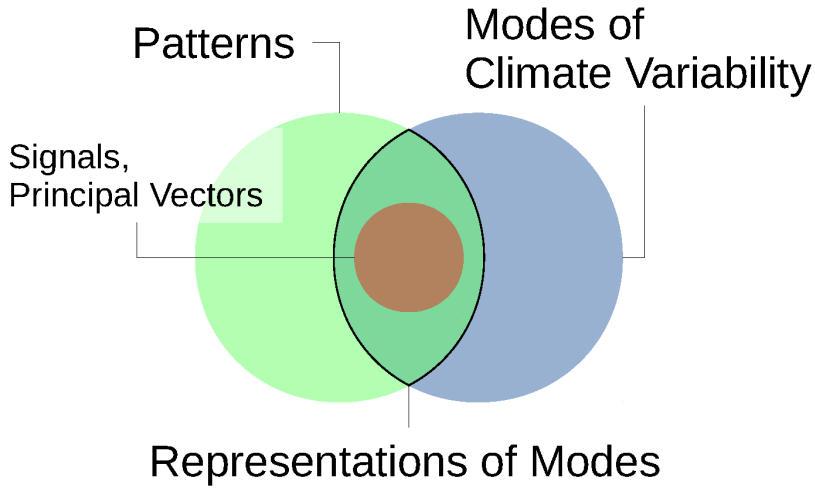

Figure 5.1: Representations of modes are found when patterns are successfully associated with modes of climate variability. In this dissertation, the PV and signals are the patterns which are found within reanalysis data, and therefore they are automatically presumed to be representations of modes.

Following the preprocessing of the datasets (section 5.3), the PV and signals are found in the reference dataset using equations 4.3 and 4.4. The algorithm used to separate the signals from the reanalysis data is presented in section 5.4.1 while a method to ensure that the results are consistent, is outlined in section 5.4.2.

5.4.1 FastICA Algorithm

In this work, signals are separated from the data using the FastICA algorithm by Hyvärinen (1999a). This algorithm was selected due to its common use which will allow comparisons to similar works (e.g.: Basak et al. (2004); Mori et al. (2006); Fodor and Kamath (2003); Westra et al. (2010)). The algorithm has an *algorithm type* parameter that dictates how it functions, with two possible arguments. The first argument is *deflation*, and instructs the algorithm to sequentially separate signals. The second argument is *parallel*. This argument causes the algorithm to separate signals in groups. The latter argument is preferred in this work, as it may avoid the accumulation of errors that can occur when the deflation argument is used (Ollila, 2010).

To implement the algorithm, the R programming language is used (Team, 2015) along with version 1.1-11 of the FastICA algorithm developed by J. Marchini and C. Heaton. Table 5.2 indicates the FastICA parameters and the corresponding arguments used in this work.

Description	Parameter (Variable)	Argument (Value)
Signal Mixture	X	$V_{n\times k}$
Number of signals to separate	n.comp	k
Algorithm type	alg.type	parallel
Contrast function	fun	logcosh
Method	method	C
Level of output information	verbose	True
Distribution Type	alpha	1*
Normalise the rows	row.norm	False
Number of iterations	maxit	200*
Convergence threshold	tol	$1e_{-4}$*
Initialise unmixing matrix	w.init	null*

Table 5.2: The parameters of the FastICA algorithm with their corresponding arguments. Default values are indicated by an asterisk (*).

Although the FastICA has its own function to perform dimension reduction, its construction of the covariance matrix with dimensions of $M \times M$, was found to be too memory intensive for this work. The alternative SVD implementation

(section 4.2) is rather adopted for the requirements of this work. As the same number of signals as mixtures is always used in this work, no additional dimension reduction is performed by the FastICA algorithm. Also, any rotation of the data is linear and therefore would create an equivalent set of mixtures.

To standardise the results an additional custom wrapper function is used prior to and after the FastICA algorithm. The wrapper function ensures that the row-means ($J_{k\times n}$) of the input matrix are saved, as they are not saved in the original FastICA implementation. The row-means are important to save as they are required in further calculations (see equation 5.6). The wrapper function also transposes the input matrix, and subsequent output matrices as the FastICA algorithm uses column-wise variables, while this work assumes row-wise variables. Lastly, the wrapper function scales the signals to have unit variance, and also scales the mixing matrix accordingly to preserve the total variance, which would otherwise be lost.

Although no noise is explicitly handled by the noise-free linear ICA model (section 4.1), the FastICA algorithm allows at most one signal to have a Gaussian distribution. Moreover, no noise is assumed to be present in the data as only the PV that are above the level of noise are retained for further analysis.

5.4.2 Ensuring Stability of FastICA Results

The FastICA algorithm (section 5.4.1) separates signals from data by iterating until a threshold of independence is reached. In the beginning the unmixing matrix (A^{-1}) is populated with random values. The use of random values as an initial estimate of the matrix can effect the final set of signals (e.g.: Kent (2011)). The outcome of this is that the signals may differ between identical runs of the FastICA algorithm when all other aspects are constant. Consequently, a solution is introduced to minimise this effect on the signals.

Once the number of signals to separate from the data has been determined (Rule-N method, section 4.3), an averaging method is employed to address dependency of the matrix on the initial values. The FastICA algorithm separates out signals

according to equation 4.4, while allowing only the initial estimate of the unmixing matrix to differ. After each run the unmixing matrix is retained along with the threshold value obtained for that run. The unmixing matrix is then averaged over a number of runs, but only using those matrices that converged. This is repeated for 10, 100, and 500 runs of the algorithm, and each set of runs is conducted 5 times. This enables the standard deviation of each set of runs at a given size to be determined. The averaged unmixing matrix corresponding to the set with the lowest standard deviation is used to separate the signals from the reference dataset.

As averaging the unmixing matrices may result in a loss of variance, the final matrix is scaled to the equal the total input variance and thereby preserve the total variance as shown in the ICA model (equation 4.1). The columns of the unmixing matrix are ordered by the absolute kurtosis of the signals to overcome the ordering ambiguity in the ICA model.

5.5 Finding Reanalysis PV and Signals in non-Reference Datasets

To find the reanalysis PV and signals in other datasets, two filters are created: one for finding the PV and another for finding the signals. The filtered results may be similar to those from the reference dataset (section 5.4) in sample number and period, but may differ in properties such as their distributions. For simplicity, the filtered results will still be referred to as PV and signals.

The PV filter is as follows:

$$D^{-1}_{k\times k} U^{+}_{k\times M} \boldsymbol{X}_{\boldsymbol{M\times n}} = \boldsymbol{V}^{\boldsymbol{T}}_{\boldsymbol{k\times n}} \tag{5.1}$$

Emboldened matrices are from non-reference datasets (e.g.: $\boldsymbol{V^T}$). $\boldsymbol{X}$ is a pre-processed non-reference dataset, while the unemboldened D and U matrices are taken from the decomposition of the reference dataset (section 4.3). As U is a non-

square matrix, its inverse is approximated using its pseudo-inverse (U^+).

To find the signals, the PV filter (equation 5.1) is extended to include the mixing matrix A and row mean matrix J taken from the reference dataset signal separation step in equation 4.4. The application of the signal filter is shown in equation 5.2 below. No reordering of the signals is required as the mixing matrix has already been ordered (section 5.4.2).

The signal filter is as follows:

$$A^{-1}_{k\times k}(\boldsymbol{V^T_{k\times n}} - J_{k\times n}) = \boldsymbol{S_{k\times n}} \tag{5.2}$$

5.6 Measure of Pattern Strength using Relative Variance

Section 5.5 discusses how the PV and signals are found in the non-reference datasets. This section examines how the degree to which the reanalysis PV and signals are manifested in the other datasets, can be measured. In this work, relative variance is selected to compare the filtered PV and signals from the non-reference datasets to the PV and signals from the reanalysis dataset. Measuring variance is seen to be important to ensure that datasets capture the correct strength of the modes. Section 5.6.1 demonstrates the use of variance and relative variance as possible measures using an artificial example, while the specific calculations are discussed in subsequent sections.

5.6.1 Demonstration Using Artificial Example

In this section, an artificial example demonstrates how the presence of reanalysis PV and signals can be measured in non-reference datasets: first using variance then using relative variance. For simplicity, it is assumed that the variance of signals can be calculated despite the ambiguity of the ICA model (section 4.2). The example is presented in figures 5.2 and 5.3. In figure 5.2, the reference dataset is presented with the variance and the percentage of variance explained (PVE) of its individual components compared to the total variance. It has a total variance

of 10 distributed between its two PV as 8 and 2, and between its signals as 4 and 6. The PVE of the PV are 80% (8/10) and 20% (2/10), and for the signals they are 40% (4/10) and 60% (6/10).

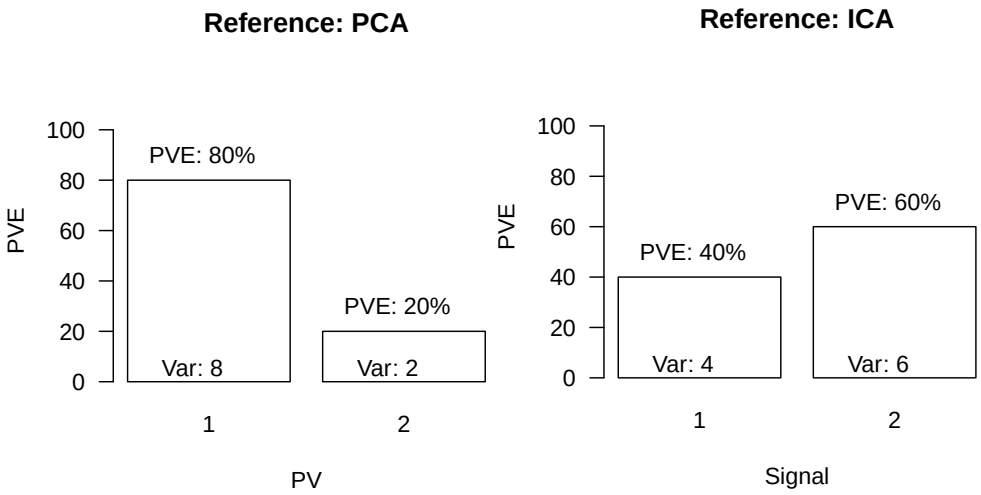

Figure 5.2: The reference dataset with the PVE of PV (*left*) and signals (*right*).

Figure 5.3 shows a dataset which has been filtered by the reference dataset. In the ideal scenario, the dataset would perfectly represent the reference dataset. So in calculating the PVE of component from a dataset, the total variance from the reference dataset (10) and not the individual dataset under consideration is used.

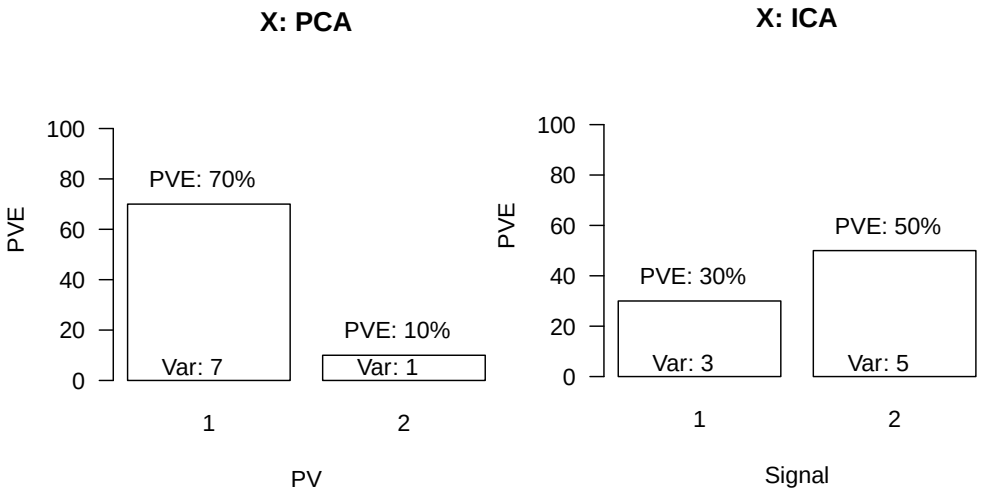

Figure 5.3: Filtered dataset X, with the PVE of PV (*left*) and signals (*right*).

In this example, the dataset is chosen to have a total variance of 8, and so only captures 80% (8/10) of the total variance of the reference dataset. The PVE of the PV are 70% (7/10) and 10% (1/10). The signals have PVE of 30% (3/10) and 50% (5/10). If total PVE were to be used as the overall measure of dataset performance, both the results from PCA and ICA would be identical, both with a total PVE of 8. They are equal in total PVE because the V^T from the SVD of the data in equation 4.3 is also used as the mixtures for ICA in equation 4.4.

The relative PVE (*RPVE*) of the PV and signals is considered as an alternative to using the PVE. The RPVE of a component is the ratio of variance between that component and its equally ranked component from the reference dataset. An artificial example of this is presented in figure 5.4. It has the same distribution of variance as the previous example in figure 5.3, but shows the RPVE of the PV and signals. For the PV, the RPVE are 88% (7/8) and 50% (1/2). The RPVE for the signals is 75% (3/4) and 83% (5/6).

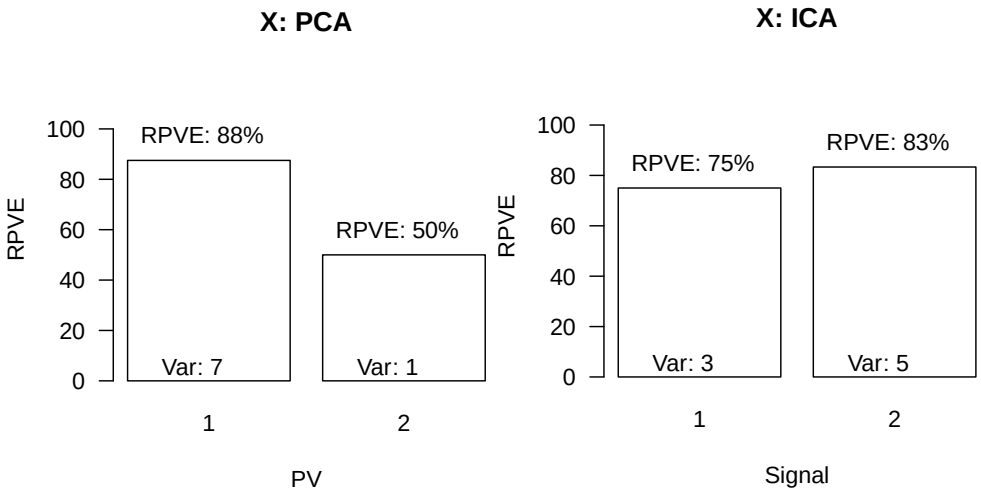

Figure 5.4: Filtered dataset X, with the RPVE of the PV (*left*) and signals (*right*).

Most importantly, the sum of the RPVE for the PV (138%) is not the same as for the signals (158%). This means that for total dataset performance, the total RPVE can differentiate between PCA and ICA results when the total PVE is unable to make the same distinction.

5.6.2 Limitations Using Relative Variance

A limitation with using the relative variance measures ($RPVE$), is that a small change in the distribution of variance amongst components may result in a marked change in the RPVE for other components. An artificial example of this can be seen in figure 5.5. This is the same dataset as the original (figure 5.3) with a total PVE of 80%. To demonstrate the effect of a change in RPVE, the variance of the PV in this example are changed to 6 and 2, producing RPVE of 75% and 100% respectively. The total RPVE is now 175%, which is an increase of 37% compared to the total RPVE of the PV in the previous example at 138%. So with just a small change in variance amongst components (PV or signals), the RPVE may differ markedly. In this work, datasets are assumed to be reasonably close to the reference dataset and are therefore unlikely to incorrectly simulate the strength of reference dataset variance to a large degree.

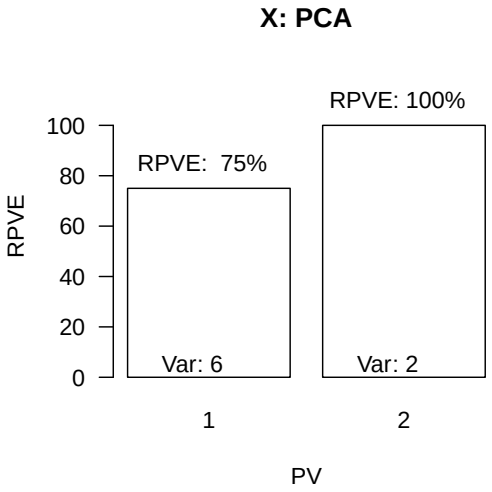

Figure 5.5: Filtered dataset X, with the same total variance as the PV in figure 5.3. The second PV now has now double the variance and RPVE, demonstrating the potential sensitivity of the RPVE calculation to the distribution of variance.

A second limitation is that the RPVE assumes that the components have the same order. If they are not ordered, potentially different patterns will be compared in terms of their variance and the wrong RPVE will be calculated for them. In the case of this work, an ordering method is imposed for the PV and signals which overcomes this problem (see section 5.4.2).

5.6.3 Percentage of Variance Calculations

While the concept of percentage of variance explained (PVE) was first discussed within the artificial example (section 5.6.1), this section shows how it is calculated. Note that the PVE is the variance of an individual PV or signal represented as a percentage of the total variance of all PV or signals and the PVE calculations are also only calculated for the PV and signals from the reference dataset. For the other datasets, relative variance ($RPVE$) is used (section 5.6.4). The variance of signals cannot be calculated directly due to the ambiguity in variance of the signals in the ICA model (section 4.2). So a proxy for them is created as PVE.

PVE is generally an approximate measure of the variance for a PV or a signal. For the PV from the reference dataset, it is an exact measure, and is given in equation 5.3. The variance of the q^{th} PV can be calculated using the square of the q^{th} non-zero singular value, divided by the sum of the square of all the non-zero singular values. The PVE is then the variance of each individual PV represented as a percentage of the total variance.

$$PVE\ of\ q^{th}\ PV = \frac{(D^q)^2}{\sum_{n} D^2} * 100\ \ : (q \in k) \tag{5.3}$$

Although signals from the reference dataset are constrained to have unit variance by design (section 4.2), their contribution in variance to the total variance can be approximated. The approximation is based on the assumption that equation 5.3 for finding the PVE for the q^{th} PV may also be expressed by both equations 5.4 and 5.5.

$$U_{M\times k} D_{k\times 1} V^{qT}_{1\times n} = Y^{q}_{(M\times n)\times 1} \tag{5.4}$$

$$\frac{\text{variance}(Y^q)}{\text{variance}(X_{(M\times n)\times 1})} * 100 \approx \frac{(D^q)^2}{\sum_{n} D^2} * 100 \tag{5.5}$$

In equation 5.4, the matrices produced from the SVD of $X_{M\times n}$ (equation 4.3)

are recombined for the q^{th} row of V^T and the q^{th} column of D into a single vector containing $M \times n$ elements denoted as Y^q. The PVE of this matrix is then defined as its variance relative to the total variance of the reference dataset (X) in equation 5.5. This new ratio is found to be very similar to equation 5.3, though there is no formal proof of this relationship.

The new ratio allows for the substitution of the of PV (V^T) with the mixing matrix (A) and row means (J) to determine the PVE of the signals from the reference dataset in equation 5.6. In the equation the q^{th} row of S and the q^{th} column of A are used. The division of the row-means matrix by the number of signals retained is to ensure that the matrix is added a total of once to the PV for all the signals. The PVE of the signals can therefore be calculated as shown in equation 5.7.

$$U_{M\times k}D_{k\times k}((A^q_{k\times 1}S^q_{1\times n}) + J_{k\times n}k^{-1}) = Z^q_{(M\times n)\times 1} \tag{5.6}$$

$$\text{PVE of Signal q} = \frac{\text{variance}(Z^q)}{\text{variance}(X_{(M\times n)\times 1})} * 100 \tag{5.7}$$

5.6.4 Relative Percentage of Variance Calculations

The artificial example (section 5.6.1) demonstrated the problem with using the total variance as the measure of overall performance for a dataset. Formally, due to equation 4.4, the total PVE of the PV is equal to the total PVE of the signals, so total PVE cannot differentiate between PCA and ICA results. To work around this problem, the relative percentage of variance explained ($RPVE$) is introduced and as it not constrained in the same way as PVE. The RPVE is the ratio between the variance of a filtered PV or signal and the variance of its equally ranked counterpart from the reference dataset.

The ratio presented in equation 5.8 extends equation 5.3 to handle PV from a non-reference dataset (emboldened) and for the PV from the referenced dataset.

This equation can be simplified to equation 5.9.

$$\text{RPVE of } q^{th} \text{ PV} = \frac{\text{variance}(U_{M\times k} D_{k\times 1} \boldsymbol{V^{qT}_{1\times n}})n}{\text{variance}(U_{M\times k} D_{k\times 1} V^{qT}_{1\times n})} * 100 \tag{5.8}$$

$$\text{RPVE of } q^{th} \text{ PV} = \frac{\text{variance}(\boldsymbol{V^{qT}_{1\times n}})}{\text{variance}(V^{qT}_{1\times n})} * 100 \tag{5.9}$$

Similarly for the signals, equation 5.7 can be extended to handle the signals from a non-reference dataset and the signals from the reference dataset. This is presented in equation 5.10 and can be simplified to equation 5.11.

$$\text{RPVE of } q^{th} \text{ Signal} = \frac{\text{variance}(U_{M\times k} D_{k\times k}((A^q_{k\times 1} \boldsymbol{S}^q_{1\times n}) + J_{k\times n} k^{-1}))}{\text{variance}(U_{M\times k} D_{k\times k}((A^q_{k\times 1} S^q_{1\times n}) + J_{k\times n} k^{-1}))} * 100 \tag{5.10}$$

$$\text{RPVE of } q^{th} \text{ Signal} = \frac{\text{variance}(\boldsymbol{S}^q_{1\times n})}{\text{variance}(S^q_{1\times n})} * 100 \tag{5.11}$$

Equation 5.11 can be further simplified, as the variance of the reference dataset signals are scaled to have unit variance (section 5.4.1). From equation 5.12, the variance of the filtered signals can now be used as their RPVE. This still produces the same RPVE as equations 5.10 and 5.11 but simplifies the calculation.

$$\text{RPVE of } q^{th} \text{ Signal} = \text{variance}(\boldsymbol{S}^q) * 100 \tag{5.12}$$

5.6.5 PCA and ICA Performance Metrics

The difference between the total RPVE possible and the total RPVE of a dataset is used as the measure of overall performance metric for a dataset. It is presented in equation 5.13 for the PV and equation 5.14 for the signals. For example if 6 PV were separated, then the total RPVE possible would be 600%. If a dataset

captured a total RPVE of 550%, then the difference would be 50%. A small difference therefore indicates that a dataset has performed well. If the dataset were to be replaced with the reference dataset, then the difference would be zero. Note that neither metric has any weightings base on the PVE of the PV (or signals), each PV (or signal) is treated equally. To demonstrate potential differences when using PCA and ICA, both a PCA based and an ICA based performance metric are presented.

$$\text{PCA Performance (PCAP)} = (k * 100) - \sum_{q=1}^{k} \frac{\text{variance}(\boldsymbol{V_{1\times n}^{qT}})}{\text{variance}(V_{1\times n}^{qT})} * 100 \tag{5.13}$$

$$\text{ICA Performance (ICAP)} = (k * 100) - \sum_{q=1}^{k} \text{variance}(\boldsymbol{S}_{1\times n}^{q}) * 100 \tag{5.14}$$

5.7 Summary of Design Steps

The preparation of data, separation of signals, and how the performance metrics are applied to the data is summarised in tables 5.3 and 5.4. A flow diagram indicates how the outputs are linked together, from dimension reduction to the separation of signals in figure 5.6.

Step 1: Preprocess Data			
Step	Datasets	Details	
1.1	M	Bilinear Interpolation	
1.2	REM	Subset the Data Temporally	
1.3	REM	Remove Linear Trend	
1.4	REM	Remove Seasonal Cycle	
1.5	REM	Weight Cells By Latitude	
Step 2: Determine Number of PV			
Step	Datasets	Details	
2	R	Determine number of PV (k) by Rule-N method	
Step 3: Find Principal Vectors			
Step	Datasets	Details	Eqn.
3.1	R	$X_{M\times n} = U_{M\times k} D_{k\times k} V^T_{k\times n}$	4.3
3.2	EMG	$D^{-1}_{k\times k} U^+_{k\times M} \boldsymbol{X_{M\times n}} = \boldsymbol{V^T_{k\times n}}$	5.1
Step 4: Stability of Results			
Step	Datasets	Details	
4	R	Minimise stochastic effect by averaging mixing matrix (A)	
Step 5: Separate Signals			
Step	Datasets	Details	Eqn.
5.1	R	$V^T_{k\times n} = (A_{k\times k} S_{k\times n}) + J_{k\times n}$	4.4
5.2	EMG	$A^{-1}_{k\times k}(\boldsymbol{V^T_{k\times n}} - J_{k\times n}) = \boldsymbol{S}_{k\times n}$	5.2

Table 5.3: A summary of the first 5 of 8 steps of the preparation of the datasets for use in the performance metrics. Each step is performed with respect to datasets: R (reference), E (ERA-40), M (member), G (Gaussian noise). Emboldened matrices are a product of a non-reference dataset.

Step 6: PVE of Reference Dataset PV and Signals			
Step	Datasets	Details	Eqn.
6.1	R	q^{th} PV PVE $= \frac{(D^q)^2}{\sum\limits_{n} D^2} * 100$	5.3
6.2	R	$U_{M\times k} D_{k\times k}((A^q_{k\times 1} S^q_{1\times n}) + J_{k\times n} k^{-1}) = Z^q_{(M\times n)\times 1}$	5.6
		q^{th} Signal PVE $= \frac{\text{variance}(Z^q)}{\text{variance}(X_{(M\times n)\times 1})} * 100$	5.7
Step 7: RPVE of Other Datasets PV and Signals			
Step	Datasets	Details	Eqn.
7.1	EMG	q^{th} PV RPVE $= \frac{\text{variance}(\boldsymbol{V^{qT}_{1\times n}})}{\text{variance}(V^{qT}_{1\times n})} * 100$	5.9
7.2	EMG	q^{th} Signal RPVE $= \text{variance}(\boldsymbol{S}^q) * 100$	5.12
Step 8: Performance Metric Calculations			
Step	Datasets	Details	Eqn.
8.a	EMG	PCAP $= (k * 100) - \sum\limits_{q=1}^{k} q^{th}$ PV RPVE	5.13
8.b	EMG	ICAP $= (k * 100) - \sum\limits_{q=1}^{k} q^{th}$ Signal RPVE	5.14

Table 5.4: As table 5.3, but for steps 6 to 8. Step 8.b is not dependent upon the completion of step 8.a.

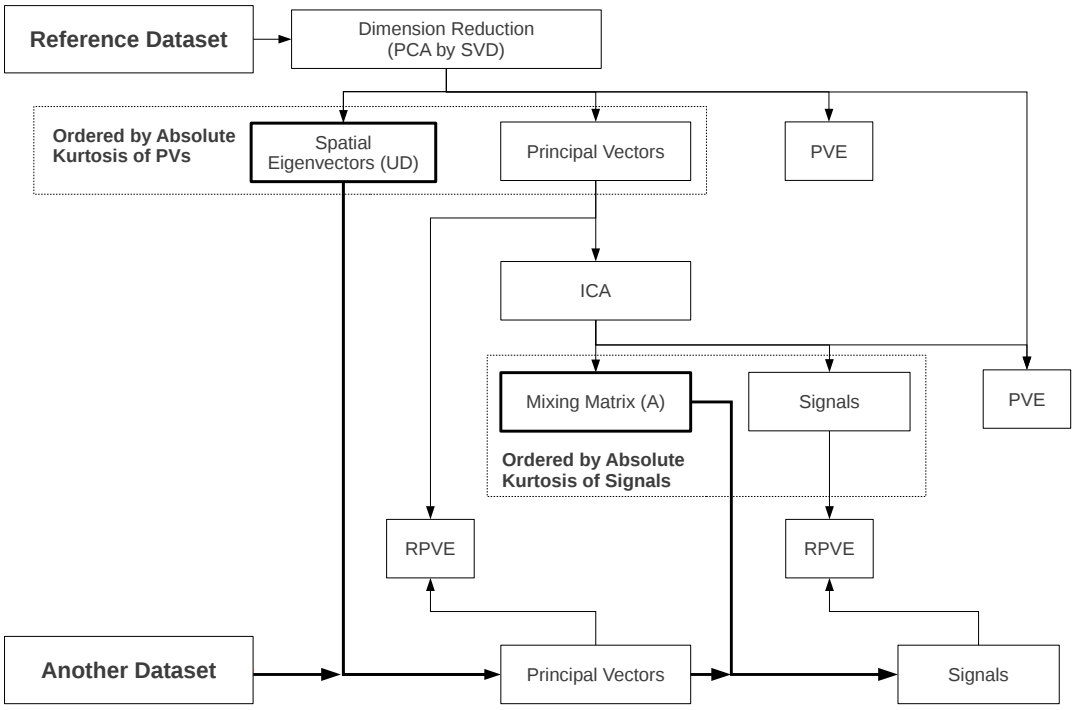

Figure 5.6: Flow diagram showing how the datasets are prepared for use by the performance metrics. This entails applying the results from the reference dataset to another dataset to find the PV and signals. The emboldened arrows indicate the results and process used for finding the PV and signals from a non-reference dataset.

5.8 Performance Metric and Data Limitations

Section 2.4 covers the limitations to model weighting methods, while section 4.5 discusses limitations to the general application of ICA and the data used in conjunction with it. This section covers the limitations specific to the performance metrics and data that are used in this work.

The performance metrics only use a single variable, geopotential height. The consequence of this selection, is that the performance results may change if a different variable is used. The same is true also if another reanalysis dataset were to be used. This makes the performance metrics sensitive to the variable and reference dataset used, as is discussed in section 2.4.1.

In addition to this sensitivity, only six climate model ensemble members are used.

A member which performs poorly, may appear as if it is an isolated case where in fact if more members were to be used then the member may not be such an isolated case as was originally thought. Therefore although the performance of a member is calculated with respect to a reference dataset, the rank (i.e.: 1^{st}, 2^{nd}, 3^{rd}, ...) of the member depends on the performance of the other members relative to it.

Due to the design of the performance metrics, the same spatial patterns that are found within the reference datasets are also used to find the PV and signals in the remaining datasets. The limitation with this approach is that it assumes that the identical spatial patterns found in the reference dataset are also found in the members, and that they are found in the exact same location. This may penalise members which simulate a mode correctly, but have a slightly different spatial location for a mode. How should a geographical shift in the mode effect the performance of a member is not discussed in this dissertation.

The preprocessing of the data (see section 5.3), may also affect the results. The extent to which this may occur is not known, but it is possible. For instance in Gleckler et al. (2008) they sometimes saw a non-negligible change in the ranking of their models when they changed the spatial grid resolution used. The linear trend and seasonal cycle were not removed from the Gaussian dataset, however these two steps would have altered its covariance structure and so could potentially impacted its performance as well. Dimension reduction could also favour models which simulated a mode incorrectly, but as the mode was not captured by the first few reanalysis PV, it would go unmeasured by the metrics.

Discounting a model result because the model performed poorly may be justifiable, but how poor is poor enough? If it is only a practical question of reducing the number of models to use, then the definition of *poor* is assessed in practical terms: Remove all the worst performing models until you are left with the desired amount. But it is unlikely that a model will fail to completely simulate a process if it is at least capable of simulating it. Therefore there may be a somewhat arbitrary threshold to decide if a model is indeed performing poor enough to warrant its exclusion. Determining the threshold may not be straight forward. Therefore, a limitation of the ICAP and PCAP metrics is that they do not provide infor-

mation on whether a member result performs poorly enough to warrant it being discarded.

5.9 Summary

The design of the performance metric is presented in this chapter. Measuring the degree to which signals in the reference dataset are found in model ensemble members provides a measure of performance for the members. Signals are found using an existing clustering technique, Independent Component Analysis (*ICA*, chapter 4). By finding the signals from reanalysis data they can automatically be assumed to represent modes, and therefore the performance metric does not require the more time consuming expert association of signals to known modes (section 5.4).

Two challenges in creating the performance metrics are discussed and addressed. The first, is that the algorithm used to perform ICA is partly stochastic. The consequence of this is that the algorithm can produce a different set of patterns when otherwise identical runs of the algorithm are used. In response to this, a method of averaging the results to minimize this concern is presented (section 5.4.2).

The second, is that Principal Component Analysis is used to reduce the dimensions of the data in preparation for its use with ICA. As the total Percentage of Variance Explained (PVE) of the principal vectors is equal to the total PVE of the signals separated from them, total PVE cannot differentiate between results produced from PCA and ICA. Therefore total PVE is an inadequate measure of performance. To solve this, the relative measure of variance formula is introduced.

The formula compares the variance of a pattern to the variance of the corresponding reference dataset pattern. It is designed to not be constrained in the same manner as the total variance measure. The PCA performance metric uses the total Relative Percentage of Variance Explained ($RPVE$) between reanalysis and dataset principal vectors. Similarly, the ICA performance metric compares the

RPVE between signals of the reference with the other datasets.

Overall performance for a model is not determined explicitly, but is rather estimated for the available simulations of the model. The total RPVE for each PV (or signal) from a dataset is used as the PCA (or ICA) performance metric. The interpretation of the performance metrics is that a smaller difference in RPVE implies that the dataset contains patterns of similar strength to those from the reference dataset. As the metrics are used for weighting datasets, they suffer from the same limitations as discussed in section 2.4.1.

Chapter 6

Dataset Performance Results

6.1 Introduction

The results of the performance metrics (chapter 5) are presented in this chapter. First, the choice on the number of signals and the robustness of them are presented in sections 6.2 and 6.3. This is followed by an investigation into how plausible the assumption is that signals and their spatial manifestations automatically represent modes (section 6.4). Section 6.5 discusses the performance of the datasets when using PCAP and ICAP for geopotential height data. Section 6.6 investigates the extent to which the performance of the alternate reanalysis dataset is sensitive to a change in geopotential height level and the number of signals separated from the data. Section 6.7 looks at performance sensitivity to a change in variable (near surface air temperature), and compares the results to a variance based metric (Fourier Distance) and a mean based metric (bias). A discussion of the results follows in section 6.8.

6.2 Number of PV and Signals

The Rule-N method (section 4.3) was applied to all the PV from the reference dataset in equation 4.3 to determine the number of PV (k) to retain for further

analysis. While the minimum number of mixtures supported by the ICA model is two (section 4.2), the potential number of PV were from 3 to 360 (n). Three was used as the minimum set size as it was deemed to be the first non-trivial set size. The Rule-N method found the first 3 to 51 PV to be above the level of noise. By design, the same number of signals would also be separated from the data. Although this work is not concerned with the association of signals to modes, the number of signals to be extracted will not be the maximum number possible. This is done in order to better facilitate comparisons to similar works. Similar works range in the number of signals separated from data, from 120 separated by Basak et al. (2004), to the more expert intensive and smaller sizes such as 10 by Aires et al. (2000) and 4 by Westra et al. (2010). The range in the number of signals separated from data appears to be due to the exploratory nature of the research in general, where even within ICA there are a number of different ways of determining the number of signals to separate (see section 4.3).

To further reduce the number of PV, an additional step was introduced which measures the degree to which the signals are independent of each other. The convergence of the signals to wards independence is measured using the neg-entropy threshold of the FastICA algorithm. Set sizes with good convergence (low thresholds) are desirable in this work as they are more independent of each other. For each set size (3 to 51), the threshold for each of the sets of signal was recorded. Each set size used 1000 runs of the FastICA algorithm.

In figure 6.1, the median thresholds for the different sizes are shown. The median thresholds generally indicate good convergence (low thresholds) from sizes 3 to 6. However, for the set sizes in the 7-40 range the sets often do not converge, which may be a result of overestimating the number of source signals in the data. Interestingly, from around a set size of 41 onwards, the median thresholds are once again close to zero. Rather than being representative of convergence, this may be an artefact of the algorithm caused by the overestimation of the number of signals in the data.

The set size corresponding to the lowest median threshold with no outliers is the set of 4. This number is similar to Kent (2011), which found a set size of 6 when using monthly mean surface temperature data over a different period. The data in

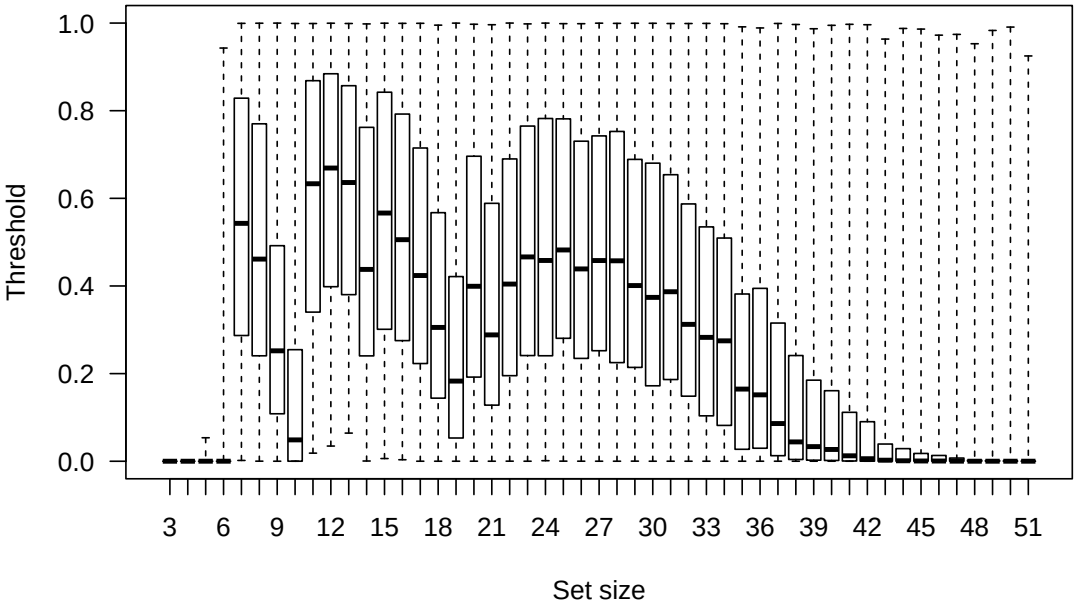

Figure 6.1: The median thresholds for the 3 to 51 PV determined by the Rule-N method. The limits of the box and whisker plots extend to include outliers, while the boxes encompass the 25th and 75th percentiles. A low threshold value indicates near independence between a set of signals.

Kent (2011) included the seasonal cycle which was represented by 2 signals. This suggests that had they removed the seasonal cycle then they would have arrived at the same number, thereby adding supporting evidence for the validity of the number found in this dissertation. Results from applying the mapping stability algorithm by Kent (2011), also support this result as the set size of 4 was found to be the most stable (not shown), while larger set sizes always produced less stable results.

6.3 A Robust Set of Signals

Having determined the number of signals to separate from the data, any effect that different initial stochastic estimates may have on the signals should be minimised as discussed in section 5.4.2. This is achieved by averaging the unmixing matrix over a number of different runs of the FastICA algorithm when only a change in the initial stochastic estimate is allowed.

The standard deviation of the 5 averaged mixing matrices for the different number of runs is shown in figure 6.2. The figure shows that with an increase in the number of runs, the standard deviation between the averaged unmixing matrices of a set decreases. This is to be expected as by averaging over more runs would better cancel out any differences in the initial stochastic estimates.

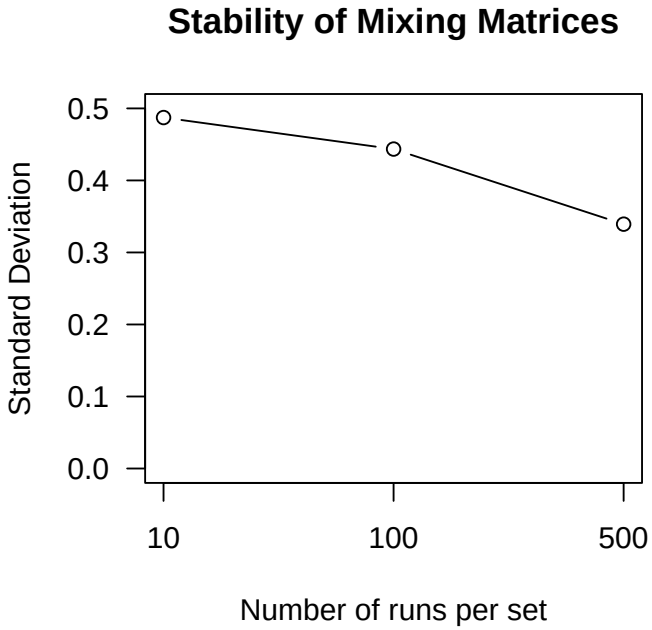

Figure 6.2: The standard deviation amongst the 5 mixing matrices that were each created by averaging the matrices over a number of runs.

For almost all runs and sets the signals converged. This may be because the threshold was shown to be reached in all runs in figure 6.1, when the number of signals was determined (section 6.2). The only exceptions to this were from the size of 500 runs, where 2 runs from one set of 500 did not converge, and for another set 3 runs did not converge. These runs were eliminated from the final averaging. As the set of averaged unmixing matrices from using 500 runs was found to have the least RMSE, one of the averaged matrices was selected to then find the signals in the reference dataset (step 4 table 5.3). As all 5 of the averaged matrices for the run are equivalent, any one from the set could have been selected.

Figure 6.3 shows the degree to which using the averaged unmixing matrix im-

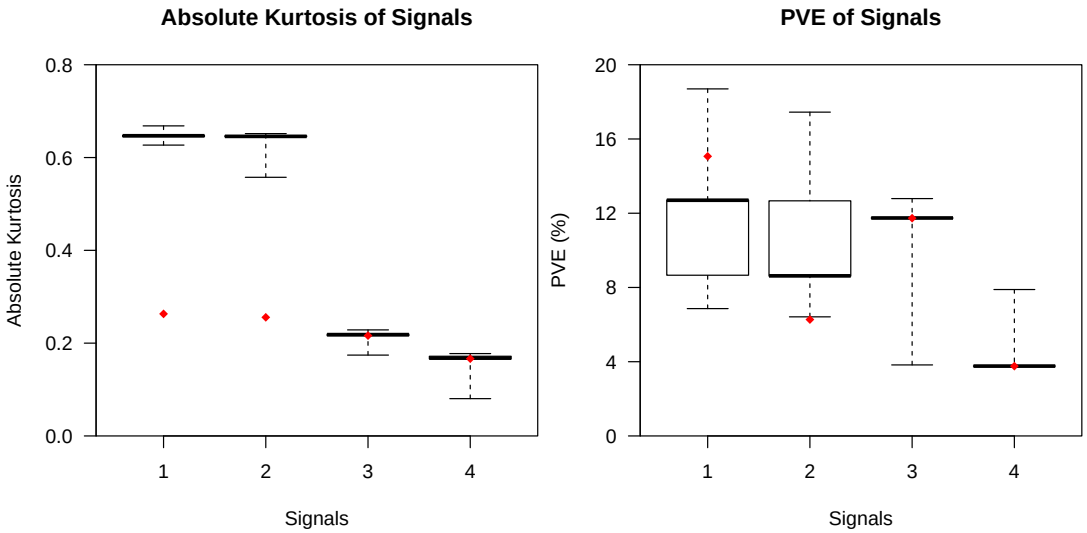

Figure 6.3: (*left*) absolute kurtosis, and (*right*) PVE of the signals for batch 1 with 500 runs. The whisker plots indicates the variability of the values over 500 runs without applying the average matrix, and their range extends to include outliers. The diamonds (red) indicate the values for the signals from the reference dataset using the average matrix. All signals are ordered by decreasing absolute kurtosis.

pacted the absolute kurtosis and PVE of the signals. The results of the averaged unmixing matrix are contrasted with the variability taken from batch 1 over 500 runs. In the figure, the absolute kurtosis for the first two signals from the reference dataset fall well below the variability of all the signals produced without the averaging process. However, the second two signals fall within the absolute kurtosis range of the signals produced without the averaging process. This shows that there can be a trade off between the absolute kurtosis of signals and their stability. Namely, a more stable set of signals can be produced but at the cost of some them being more Gaussian.

In contrast, the PVE of the signals using the averaging process from the reference dataset fall mainly within the variability of the signals produced without the averaging process. Signal 2 from the reference dataset falls just below the range of the non-averaged signals, which may be a product of having too few runs to with which to construct the PVE variability of the signals. As total PVE for the set of signals is constrained by design (section 5.4.2), it is not surprising that the

PVE of the signals produced using the averaged unmixing matrix, falls within the range of the runs.

6.4 Plausibility of Patterns

In this dissertation, the results (*patterns*) from applying PCA and ICA to reanalysis data are automatically assumed to be representations of modes of climate variability (section 5.4). The motivation for this assumption, is that associating patterns to modes of climate variability is a complex and subjective task (see section 4.4).

While this assumption is not proved in this dissertation, this section provides evidence for its plausibility. As the focus of this work is on developing an ICA-based performance metric (section 5), only the ICA related patterns are discussed. PCA related patterns are left for future work.

The signals and their corresponding spatial manifestations are the ICA related patterns. The signals are maximised to be non-Gaussian by design, and therefore they are plausible in the sense that they are constructed to be unlike Gaussian noise. The absolute kurtosis of the signals separated from the reanalysis dataset, show in figure 6.3 that they are to some extent unlike Gaussian noise and are therefore at least plausible.

The spatial manifestation that accompanies a signal is termed a *static map* in this work. Each static map is calculated as follows:

$$L^q_{M\times 1} = U_{M\times k} D_{k\times k} A^q_{k\times 1} \tag{6.1}$$

In equation 6.1, the q^{th} static map (L^q) is calculated using the U and D matrices (equation 4.3) and the q^{th} column of the averaged unmixing matrix (A, section 5.4.2). The resulting $L^q_{M\times 1}$ matrix is then transformed into a 2D matrix (144×73, see section 5.2). The latitude weighting from step 4 of the preprocessing steps (section 5.3) is then reversed, and the resulting matrix is plotted as an image.

The static maps for signals 1 and 2 can be seen in figure 6.4, while the static maps for signals 3 and 4 can be seen in figure 6.5.

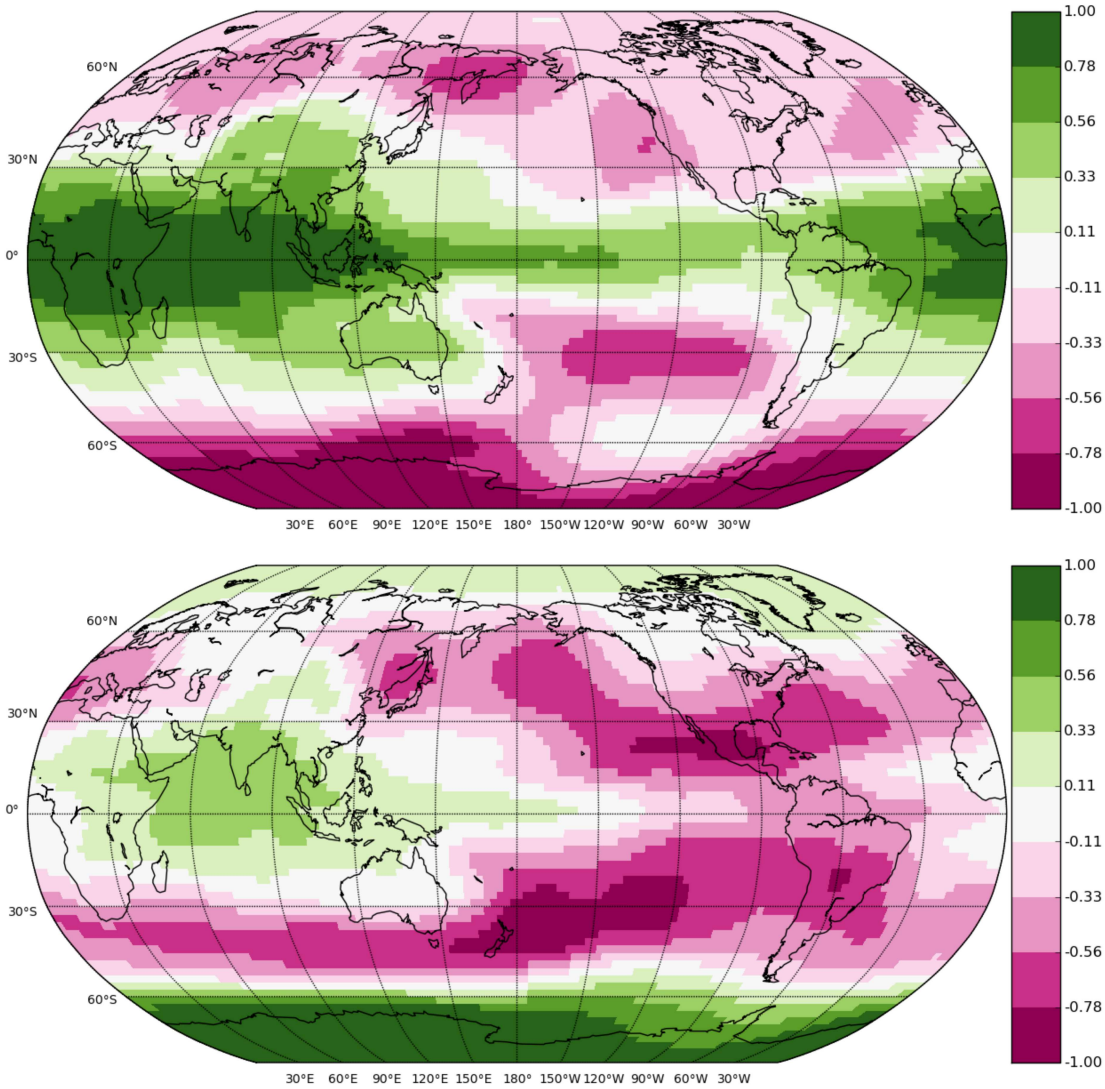

Figure 6.4: The static maps corresponding to signals 1 (*top*) and 2 (*bottom*) from the reference dataset for 700 hPa geopotential height. Images are scaled to have a range of between -1 and 1 due to sign and variance ambiguity (section 4.2).

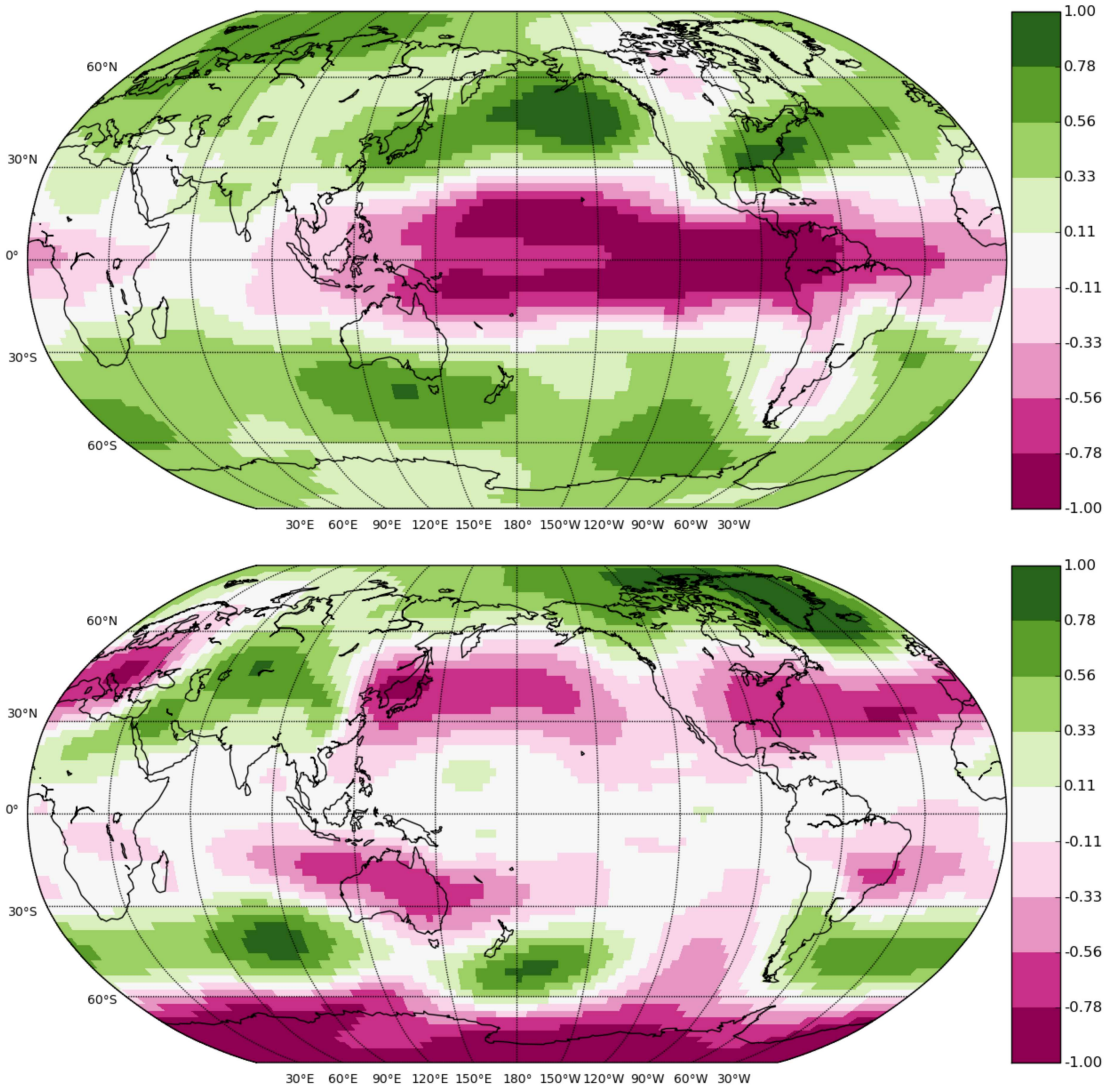

Figure 6.5: As figure 6.4 but for the static maps of signals 3 (*top*) and 4 (*bottom*).

In terms of the plausibility of the static maps (figures 6.4 and 6.5), the maps show geographical groupings of points with some hemispherical symmetry. These suggest that the images are not representative of Gaussian noise, which would show images with uniform distributions of points. Additionally, Fodor and Kamath (2003) show in their figure 7 (reproduced in figure 6.6), the result of static map equivalent images with artefacts due to overlearning. As the static maps in this work do not exhibit the same artefacts, and are not visually similar to Gaussian noise, they are at least plausible representations of modes.

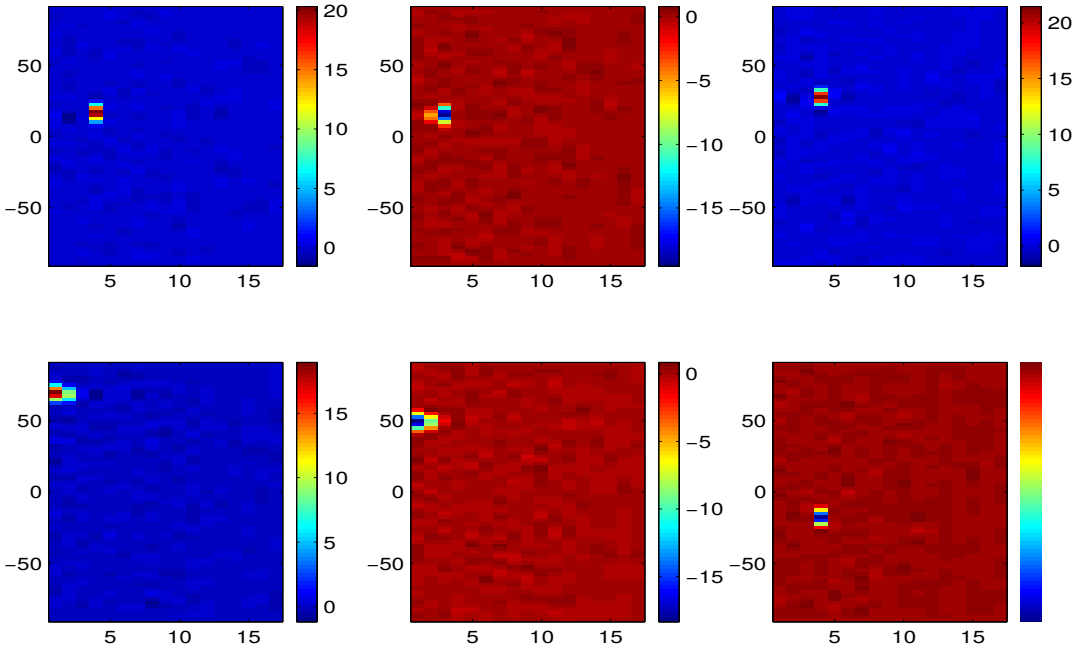

Figure 6.6: The six independent component basis (static map equivalents) obtained from the full-dimensional anomaly data. Caption and image from figure 7 of Fodor and Kamath (2003).

6.5 PCA and ICA Performance Metric Results

A plot depicting the performances of the datasets is shown in figure 6.7 and the corresponding values are in table 6.1.E4 and G obtained the best and worst performance respectively by PCAP and by ICAP. This was to be expected based on the assumptions that E4 should closely reflect the real climate and G contained only noise. The multi-model mean performed poorly in ICAP and PCAP as it does not capture variance by design, and so it is not shown in the figure.

PCAP and ICAP rank the members very similarly except for M1 and M2. The reason for the difference of the two members is unknown. M1 and M2 have the highest total PVE out of the members which may allow them to capture the signals better than the other members. However, total PVE alone does not account for E4 having less total PVE than M1 and M2 while still out performing them. What this does show is that despite variance being preserved (section 5.6) PCAP and ICAP may not only differ in theory but can also differ in practice.

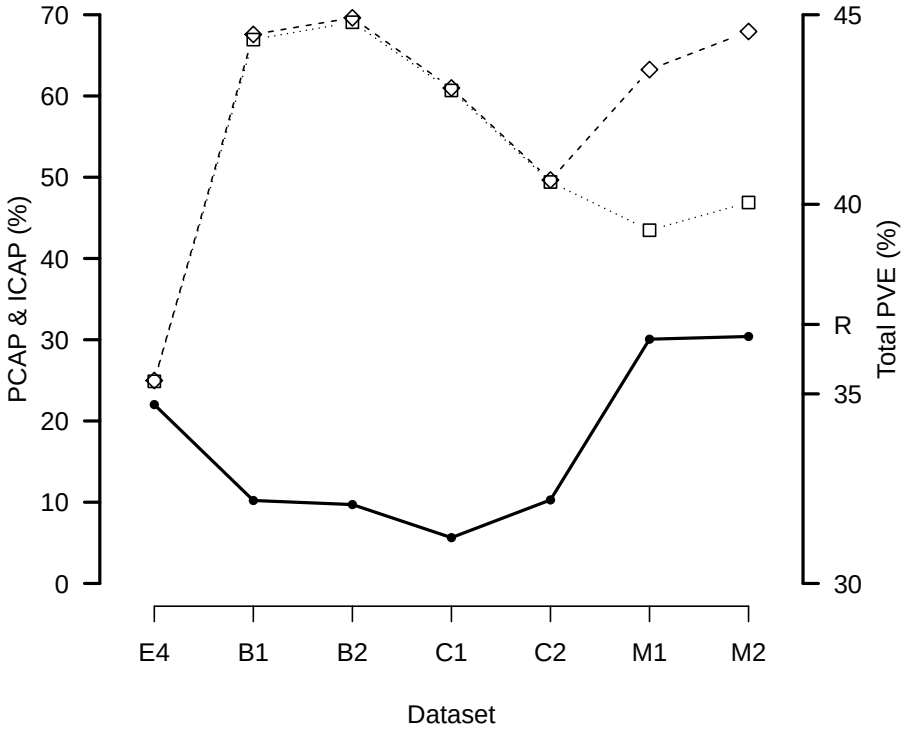

Figure 6.7: PCAP (diamonds) and ICAP (squares) results. The PVE are indicated by the solid points. Dataset order is arbitrary and the Noise and MM datasets are not shown due to their poor performance. "R" Indicates the total PVE of the reference dataset (37%).

Dataset	**Total PVE (%)**	**ICAP (%)**	**PCAP (%)**
R	36.83	-	-
E4	34.72	24.88	24.98
B1	32.19	66.95	67.57
B2	32.08	69.08	69.62
C1	31.21	60.68	60.99
C2	32.20	49.41	49.66
M1	36.44	**43.48**	**63.25**
M2	36.51	**46.89**	**67.94**
MM	7.59	331.47	331.56
G	0.04	399.28	399.28

Table 6.1: The total PVE, ICAP, and PCAP results. M1 and M2 (emboldened) differ the most when changing between ICAP and PCAP. The G and MM datasets results are included. Order of results is the same as in figure 6.7.

C1 and C2 have the greatest performance difference between members from the same model. To investigate this further, the Fourier decomposition of their spatially averaged time series (30 years, n=360 months) using preprocessed data, is shown in figure 6.8. While figure 6.8 shows that C1 and C2 have the highest variance occurring at the same frequency (1), they do have difference frequencies for the second highest variance (13 and 9 respectively).

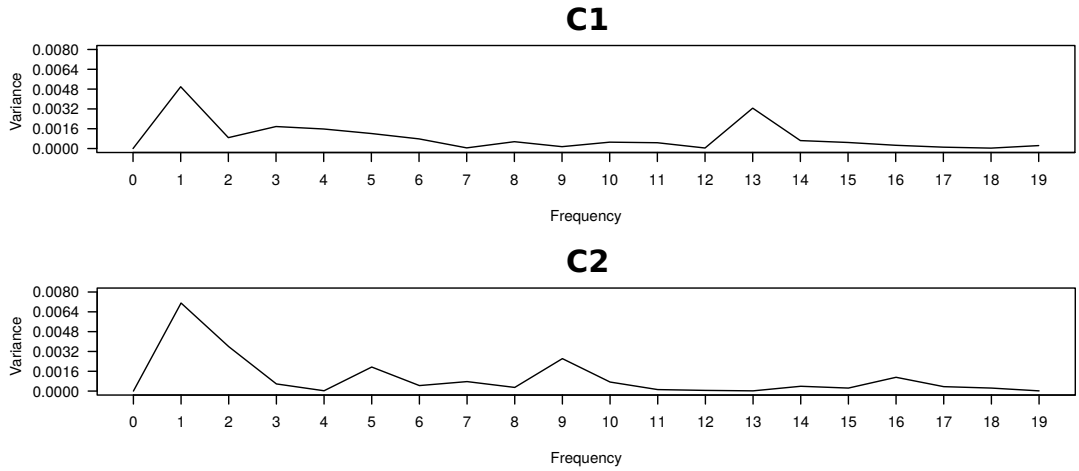

Figure 6.8: The Fourier decompositions of the spatially averaged time series for C1 (*top*) and C2 (*bottom*) over the 30 year period (360 months). The first 20 frequencies (0-19) and corresponding variances are shown, the remaining frequencies have negligible variance.

The variance for each grid point is used to plot geographical distributions of the variance at the different frequencies in figure 6.9 to highlight potential spatial differences. From the figure, the variance from C1 is primarily focused around the equator while C2 has a greater geographical representation of variance. This difference in geographical representation may be a factor behind the difference in their performance.

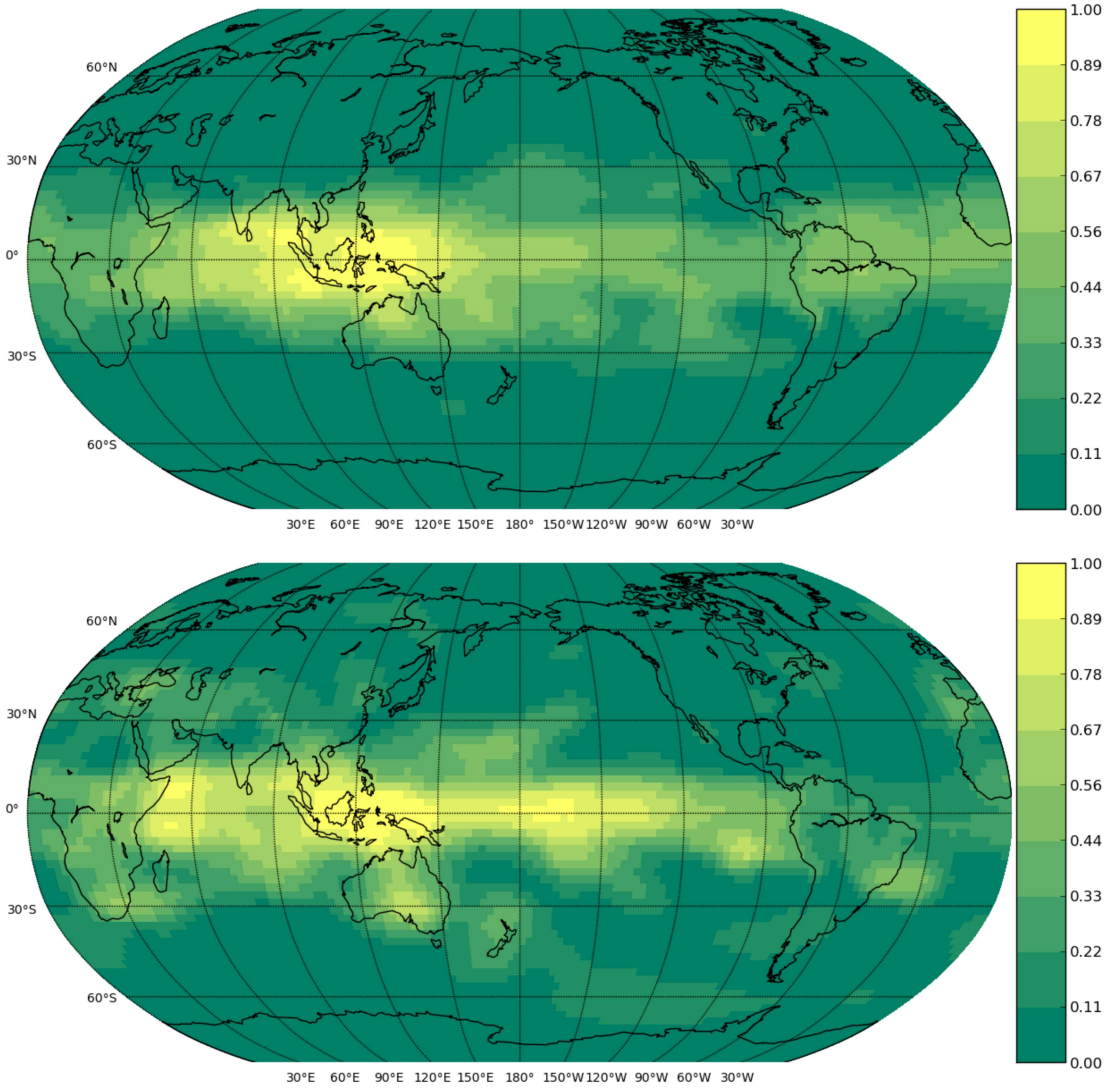

Figure 6.9: The geographical distributions of the variance at frequency 13 from C1 (*top*), and for frequency 9 from C2 (*bottom*). Images are scaled to unit variance.

6.6 Reproducibility of Reanalysis Performance

A consequence of the design of the PCAP and ICAP metrics (section 5), is that an alternative reanalysis dataset should generally perform better than other model datasets, as reanalysis data is generally more constrained by observations than model simulations. Evidence for this performance difference can be seen in ta-

ble 6.1, where the alternative reanalysis dataset (E4) performed better than the model simulations for both PCAP and ICAP.

However, to ensure that this better performance is not an isolated case, this section investigates the degree to which PCAP and ICAP results are sensitive to changes. Specifically, a change in the geopotential height level used, and the number of PV and signals that are separated from the data. Demonstrating that the metrics have a low sensitivity to these changes will help show that they are consistent and therefore potentially useful as model performance metrics beyond this work.

For the data, the reference and alternative reanalysis dataset are the same as those in section 5.2. However, the period under investigation is reduced by a year due to data availability. The available period is from Jan 1961 to Dec 1989 (29 years, 348 months). 20 additional initial condition ensemble members are used (HadCM3 from CMIP5, Met Office Hadley Centre (2017)) in combination with the existing datasets discussed in section 5.2. The additional members consist of two sets of ten initializations. Each set uses a different initialization, either number 2 or 3 (e.g: r1i2p1 or r1i3p1). All datasets are preprocessed as in section 5.3, but are done using the shorter time period. The Gaussian noise and multi-model mean datasets are also recalculated. Shortened dataset nomenclature is also used for the additional datasets, for example with H6:3 referring to r6i3p1 and H9:2 referring to r9i2p1.

500 hPa and 700 hPa are used for the two different geopotential height levels. 700 hPa is used to facilitate any comparisons to the previous section (section 6.5), while 500 hPa is used due to its availability and its adoption in other works such as Wallace and Gutzler (1981) and Christiansen (2009). Set sizes of 4 and 6 are extracted from both the 500 hPa and 700 hPa levels. The set sizes are determined using the same approach as described in section 6.2, with the resulting median thresholds presented in figure 6.10.

Figure 6.10 shows that for 700 hPa the largest set size with the lowest median value was 5. However, to facilitate comparisons to the original reference dataset at 700 hPa, a set size of 4 was also chosen for use in this section. The mapping

stability from Kent (2011) (not shown) also showed that the most stable set occurred at a size of 4. For 500 hPa, the largest set size with the lowest median threshold was 6 and this set size was also confirmed by the mapping stability calculation (not shown).

PCAP and ICAP were calculated for each dataset, using the set sizes of 4 and 6 and the geopotential height levels of 700 hPa and 500 hPa. In figures 6.11 and 6.12 the PCAP and ICAP of the datasets are indeed shown to be sensitive to changes in the set size and level used. This is supported by the general clustering of the datasets by set size and some changing of performance for a dataset between levels. For example, H7:3 differs between 700 hPa and 500 hPa when using a set size of 6.

In terms of the performance of the alternate reanalysis dataset (E4), it generally performed better than other member datasets. The exception to this can be seen in the 500 hPa plot (figure 6.12), with a set size of 4. For example, member M1 performed better by ICAP than E4. This may be because the set size of 4 is not the most stable for the 500 hPa level. This suggests that PCAP and ICAP can generally differentiate between reanalysis data and member datasets.

Visual clustering of the members shows that they strongly clustered by set size, with no overlap between members of the two different set sizes. For 700 hPa and a set size of 4, the members generally perform slightly better in PCAP than in ICAP, while for a set size of 6 they perform notably better in PCAP than ICAP. This difference between PCAP and ICAP is also present for the 500 hPa level using a set size of 4. However, this time the members generally perform better in ICAP than PCAP. For a set size of 6, the members are clustered approximately equally between the two metrics. This suggests that while the member datasets are impacted by changes in the set size and level used, PCAP and ICAP can be used to differentiate between member datasets less well when stable set sizes are used.

The visual clustering of members by their constituent models also shows that they are sensitive to set size and level changes. For example, cluster C for 700 hPa contains members C1 and C2 and out performs cluster B in terms of PCAP

with a set size of 6. This same performance difference cannot be seen for the 500 hPa level.

Although not explicitly investigated, ICAP was also found to be sensitive to a change in the period used. For the time period of 30 years (360 months) M1 has an ICAP of approximately 43% (table 6.1) making it the best performing member. However, for the reduced period of 29 years (348 months), the member performs worse with approximately 60%, changing its rank from first to 4^{th} behind C2, M1, and C1.

The performance of the datasets in this section are shown to be sensitive to changes in the set size and geopotential height level used. This is consistent with other research (section 2.4.1), which states that performance can be strongly tied to the variable used in the analysis. Additionally, the ranking of member datasets by ICAP was shown to be sensitive to a change in the data period used. Despite these sensitivities, the alternate reanalysis dataset consistently performed better than the majority of members for both set sizes and levels.

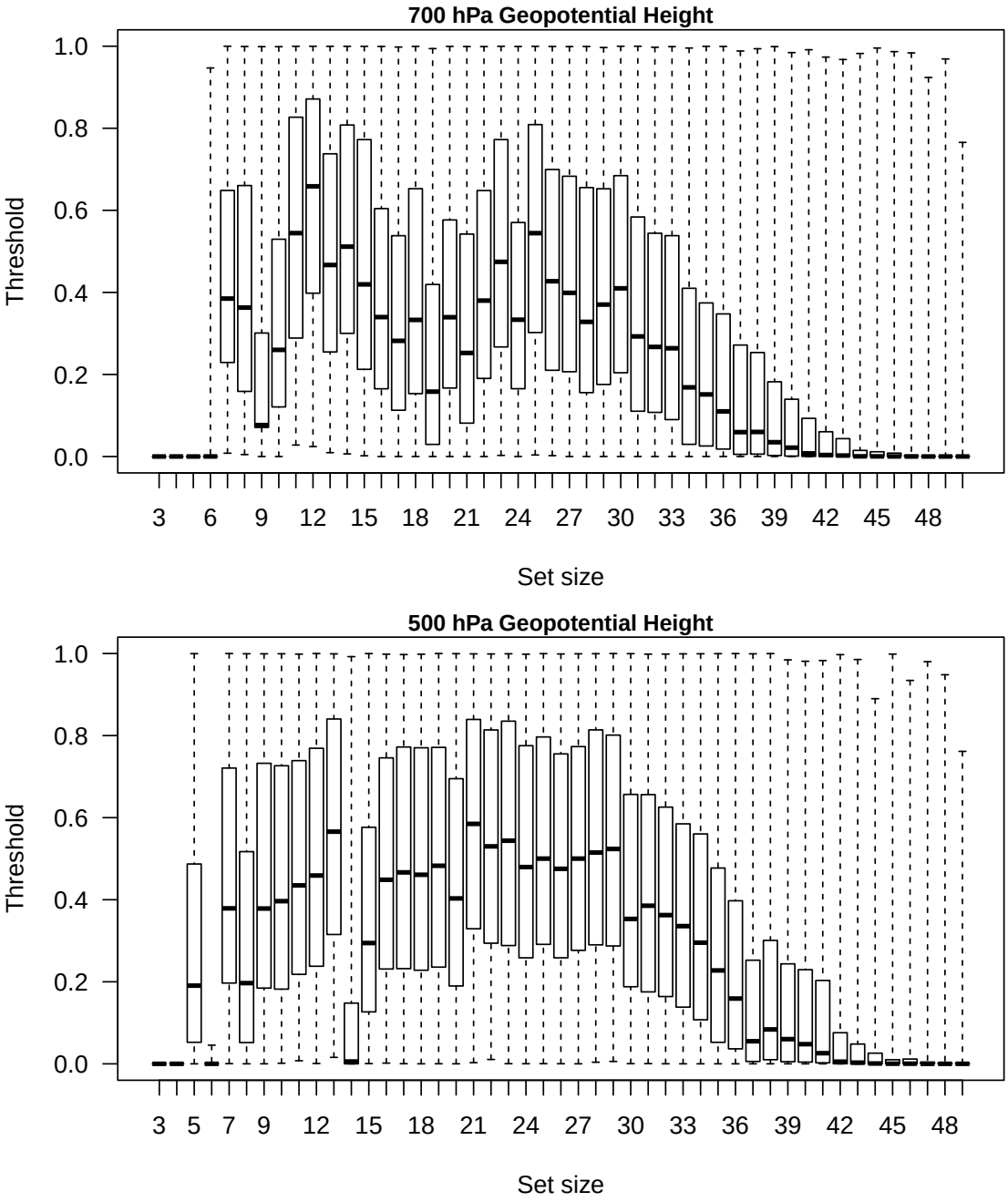

Figure 6.10: Thresholds for the different geopotential height levels, 700 hPa (*top*), and 500 hPa (*bottom*). The datasets have a maximum set size of 50 and 49 respectively. The limits of the box and whisker plots extend to include outliers, while the boxes encompass the 25th and 75th percentiles. A low threshold value indicates near independence between a set of signals.

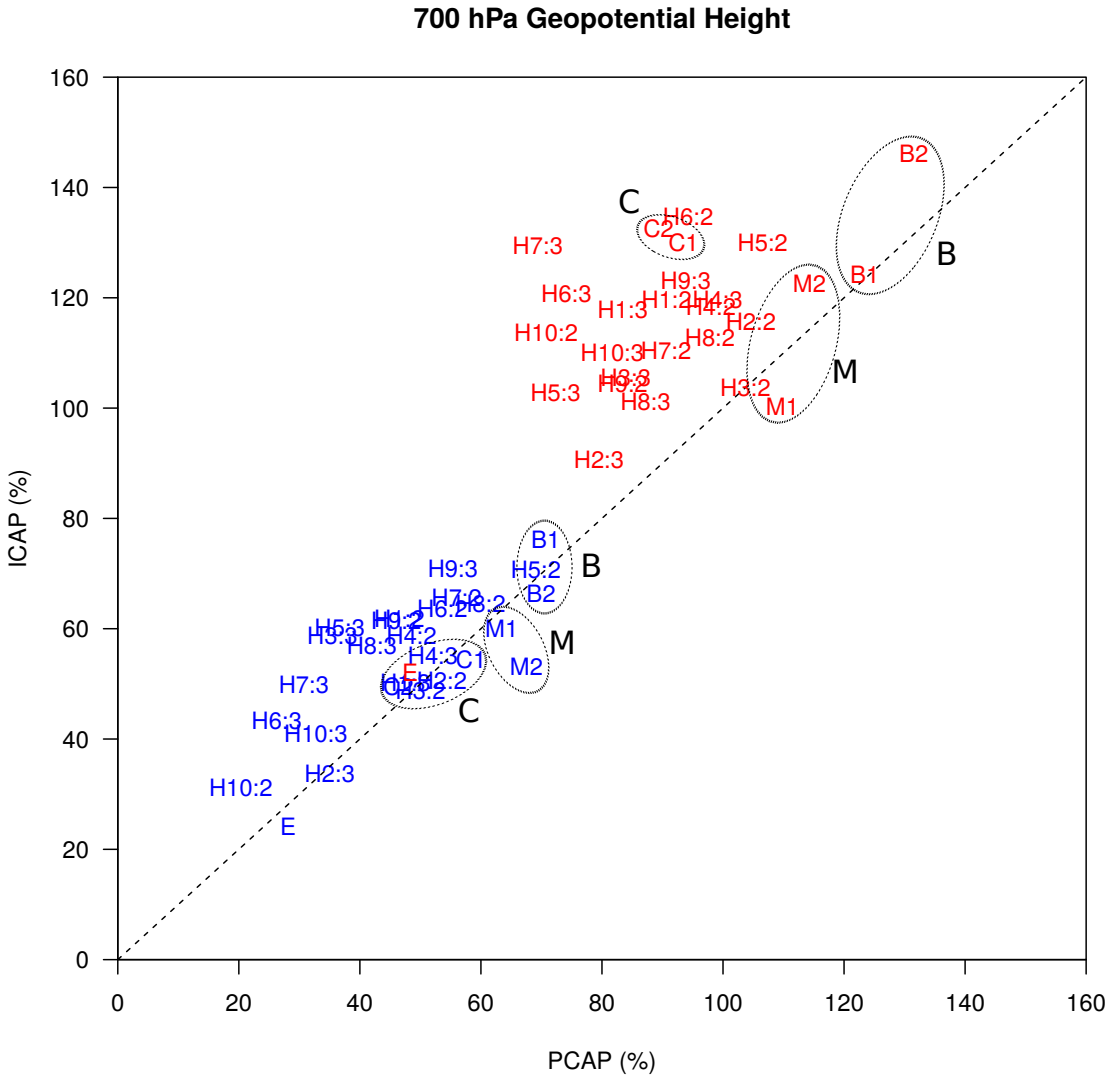

Figure 6.11: The PCAP and ICAP of each dataset for 700 hPa geopotential height. Dashed diagonal line indicates where PCAP equals ICAP. Dataset colour refers to PV and signals which use set a size of 4 (blue) and 6 (red). Clusters are visual groupings of members from the same model. For example, B refers to a cluster of B1 and B2 but may contain other datasets. MM and G performed poorly and are therefore not shown.

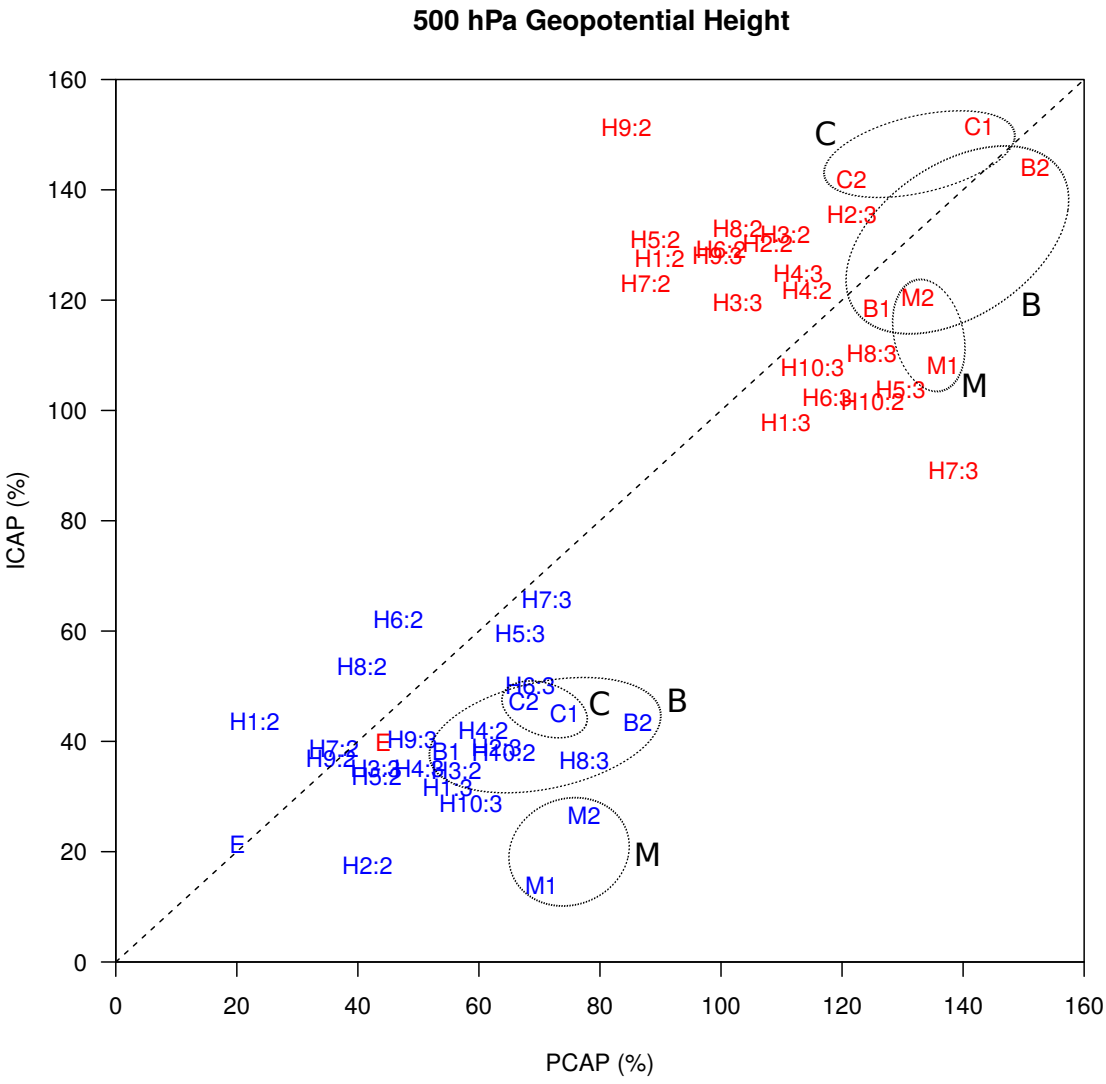

Figure 6.12: As figure 6.11 but for 500 hPa geopotential height.

6.7 Metric Sensitivity

As the performance of datasets may be dependent on the variable used (section 2.4), this section examines how sensitive the performance of datasets are when using PCAP and ICAP and compares them to the performance sensitivities of two other metrics. Specifically, sensitivity will be examined primarily in terms of dataset rank, and to a lesser extent their relative performance. Using two additional metrics will help determine how much more or less robust PCAP and ICAP are compared to the other metrics. Ideally, developing more robust metrics would help by removing the choice around which variable should be used to assess dataset performance.

The alternate variable and the application of PCAP and ICAP to it, are discussed in section 6.7.1. Sections 6.7.2.1 and 6.7.2.2 discuss the two other metrics and their comparison to PCAP and ICAP.

6.7.1 PCAP and ICAP with Near Surface Air temperature Data

For the alternative variable, near surface air temperature data (TAS) is used. All the datasets are the taken from the same sources as in section 5.2 and are preprocessed in the same manner as the GHT variable (700mb, section 5.3). As the performance of the noise and multi-model mean have already been examined (section 6.5), they are excluded from further analysis.

The number of PV above the level of noise was determined to be 68, and the set size with the lowest convergence threshold was 3 (not shown).The PV and signals were separated from the data using an averaged unmixing matrix (batch 1 with a run size of 10), and PCAP and ICAP were calculated. The performance of the datasets can be seen in figure 6.13.

Figure 6.13 shows that the performance of the datasets are generally the same for both PCAP and ICAP, with only M1 and M2 being different. Compared to the GHT performances (section 6.7), there are a few differences. The first is that

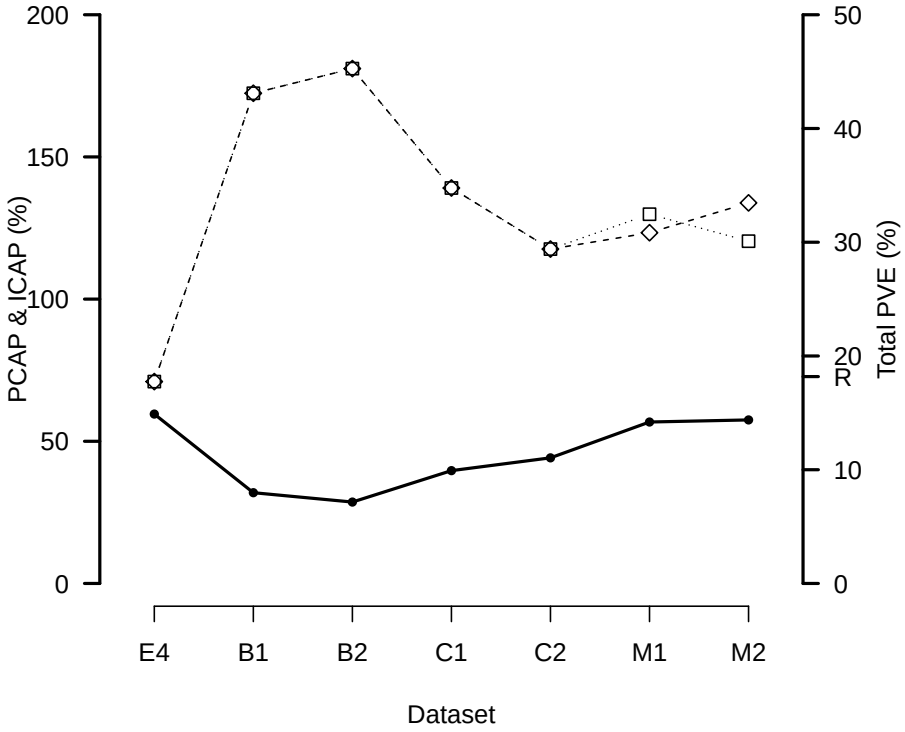

Figure 6.13: PCAP (diamonds) and ICAP (squares) results for TAS data. The PVE are indicated by the solid points. Dataset order is arbitrary and "R" Indicates the total PVE of the reference dataset (18.5%).

less variance (PVE) was captured by the PV of TAS than those of GHT (18.5% vs 37%). This is to be expected as there were less PV found to be stable for TAS than for GHT. The second, is that the datasets all performed worse (higher PCAP and ICAP values) compared to GHT performance.

Figure 6.14 shows PCAP and ICAP values for both GHT and TAS. E4 consistently performs better than the other datasets for both metrics and variables. This is to be expected, as by design it is more constrained by observations than the member datasets. With the exception of M1 and M2, all the remaining datasets remain in the same order for both metrics and variables. The differences between M1 and M2 are investigated further in section 6.7.2.1.

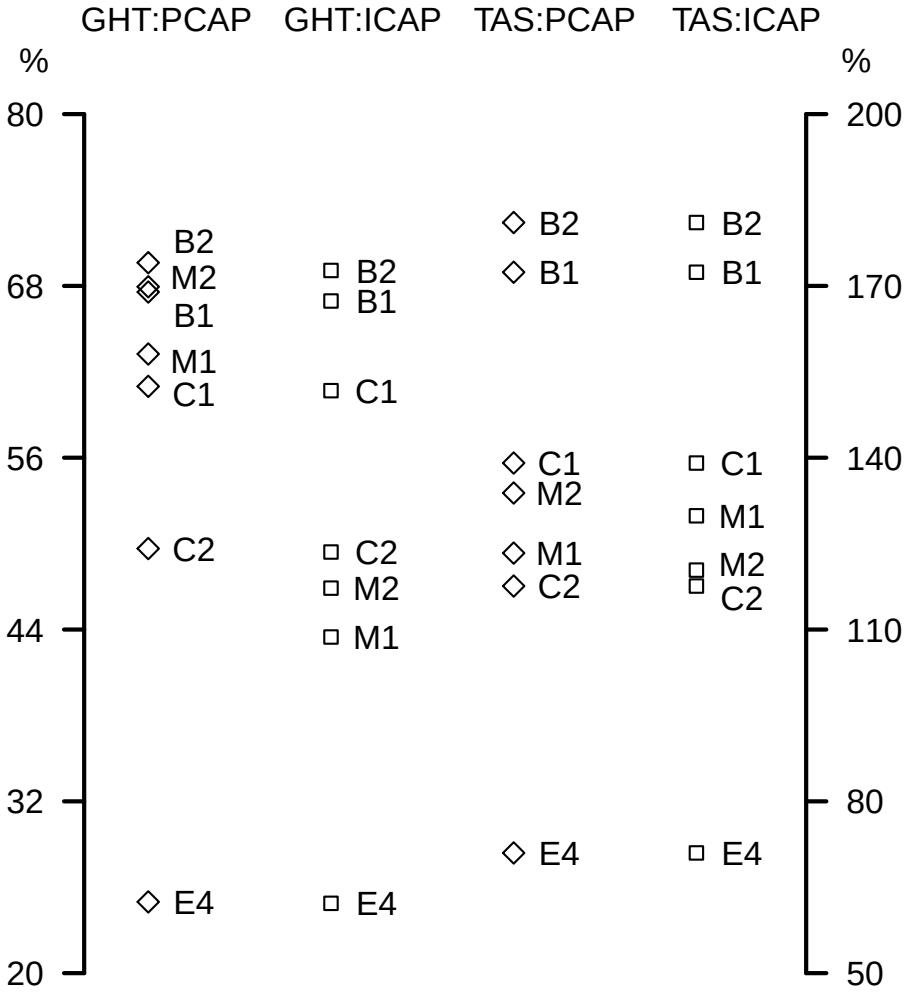

Figure 6.14: PCAP and ICAP for both GHT and TAS data. Note that the GHT uses left the y-axis, while TAS uses the right y-axis.

6.7.2 Alternative Metrics

The decision of which metrics should be used to compare against PCAP and ICAP is not clear. While some of the potential existing metrics are discussed in section 2.3, in order to attribute any performance differences to only be a product of the metric used, the same preprocessed data should also be used by the two other metrics. Further complicating this issue, is that the metrics may measure fundamentally different aspects of the climate. For example PCAP and ICAP focus on variance, while the REA method (section 2.3.1) focuses on the mean

state of the climate.

Due to these complexities, the first alternate metric is a novel measure designed to assess dataset performance using variance like PCAP and ICAP. It provides a potential solution to the above mentioned problems as it also uses the same preprocessed data and uses variance as part of the assessment process as well. However, as it is a novel measure and therefore not in the existing literature, a second metric is also used. While the second metric uses the mean state of the climate and not variance, it is well documented in the existing literature. Therefore, it will serve as more recognised metric to contrast PCAP and ICAP sensitivity results against.

6.7.2.1 Fourier Distance

The first metric is termed the *Fourier Distance* (*FD*). To determine the FD of a dataset, a timeseries of length n is made by taking the spatial average of the pre-processed data for each point in time. The Fourier decomposition is then created from the timeseries. The covariance of the dataset with timeseries of the reference dataset is calculated. Lastly, the covariance of the Fourier decomposition of the reference dataset with itself, is then subtracted from that of the dataset. The result is a measure of how closely the variance of the frequencies from the dataset match those of the reference dataset. The less the FD of a dataset, the more closely it resembles the reference dataset.

Fourier decomposition is used, as Taylor et al. (2012) state that model data cannot be expected to match the timing of events in reanalysis data. This is because climate model simulations are typically influenced by both boundary conditions (e.g.: solar forcing) and a prior determined state that is taken as the model has been run to reach equilibrium. The influence of the prior state on a model, makes it unlikely for its events to occur at the same time as those in the reference dataset (e.g.: ENSO events). The consequence of this is that while a timeseries from dataset may have a similar standard deviation when compared to the reference dataset, its correlation with respect to it may be low due to the non-synchronous timing of events. The Fourier decomposition of the timeseries,

therefore allows for the frequency of events to be compared with out worrying about potential phase (timing) differences.

The absolute FD is used as it functions in a similar manner to PCAP and ICAP, giving better performing datasets a lower score. The main data differences with FD compared to PCAP and ICAP, is that it uses globally averaged data while PCAP and ICAP do not. FD also uses the preprocessed data (steps 1 to 5) while PCAP and ICAP use the preprocessed data with dimension reduction. It is similar to PCAP and ICAP, in that it examines variance compared to the reference dataset. The FD for both the GHT and TAS data is presented in figures 6.15 and 6.17.

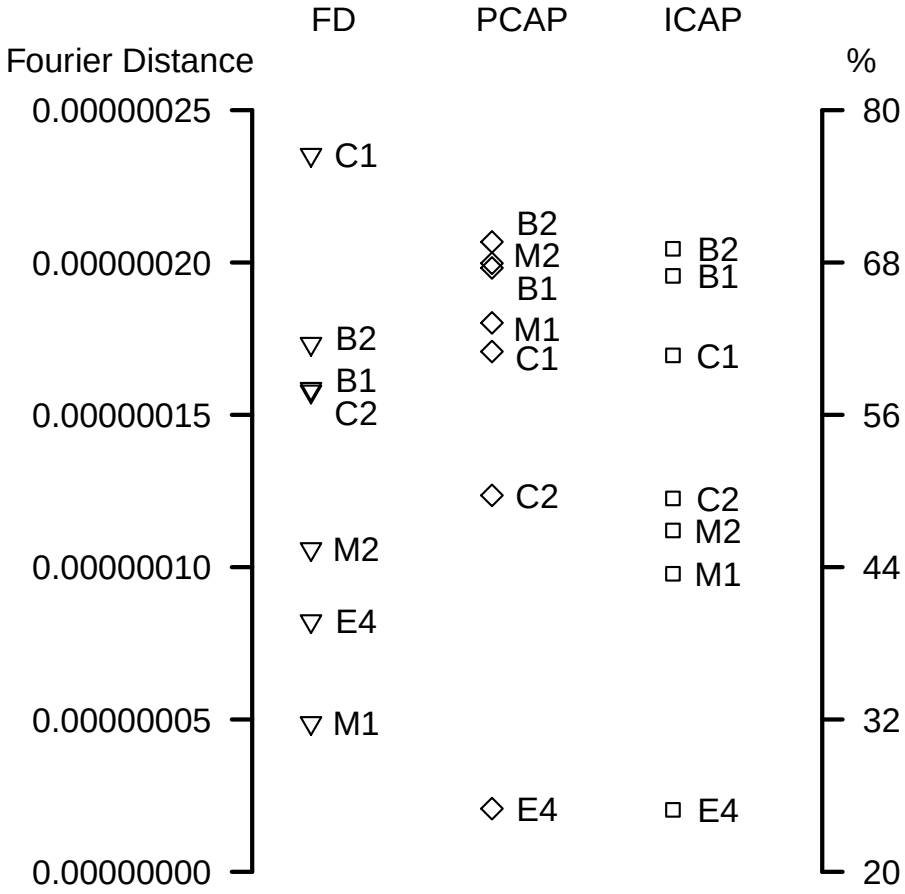

Figure 6.15: The Fourier Distance, PCAP, and ICAP results using GHT data. FD uses the left y-axis, while PCAP and ICAP use the right y-axis.

In figure 6.15, M1 is shown to out perform E4 in terms of FD. The reason for this

can be seen in figure 6.16, where M1 has a much greater variance at frequency 2 than both E4 and R. This overestimation in variance results in its improved FD performance over E4. If just the datasets from models are examined, the order by FD and ICAP are very similar. The only exception to this is C1, which has a lower rank by FD compared to PCAP and ICAP. The comparison between C1 and C2 in section 6.5 indicates that the difference in their ICAP performance may be due to differences in their spatial representations of variance. As the FD metric uses a timeseries which is constructed using a spatial average, this spatial difference may be lost in FD as opposed to PCAP and ICAP. M2 has a similarly large variance for frequency 2, but does not capture the remaining frequencies at the same variances as M1. This difference may account for it performing worse than M1.

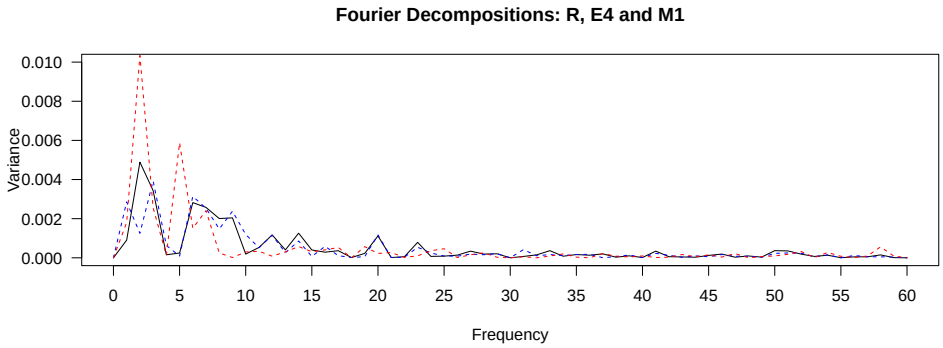

Figure 6.16: The Fourier decomposition of the GHT timeseries from R (*black*), E4 (*dashed blue*) and M1 (*dashed red*) datasets. Only the first 61 (0-60) frequencies are shown as the remaining frequencies have negligible variance.

In figure 6.17, E4 out performs most of the other datasets using TAS data. The exception to this is M1, which like the FD performance using GHT data, is due to the overestimation of some frequencies by M1 (not shown). In terms of variable sensitivity for FD, some the datasets change their rank. For example, C1 has a better rank using TAS than GHT, while B1 has a worse rank. M1 and M2 remain the same.

To easily compare the rank differences between the metrics and variables, table 6.2 combines the ranks from FD, PCAP, and ICAP when using both variables. This table provides a simple method for assessing the sensitivity of the metrics to

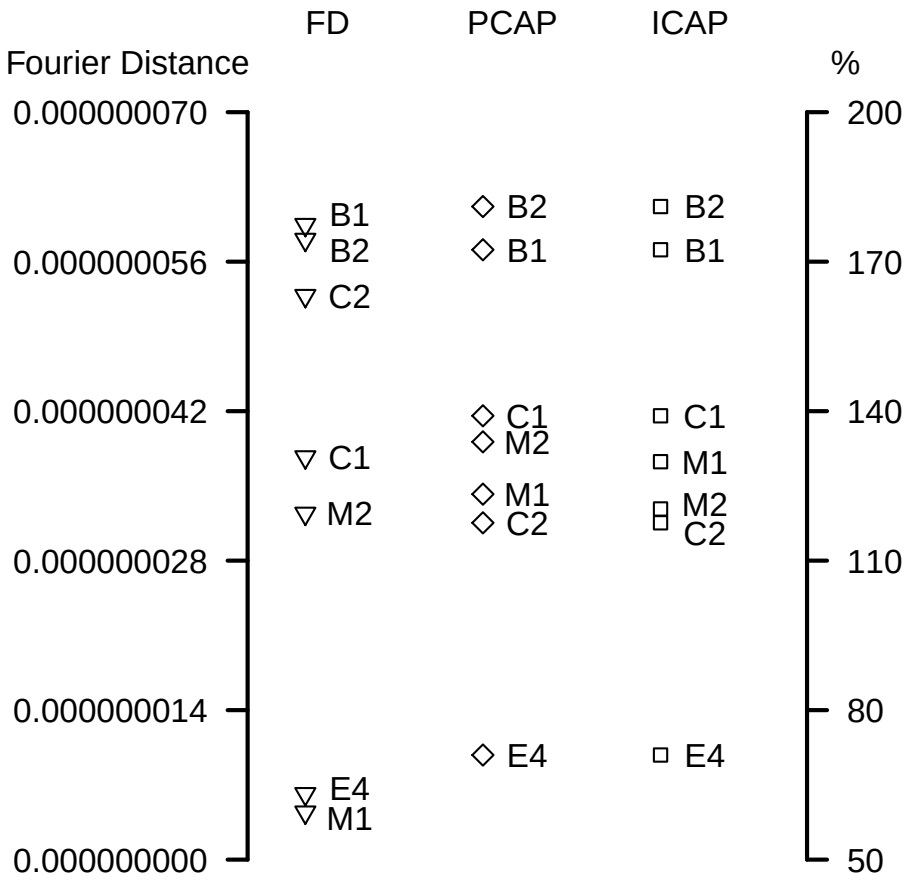

Figure 6.17: As figure 6.15 but for TAS.

a change in the variable used. This is accomplished by counting how many times the same dataset is found at the same rank for both variables and is termed: *rank sensitivity*. For example with FD, B2 is at rank 6 for both variables, while at rank 7 there are different datasets for the variables (C1 for GHT and B1 for TAS). In total, the rank sensitivities are: 4 for FD, 3 for PCAP, and 5 for ICAP. Therefore in terms of maintaining the same order of the datasets between GHT and TAS, ICAP is the least sensitive.

	GHT	TAS	GHT	TAS	GHT	TAS
Rank	**FD**	**FD**	**PCAP**	**PCAP**	**ICAP**	**ICAP**
7	C1	B1	B2	B2	B2	B2
6	B2	B2	M2	B1	B1	B1
5	B1	C2	B1	C1	C1	C1
4	C2	C1	M1	M2	C2	M1
3	M2	M2	C1	M1	M2	M2
2	E4	E4	C2	C2	M1	C2
1	M1	M1	E4	E4	E4	E4

Table 6.2: The ranks of the datasets for FD, PCAP and ICAP when measuring performance using GHT and TAS.

6.7.2.2 Bias

The second metric to rank datasets is known as (*bias*) and it measures the mean state of the climate as opposed to variance. It is not a novel metric, and has been used in other works such as Suppiah et al. (2007) and in those discussed in section 2.3. The bias of a gridded dataset is calculated by finding the average for each grid cell over the period. Following which, the total absolute difference between the mean of the dataset and the mean of the reference dataset serves as the measure of performance.

The bias metric differs to FD (section 6.7.2.1), PCAP and ICAP in that the data used cannot be preprocessed in the same manner as those in section 5.3 which removes the mean in step 3 (as part of trend removal). Rather, only the subset data from step 1 is used. While bias uses different data and measures a different aspect of the climate, it is much more common than FD, PCAP and ICAP. Therefore, it serves as as more recognisable metric to compare against the rankings of FD, PCAP, and ICAP.

The bias for each dataset is calculated for the GHT (section 5.2) and TAS (section 6.7.1) data. G and MM datasets are not used in this section, though MM would likely perform very well (Randall et al., 2007). The biases for GHT and TAS are shown in figure 6.18.

In figure 6.18, E4 performs the best for both datasets as it is constrained by

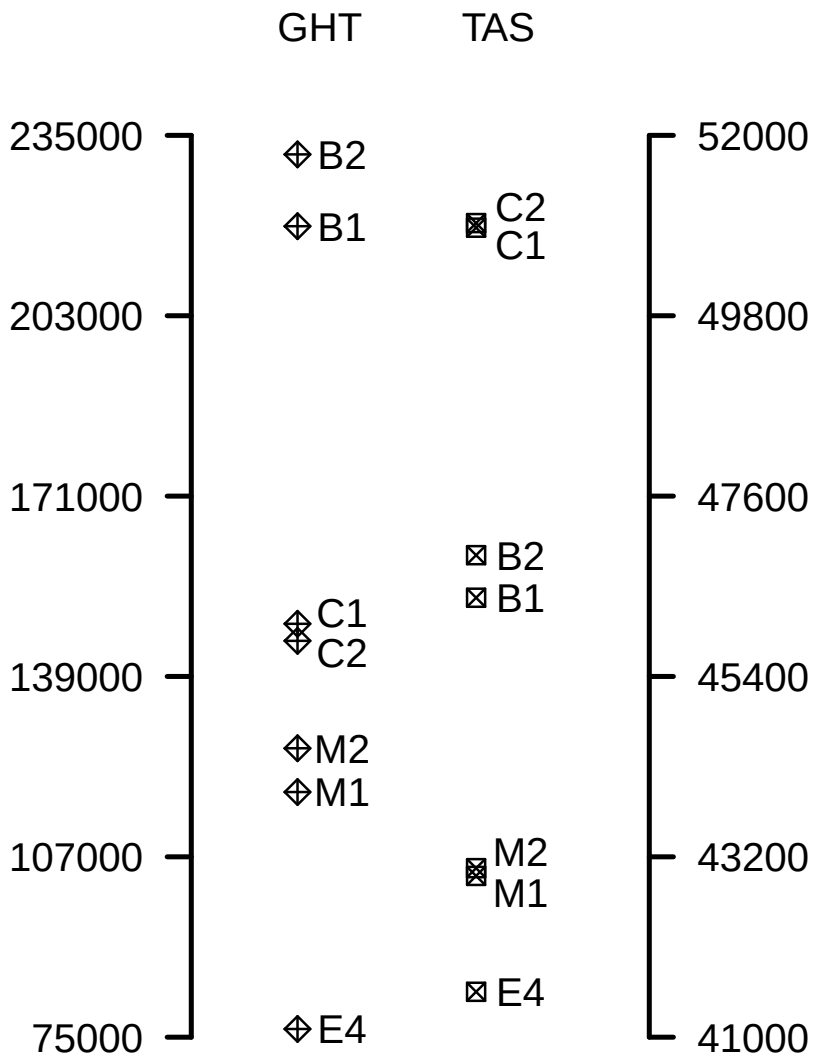

Figure 6.18: The biases for the datasets, GHT with y-axis (*left)* and TAS with y-axis (*right*). The GHT bias range is much greater than that of TAS.

observations. In terms of member performance, each member visually clusters closer to members from the same model, than to members from other models. The main difference between GHT and TAS for bias, is that members C1 and C2 have a better rank (lower) for GHT than for TAS. Figure 6.19 indicates the geographical differences in mean for GHT and TAS. For example, C1 underestimates the mean over the Antarctic region for GHT while generally overestimating it in TAS. As C2 has a similar bias to C1, it is not investigated.

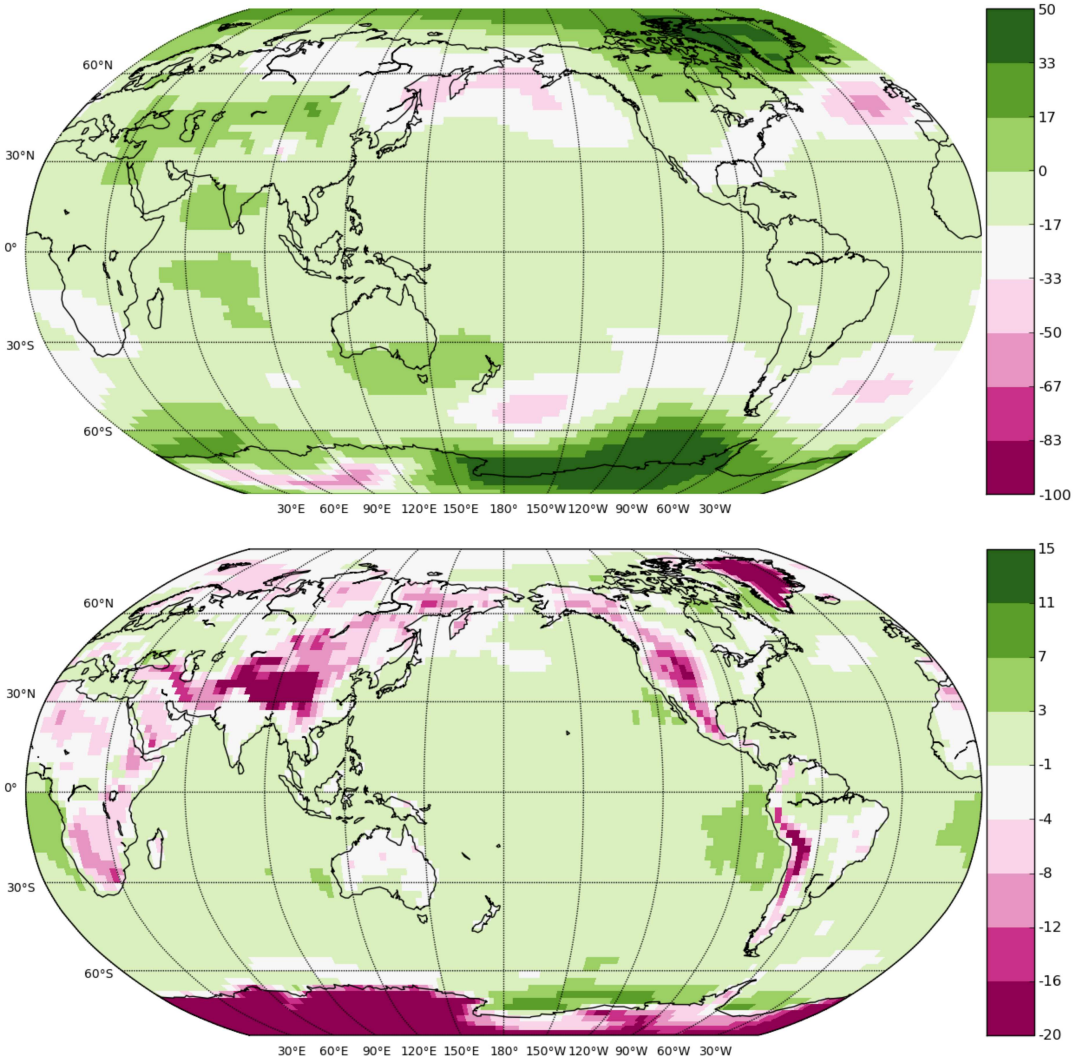

Figure 6.19: The biases for C1 with respect to the reference datasets for GHT (*top*) and TAS (*bottom*). The plots have different ranges to better highlight regional differences.

The same approach to measuring the sensitivity of datasets to a change in variable used for FD, PCAP, and ICAP (section 6.7.2.1) is also used for bias. The rank sensitivity of bias is therefore 3, which is mainly due to the change in performance for C1 and C2 between variables. This makes the bias metric as sensitive to a change in rank as PCAP, and more sensitive than ICAP with 5.

6.8 Discussion

ICA was used to construct a novel performance metric ($ICAP$). Signals are separated from reanalysis data and the degree to which member datasets manifest the same signals is used as the metric. The signals are automatically assumed to represent variations in the atmosphere which are related to modes of variability. The specific limitations to ICA are discussed in section 4.5, general limitations to performance metrics in section 2.4.1, while specific limitations with regards to this work are discussed in section 5.8.

As ICAP is dependent upon the results from PCA (section 5.6), a performance metric was constructed using PCA as well ($PCAP$). From the artificial example in section 5.6.1. it was known that the two metrics could potentially have different relative variances ($RPVE$) which could result in different rankings depending on whether PCAP or ICAP is used. However the fact that the two orderings do differ, shows that ranking the members by RPVE can produce different rankings despite the total variance between the two methods being preserved.

Sections 6.3 and 6.4 showed that the signals and their corresponding spatial maps from the reference dataset were unlike noise and artefacts. This is important, as it adds supporting evidence for the assumption in section 5.4 which stated that signals and static maps are automatically assumed to represent modes in this work. While the signals and their spatial maps were not associated to modes due to the complexities of the association process (section 4.4), their plausibility suggests that the automatic association to modes is at least to some extent valid. Further research will be needed to determine the full validity of this assumption, and whether it holds for PCAP as well.

Section 6.6 examined the sensitivity of PCAP and ICAP when the set size and geopotential height level were changed. Additional ensemble members were also used to ensure that the performance rankings were less dependent upon the number of datasets used. The performance of the datasets showed that PCAP and ICAP could generally identify the alternative reanalysis dataset (E4) from the other members, as well as showing the noise dataset performed poorly. As E4 and the noise dataset consistently had good and bad performances respectively,

this shows that PCAP and ICAP are to some degree consistent in their rankings of datasets. This does not however make these metrics completely insensitive to changes in set sizes and levels, as the performance of some datasets were shown to be impacted by the changes.

To gauge the extent to which dataset performance is impacted by a change in the variable measured, an additional experiment was conducted. In this experiment, PV and signals were separated from near-surface temperature data (section 6.7) and compared to the geopotential height results (section 6.5). The sensitivity to the change was measured by counting how many times each dataset kept its rank between different variables for a given metric. This was to provide a simple measure of sensitivity, and it was termed: rank sensitivity.

However, determining the rank sensitivity of PCAP and ICAP in isolation does not provide information about how much better or worse their rank sensitivity is, compared to other metrics. To address this concern, two other metrics were introduced. The Fourier Distance (FD) measures variance like PCAP and ICAP, while bias uses the mean state of the climate. Both these metrics rank models with respect to the reference dataset. In ranking models relative to a reference dataset, they are similar to PCAP and ICAP. FD is a novel metric used solely in this work. It uses similar data to PCAP and ICAP but it is globally averaged and is not dimensionally reduced. FD also ranks data using the variance of the datasets like PCAP and ICAP. Bias on the other hand is a common metric in the existing literature and it serves as a benchmark of rank sensitivity in this work. The results from sections 6.7.2.1 and 6.7.2.2 show that PCAP and ICAP are at worst as sensitive as FD and bias, with ICAP being slightly less rank sensitive than PCAP. However different variables or datasets may result in different rank sensitivities.

The limitations and complexities with PCAP and ICAP have been discussed in section 4.5 and 5.8. Of particular note, is that like other metrics, the two metrics in this work are still sensitive to changes, such as changes in the variable, geopotential height level, time period, and the number of PV and signals used. What the results from this chapter do indicate, is potential. ICAP produced plausible static maps and both metrics were generally consistent in their ranking

of datasets. Their ranking of datasets was also only as sensitive as an existing metric (bias).

Chapter 7

Conclusions

7.1 Overview

This dissertation is concerned with the design and demonstration of a novel metric for evaluating climate models, and the primary contribution is a methodology which develops an ICA model performance metric (*ICAP*). The metric is used to determine the performance of model results by determining how well models have simulated global patterns found in present day climate data (reanalysis data). As assessing model performance necessitates the consideration of multiple aspects of model performance, the metric is designed to be used in conjunction with other evaluation methods.

As the metric targets the fundamental modes of the climate system, in multi-model contexts it provides a basis for choosing to remove or weight models that poorly simulate the present day climate. The rational is that models which poorly simulate the present day are considered less likely to simulate the future climate well and should therefore be discounted in further analysis.

Patterns that may represent modes of climate variability (*modes*) are found in reanalysis data using a novel application of Independent Component Analysis (*ICA*). The specific patterns found using ICA (*signals*) are identified from reanalysis data and therefore represent the observationally constrained fundamen-

tal modes. These modes are then assessed in the model data to inform on the evaluation of whether the model results contain credible information, and for determining if the models are justifiably producing correct results for the correct reason. ICA is selected as a new technique to find representations of modes in data, as it presents potential advantages in identifying the modes and associated patterns in noisy data. The importance of understanding modes and the complexities in associating patterns to modes, are further discussed in section 7.2.

Part of the performance metric application employs preprocessing of data using Principal Component Analysis (PCA). This inclusion of PCA creates a constraint that makes the raw results from PCA and ICA appear identical in terms of their total variance. However the total variance is not indicative of the signals, and so to differentiate between the results from the two methods, a novel measure is developed, termed the relative percentage of variance explained ($RPVE$). The performance metric calculates the total RPVE of the reanalysis signals found in a model result. Due to the dependence of ICA on PCA in this work, a similar performance metric based on PCA results is also developed ($PCAP$) to investigate the dependency.

There are many limitations with PCAP and ICAP (sections 2.4.1, 4.5, and 5.8). These include the potential sensitivity of performance to the variable used, the limited number of datasets generally used in this work, and how to determine the number of PV to extract from the data. To assist in determining how many PV should be extracted, the rule N method was applied to determine which ones were above the level of noise (section 4.3). To help ensure that the signals separated were consistent, an average of the unmixing matrix was demonstrated in section 5.4.2.

As model performance been shown to be dependent on the variable used (section 2.4.1), PCAP and ICAP were also assessed when first using different geopotential height levels (700mb and 500mb), additional members and two different set sizes (section 6.6). Both metrics were able to consistently identify the alternative reanalysis dataset from almost all the remaining members at both levels and when using both set sizes. Secondly, the metrics were investigated when changing the

variable used (geopotential height data at 700mb and near surface temperature data). Potential changes in the order (ranking) of the datasets was measured to determine the sensitivity of dataset performances to a change in the variable used. Compared to the rank sensitivities of two other metrics (Fourier Distance and bias), PCAP and ICAP has similar sensitivities. ICAP was found to be the least sensitive of the three metrics, but only by a small amount (section 6.7.2). The patterns found using the ICA-based approach were also shown to be plausible representations of the modes, with a full discussion of pattern association in section 7.2.

The ICAP metric addresses the following:

1. **Model Performance Metric**
   The ICA based model performance metric is developed to rank multi-model ensemble members according to how well they capture the representations of global modes found in reanalysis data. Figure 6.7 is represented below in figure 7.1, and it demonstrates how the metric was able to rank the members used in this work.

   Of particular importance is that the ordering of the members differs depending on which metric is used. For example, member M1 performed best by ICAP (lowest value) while member C2 performed best by PCAP. This shows that despite the dependency of the ICA approach on using PCA for preprocessing, the ranking of the members using ICAP compared to when using PCAP brings additional power to discriminate between signals in the simulations. This may be due to different ways the techniques function, maximising variance (PCA) and maximising independence (ICA).

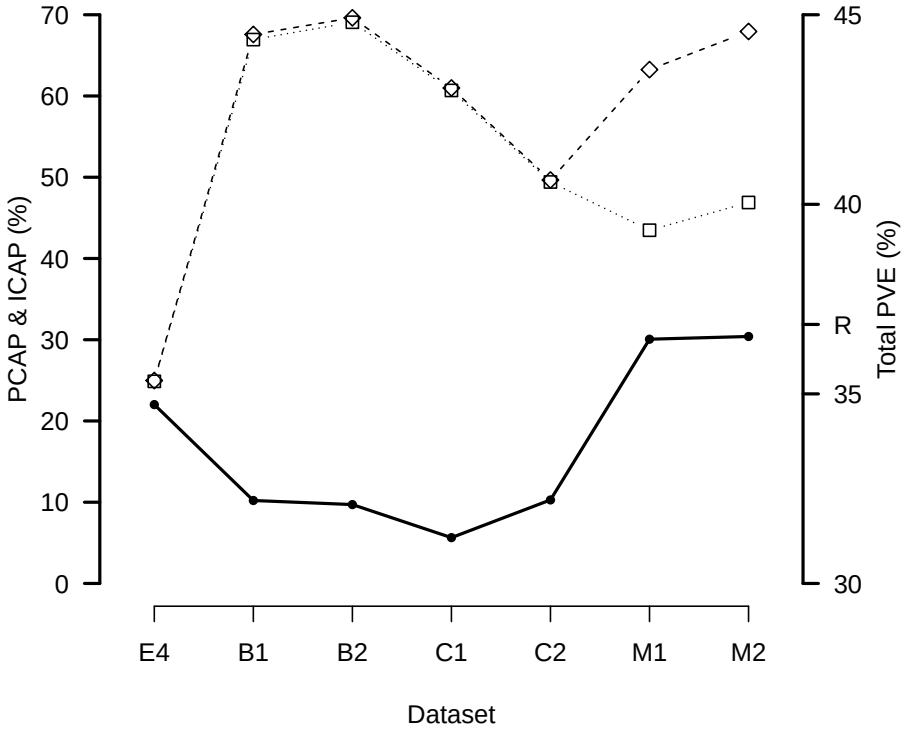

Figure 7.1: PCAP (diamonds) and ICAP (squares) results. The PVE are indicated by the solid points. Dataset order is arbitrary and the Noise and MM datasets are not shown due to their poor performance. "R" Indicates the total PVE of the reference dataset (37%), while the list of the full dataset names can be found in table 5.1 (See section 6.5 for more details).

2. **ICA Application Extended to Multiple Datasets**
   The ICA performance metric presented in this work removed the need to subjectively associate signals to modes. The consequence of this is that the methodology is applicable to multiple datasets without having the overhead of associating patterns to modes.

   To demonstrate the application of the performance metric, a total of ten datasets were used in section 5.2, and more than 20 in section 6.6. Therefore the performance metric allows for the novel extension of ICA to multiple datasets.

3. **Differentiate Between Noise and Climate**
   To show that the model performance metric can differentiate between re-

sults from the true climate and those from noise, two test datasets were used: an alternative reanalysis which should perform well, and a dataset containing only noise which should perform poorly. The results were as expected, with the alternative reanalysis dataset performing well, and the dataset containing only noise performing poorly (see table 6.2). Therefore the contribution of this dissertation is a metric which can differentiate between noise and signals within simulations of the climate.

7.2 Case for Understanding Modes of Climate Variability

Understanding the behaviour of global modes is a necessary part of understanding the overall climate, as they describe the temporal and spatial distribution of variables, such a temperature and rainfall. As modes can be global in their spatial influence, they can also be responsible for affecting regional climates. An example of this is presented by Hart et al. (2013) who show that the Madden-Julian Oscillation (a mode) has a weak but significant affect on South African rainfall through its modulation of tropical temperate troughs. Smith and Chandler (2010) also demonstrate the influence of a global mode on a regional climate when they account for the El Niño Southern Oscillation amongst other factors in their assessment of global models to capture Australian rainfall.

The importance of understanding modes is also shown by Leary et al. (2009), who are interested in deriving relevant information for users from climate data. They state that before the data can be interpreted as information it should portray a realistic climate and there should be a "*clear process-based understanding of the response of the physical and social systems to climate and other pressures*" (see section 1.2). In terms of modelling importance, Tebaldi and Knutti (2007) state that models should obtain the right result (appropriate variability on the range of spatial and temporal scales) for the right reason, such as through the proper simulation of modes rather than by factors such as tuning (section 2.4.2).

To understand the behaviour of modes within model simulations, potential representations of modes have to be found in the data and then successfully associated to modes. There are a number of different approaches to accomplish this task which are reviewed in chapter 3. However, the association process contains many complexities (section 4.4). For example, one pattern could represent multiple modes or how to associate patterns found using one technique to a climate index (indicative of a mode) which is produced using another technique that makes fundamentally different assumptions about the data.

This dissertation assumes that representations found within reanalysis data automatically represent valid modes (section 5.5). This assumption removes the task of associating potential representations to modes, which is currently performed manually by experts. Expert association may be timely and subjective and therefore increasingly infeasible for application to many datasets. This work finds that the representations from the reanalysis dataset are plausible, as their spatial manifestations may represent modes and they did not represent artefacts or noise (section 6.4). Additional work would be needed to determine the full validity of this assumption.

The contribution of this dissertation to the existing approaches is the ability of the ICA based metric to differentiate between patterns inherent in records of the climate (reanalysis data) from those seen in noise (Gaussian). This can be seen in table 7.1 (table 6.1 represented), where the alternative reanalysis dataset (E4) performed better (smaller value) than the noise dataset (G) when using ICAP to rank the datasets with respect to another reanalysis dataset (R). The table also shows the difference between the ICA and PCA based metrics, namely that ordering of members can change. For example, M1 and M2 change their values the most between ICAP and PCAP compared to the other datasets used in this work.

The plausibility, consistency, and sensitivity of ICAP demonstrates that it can offer a novel contribution to the existing tools available for understanding the behaviour of modes, and how they are represented by models.

Dataset	Total PVE (%)	ICAP (%)	PCAP (%)
R	36.83	-	-
E4	34.72	24.88	24.98
B1	32.19	66.95	67.57
B2	32.08	69.08	69.62
C1	31.21	60.68	60.99
C2	32.20	49.41	49.66
M1	36.44	**43.48**	**63.25**
M2	36.51	**46.89**	**67.94**
MM	7.59	331.47	331.56
G	0.04	399.28	399.28

Table 7.1: The total PVE, ICAP, and PCAP results. M1 and M2 (emboldened) differ the most when changing between ICAP and PCAP. The G and MM datasets results are included. Order of results is the same as in figure 6.7 (see section 6.5 for more details)

7.3 Developing a Context for Application

7.3.1 Meta-Metric

This dissertation demonstrates a novel performance metric for evaluating climate models according to how well they have simulated representations of modes of climate variability found within reanalysis data using Independent Component Analysis. However, the current applications of similar techniques to climate data involves expert interpretation to determine if the patterns found by the techniques are indeed representative of modes (section 4.4). The subsequent success of the application and the value of the technique is therefore dependant upon the quality of the association.

There are two problems in particular that require further investigation for this method of discounting models. Firstly, by removing the need for expert assessment, the traditional means for justifying the selection of the technique is also removed. This makes all techniques valid to some degree for finding representations in reanalysis data. So multiple techniques will be required to provided a more justifiable set of model result rankings rather than using a single one.

Secondly, the volume of data from climate models to analyse is set to grow in the future (Overpeck et al., 2011) and could be further increased if aspects such as sub-regions are also considered for analysis. Applying a variety of techniques to large volumes of data will further compound the situation by potentially creating a large growth in the amount of metric results to interpret.

One solution to these problems may be to create a meta-metric, a metric which summarises the overall performance of multiple model results by multiple techniques. This would extend the concept by Gleckler et al. (2008) who created the Model Climate Performance Index. The metric is the average root mean square error over a number of variables (fields) and spatial domains. It allows for the overall assessment of multiple model results, offering the trade of added simplicity for a loss in error information (e.g.: errors over different spatial domain are averaged).

The meta-metric would address part of the wider set of recommendations by Knutti (2010) by functioning as a broad brush metric which could be used to identify poorly performing model results. The poorly performing results could then be analysed further using other evaluation methods to determine if they should be discounted or retained. By focusing only on poorly performing model results, the addition analysis would be isolated to only those model results. This would reduce the total amount of analysis required. How poor is poor enough to discard a result remains an open question, and the meta-metric would require the creation of additional performance metrics.

Given the foreseen growth in the volume of climate data, a meta-metric followed by a more detailed evaluation of poor performing results, may be one of the ways to help determine if climate data can serve as credible information for users.

7.3.2 Model Discounting Framework

Discounting model results is one of the methods reviewed for having potential to reduce the spread in model results (see section 2.3.3). As there is no one best method or set of methods for evaluating models, selecting a method is ultimately

subjective. This section introduces a new approach to assist experts in the task of discounting model results.

The approach is based the concept of an issue tree which is taken from the field of management consultancy (e.g.: Cheng (2012)). The issue tree is a tool used by consultants to solve a specific problem. To solve the problem, a consultant creates a hypothesis and seeks to validate it against the available data. Extending this approach to discounting model results, the expert would decide if a result should be discounted based on the condition the model result has failed. An example of an issue tree is presented in figure 7.2.

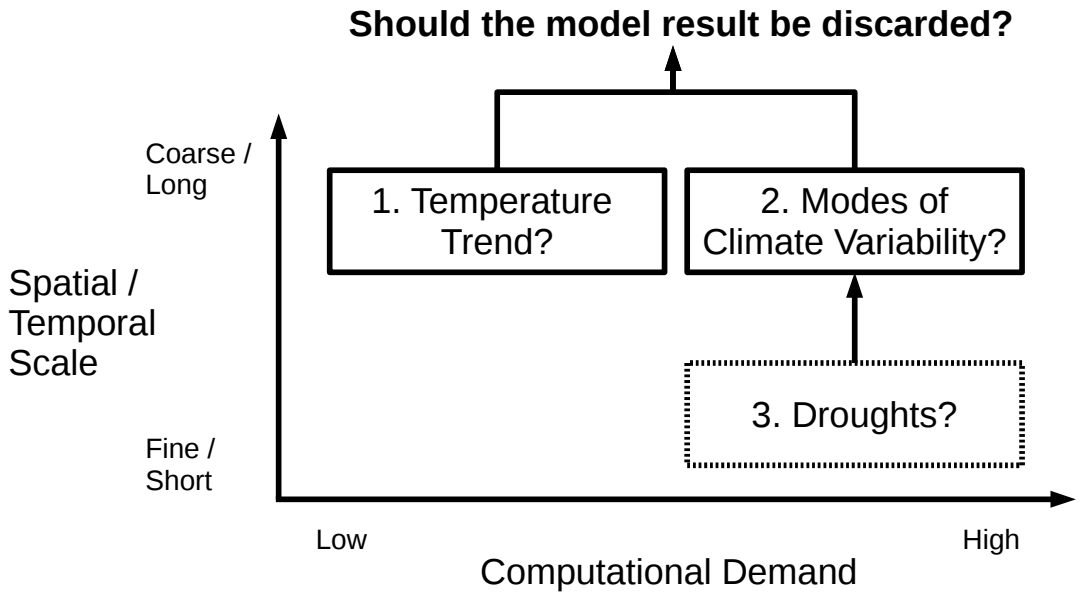

Figure 7.2: An example of an issue tree for discounting a model result. Conditions 1 and 2 are the Model Discounting Framework, while condition 3 customises the framework to become an issue tree for a specific case. Numbers indicate the order in which the conditions should be used to evaluate a model result.

In the figure, the first condition to pass could be determining if the model result had correctly simulated the air surface temperature trend compared to observational data. If the model result passed this condition then the next condition could be determining if the model had correctly simulated modes of climate variability. If a result fails a condition, then it is discounted and no further conditions

are required to be checked. The advantage of this framework is that it doesn't have to comprehensively include every evaluation method as a condition: Merely failing one condition in the issue tree is sufficient to discount a model result.

The tree is constructed to have two axes: computational demand and spatial / temporal scale. This design enables an efficient fail-fast approach to discounting a model result. Firstly, conditions such as testing the sign of the temperature trend are relatively computationally inexpensive, and therefore a model result should be tested against this type of condition first. More computationally demanding conditions (such as ICAP) should only be used to test a model result if less computationally demanding conditions did not discount the result.

Secondly, if a model result fails to capture the modes of climate variability on a global scale, then it is unlikely that any skill in simulating more regional patterns such as rainfall, can be justified. This may be the case in regional models which are forced by global models. Therefore conditions which test global scale over long periods of time should be applied to data first before finer scale conditions are applied to the result. The issue tree is therefore designed to optimise the use of existing methods for discounting model results.

The recommendation is to take the concept of the issue tree presented in this work and create a more general model discounting framework (MDF). This would provide a standard set of conditions for evaluating a model result against. In doing so, it would ensure a minimum level of performance for each model result that is not discounted. At the same time, the framework could be taken by individual experts and extended to meet the discounting needs of their specific work. For example by further testing a result against drought frequency.

The concept presented in this section is a very general example of what the framework may be able to offer experts. Adapting the framework for global or regional model results would be needed, as well as quantifying what is poor performance. The construction of the framework will also have practical constraints such how a model result is preprocessed. For example, if the trend is not removed from the model result between testing the first two conditions in figure 7.2, then the condition for testing the modes of climate variability may also be partly affected by

trends. Therefore ensuring that the right conditions (e.g.: 1 and 2) are mutually independent of each other will also be needed.

References

F. Aires, A. Chédin, and J.P. Nadal. Independent Component Analysis of Multivariate Time Series: Application to the Tropical SST Variability. *Journal of Geophysical Research*, 105(D13):17,437–17,437, 2000.

F. Aires, W.B. Rossow, and A. Chédin. Rotation of EOFs by the Independent Component Analysis: Toward a Solution of the Mixing Problem in the Decomposition of Geophysical Time Series. *Journal of the Atmospheric Sciences*, 59 (1):111–123, 2002.

J.D. Annan and J.C. Hargreaves. Reliability of the CMIP3 Ensemble. *Geophysical research letters*, 37(2):L02703, 2010. doi: 10.1029/2009GL041994.

M.P. Baldwin, D.B. Stephenson, and I.T. Jolliffe. Spatial Weighting and Iterative Projection Methods for EOFs. *Journal of Climate*, 22(2):234–243, 2009.

A.G. Barnston and R.E. Livezey. Classification, Seasonality and Persistence of Low-Frequency Atmospheric Circulation Patterns. *Monthly Weather Review*, 115(6):1083–1126, 1987.

J. Basak, A. Sudarshan, D. Trivedi, and M.S. Santhanam. Weather Data Mining Using Independent Component Analysis. *Journal of Machine Learning Research*, 5:239–253, 2004.

W.J Burroughs. *Weather Cycles: Real or Imaginary?* Cambridge University Press Cambridge, 2003.

J.F. Cardoso. Blind Signal Separation: Statistical Principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

D.W. Cash, W.C. Clark, F. Alcock, N.M. Dickson, N. Eckley, D.H. Guston, J. Jäger, and R.B. Mitchell. Knowledge Systems for Sustainable Development. *Proceedings of the National Academy of Sciences*, 100(14):8086–8091, 2003.

R.E. Chandler and B. Bates. Exploiting Strength, Discounting Weakness: Combining Information from Multiple Climate Simulators. *Statistical Science*, (311):1–25, 2011.

V. Cheng. *Case Interview Secrets: A Former McKinsey Interviewer Reveals How to Get Multiple Job Offers in Consulting*. Innovation Press, 2012.

E.C. Cherry. Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

B. Christiansen. Atmospheric Circulation Regimes: Can Cluster Analysis Provide the Number? *Journal of Climate*, 20:2229–2250, 2007.

B. Christiansen. Is the Atmosphere Interesting? A Projection Pursuit Study of the Circulation in the Northern Hemisphere Winter. *Journal of Climate*, 22 (5), 2009.

R.H. Compagnucci and M.B. Richman. Can Principal Component Analysis Provide Atmospheric Circulation or Teleconnection Patterns? *International Journal of Climatology*, 28(6):703–726, 2008.

P.N. Edwards. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Mit Press, 2010.

P.N. Edwards. History of Climate Modeling. *Wiley Interdisciplinary Reviews: Climate Change*, 2(1):128–139, 2011.

B. Enserink, J.H. Kwakkel, and S. Veenman. Coping with Uncertainty in Climate Policy Making: (Mis)Understanding Scenario Studies. *Futures*, 53:1–12, 2013. doi: 10.1016/j.futures.2013.09.006.

L. Ertöz, M. Steinbach, and V. Kumar. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. pages 47–58. SIAM, 2003.

G. Flato, J. Marotzke, B. Abiodun, P. Braconnot, S.C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen. IPCC, 2013: Annex III: Glossary [Planton, S. (ed.)]. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.*, pages 1447–1465, 2013a.

G. Flato, J. Marotzke, B. Abiodun, P. Braconnot, S.C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen. Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.*, 2013b.

I.K. Fodor and C. Kamath. Using Independent Component Analysis to Separate Signals in Climate Data. In *the Climate Data, Proceedings of the SPIE*, volume 5102, pages 25–36, 2003.

J.H. Friedman and J.W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *Computers, IEEE Transactions on*, 100(9):881–890, 1974.

P.R. Gent, F.O. Bryan, G. Danabasoglu, S.C. Doney, W.R. Holland, W.G. Large, and J.C. McWilliams. The NCAR Climate System Model Global Ocean Component. *Journal of Climate*, 11(6):1287–1306, 1998.

M.A. Giorgetta, J. Jungclaus, C. Reick, B. Stevens, J. Marotzke, M. Claussen, E. Roeckner, T. Mauritsen, T. Crueger, H. Schmidt, et al. Climate Variability and Climate Change in MPI-ESM CMIP5 Simulations. *J. Adv. Model. Earth Syst., in revision*, 2012.

F. Giorgi and E. Coppola. Does the Model Regional Bias Affect the Projected Regional Climate Change? An Analysis of Global Model Projections. *Climatic change*, 100(3):787–795, 2010.

F. Giorgi and L.O. Mearns. Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging (REA) Method". *Journal of Climate*, 15(10): 1141–1158, 2002.

P.J. Gleckler, K.E. Taylor, and C. Doutriaux. Performance Metrics for Climate Models. *J. Geophys. Res*, 113:D06104, 2008.

A. Hannachi, S. Unkel, N.T. Trendafilov, and I.T. Jolliffe. Independent Component Analysis of Climate Data: A New Look at EOF Rotation. *Journal of Climate*, 22(11):2797–2812, 2009. doi: 10.1175/2008JCLI2571.1.

N.C.G Hart, C.J.C Reason, and N. Fauchereau. Cloud Bands over Southern Africa: Seasonality, Contribution to Rainfall Variability and Modulation by the MJO. *Climate Dynamics*, 41(5-6):1199–1212, 2013.

B.C. Hewitson and R.G. Crane. Self-organizing Maps: Applications to Synoptic Climatology. *Climate Research*, 22(1):13–26, 2002.

A. Hyvärinen. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, 1999a.

A. Hyvärinen. Survey on Independent Component Analysis. *Neural computing surveys*, 2(4):94–128, 1999b.

A. Hyvärinen and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural networks*, 13(4):411–430, 2000.

A. Hyvärinen, Jaakko Särelw, and Ricardo Vigário. Spikes and Bumps: Artefacts Generated by Independent Component Analysis with Insufficient Sample Size. In *First International Workshop on Independent Component Analysis and Signal Separation*, 1999.

A. Ilin, H. Valpola, and E. Oja. Semiblind Source Separation of Climate Data Detects El Niño as the Component with the Highest Interannual Variability. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, volume 3, pages 1722–1727. IEEE, 2005.

A. Ilin, H. Valpola, and E. Oja. Exploratory Analysis of Climate Data Using Source Separation Methods. *Neural Networks*, 19(2):155–167, 2006.

H. Itoh, A. Mori, and S. Yukimoto. Independent Components in the Northern Hemisphere Winter: Is the Arctic Oscillation Independent? 気象集誌 *(Journal of the Meteorological Society of Japan)*, 85(6):825–846, 2007.

R.A. Jarvis and E.A. Patrick. Clustering Using a Similarity Measure Based on Shared near Neighbors. *IEEE Transactions on Computers*, 22(11):1025–1034, 1973.

I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

P.D. Jones, K.E. Trenberth, P.G. Ambenje, R. Bojariu, D.R. Easterling, A.M.G. Klein, D.E. Tank, J.A. Renwick, F. Rahimzadeh, M.M. Rusticucci, et al. Observations: Surface and Atmospheric Climate Change. *IPCC, Climate change*, pages 235–336, 2007.

M.W. Jury, A.F. Prein, H. Truhetz, and A. Gobiet. Evaluation of CMIP5 Models in the Context of Dynamical Downscaling over Europe. *Journal of Climate*, 28 (14):5575–5582, Mar 2015. ISSN 0894-8755. doi: 10.1175/JCLI-D-14-00430.1. URL http://dx.doi.org/10.1175/JCLI-D-14-00430.1.

E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, H. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J Janowiak, K.C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Jospeh. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, 77:437–471, 1996.

M. Kent. The Value of Independent Component Analysis in Identifying Climate Processes. Master's thesis, University of Cape Town, 2011.

D.G.C. Kirono and D.M. Kent. Assessment of Rainfall and Potential Evaporation from Global Climate Models and Its Implications for Australian Regional Drought Projection. *International Journal of Climatology*, 31(9):1295–1308, 2011.

R. Knutti. Should We Believe Model Predictions of Future Climate Change? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1885):4647–4664, 2008. doi: 10.1098/rsta.2008.0169.

R. Knutti. The End of Model Democracy? *Climatic Change*, 102(3-4):395–404, jan 2010. ISSN 0165-0009. doi: 10.1007/s10584-010-9800-2. URL http://www.springerlink.com/index/10.1007/s10584-010-9800-2.

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

I. Koch and K. Naito. Dimension Selection for Feature Selection and Dimension Reduction with Principal and Independent Component Analysis. *Neural computation*, 19(2):513–545, 2007.

K.B. Korb and A.E. Nicholson. *Bayesian Artificial Intelligence*. CRC Press, 2004.

R.H. Langland, R.N. Maue, and C.H. Bishop. Uncertainty in Atmospheric Temperature Analyses. *Tellus A*, 60(4):598–603, 2008. ISSN 1600-0870. doi: 10.1111/j.1600-0870.2008.00336.x. URL http://dx.doi.org/10.1111/j.1600-0870.2008.00336.x.

N. Leary, K. Averyt, B. Hewitson, and J. Marengo. Crossing Thresholds in Regional Climate Research: Synthesis of the IPCC Expert Meeting on Regional Impacts, Adaptation, Vulnerability, and Mitigation. *Climate Research*, 40:121–131, 2009. doi: 10.3354/cr00832.

Y. Liu, R.H. Weisberg, and C.N.K. Mooers. Performance Evaluation of the Self-Organizing Map for Feature Extraction. *Journal of geophysical Research*, 111 (C5):C05018, 2006.

A. Lotsch, M. A. Friedl, and J. Pinzón. Spatio-Temporal Deconvolution of NDVI Image Sequences Using Independent Component Analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 41(12):2938–2942, 2003.

R.A. Madden and P.R. Julian. Detection of a 40-50 Day Oscillation in the Zonal Wind in the Tropical Pacific. *Journal of the Atmospheric Sciences*, 28(5):702–708, 1971.

V. Masson-Delmotte, M. Schulz, A. Abe-Ouchi, J. Beer, A. Ganopolski, J.F. González Rouco, E. Jansen, K. Lambeck, J. Luterbacher, T. Naish, T. Osborn, B. Otto-Bliesner, T. Quinn, R. Ramesh, M. Rojas, X. Shao, and A. Timmermann. 2013: Information from Paleoclimate Archives. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker,T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. *Climate change*, pages 383–464, 2013.

Met Office Hadley Centre. WCRP CMIP5: Met Office Hadley Centre (MOHC) HadCM3 model output for the decadal1960 experiment. Centre for Environmental Data Analysis. 2017. URL http://catalogue.ceda.ac.uk/uuid/270018c8110449a1a6186333737e86a8.

A. Mori, N. Kawasaki, K. Yamazaki, M. Honda, and H. Nakamura. A Reexamination of the Northern Hemisphere Sea Level Pressure Variability by the Independent Component Analysis. *SOLA*, 2(0):5–8, 2006.

J.M. Murphy, B.B.B. Booth, M. Collins, G.R. Harris, D.M.H. Sexton, and M.J. Webb. 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, 2007.

J.P. Nadal, E. Korutcheva, and F. Aires. Blind Source Separation in the Presence of Weak Source. *Arxiv preprint cond-mat/0005258*, 2000.

E. Ollila. The Deflation-Based FastICA Estimator: Statistical Analysis Revisited. *Signal Processing, IEEE Transactions on*, 58(3):1527–1541, 2010.

J.E. Overland and R.W. Preisendorfer. A Significance Test for Principal Components Applied to a Cyclone Climatology. *Monthly Weather Review*, 110:1, 1982.

J.T. Overpeck, G.A. Meehl, S. Bony, and D.R. Easterling. Climate Data Challenges in the 21st Century. *Science(Washington)*, 331(6018):700–702, 2011.

M.S. Pedersen, J. Larsen, U. Kjems, and L.C. Parra. A Survey of Convolutive Blind Source Separation Methods. *Multichannel Speech Processing Handbook*, 2007.

D.W. Pierce, T.P. Barnett, B.D. Santer, and P.J. Gleckler. Selecting Global Climate Models for Regional Climate Change Studies. *Proceedings of the National Academy of Sciences*, 106(21):8441, 2009.

J. Räisänen. How Reliable Are Climate Models? *Tellus A*, 59(1):2–29, 2007. doi: 10.1111/j.1600-0870.2006.00211.x.

J. Räisänen and J.S. Ylhäisi. Can Model Weighting Improve Probabilistic Projections of Climate Change? *Climate Dynamics*, pages 1–18, 2011. doi: 10.1007/s00382-011-1217-8.

J. Räisänen, L. Ruokolainen, and J. Ylhäisi. Weighting of Model Results for Improving Best Estimates of Climate Change. *Climate dynamics*, 35(2):407–422, 2010.

D.A. Randall, R.A. Wood, S. Bony, R. Colman, T. Fichefet, J. Fyfe, V. Kattsov, A. Pitman, J. Shukla, J. Srinivasan, R.J. Stouffer, A. Sumi, and K.E. Taylor. Climate Models and Their Evaluation In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. 323:589–662, 2007.

D.R. Randall, M. Khairoutdinov, A. Arakawa, and W. Grabowski. Breaking the Cloud Parameterization Deadlock. *Bulletin of the American Meteorological Society*, 84(11):1547–1564, 2003.

D.B. Reusch, R.B. Alley, and B.C. Hewitson. North Atlantic Climate Variability from a Self-Organizing Map Perspective. *Atlantic*, 112:1–20, 2007. doi: 10.1029/2006JD007460.

M.B. Richman. Rotation of Principal Components. *Journal of climatology*, 6(3): 293–335, 1986.

B.M. Sanderson and R. Knutti. On the Interpretation of Constrained Climate Model Ensembles. *Geophysical Research Letters*, 39(16):L16708, 2012. doi: 10.1029/2012GL052665.

J. Särelä and H. Valpola. Denoising Source Separation. *Journal of Machine Learning Research*, 6:233–272, 2005.

M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, and J. Selbig. Metabolite Fingerprinting: Detecting Biological Features by Independent Component Analysis. *Bioinformatics*, 20(15):2447–2454, 2004.

J. Slingo and T. Palmer. Uncertainty in Weather and Climate Prediction. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 369(1956):4751–4767, 2011.

I. Smith and E. Chandler. Refining Rainfall Projections for the Murray Darling Basin of South-East Australia the Effect of Sampling Model Results Based on Performance. *Climatic Change*, 102(3):377–393, 2010.

M. Steinbach, P. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of Climate Indices Using Clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 446–455, New York, NY, USA, 2003. ACM. ISBN 1-58113-737-0. doi: 10.1145/956750.956801. URL http://doi.acm.org/10.1145/956750.956801.

K. Steinhaeuser, N.V. Chawla, and A.R. Ganguly. An Exploration of Climate Data Using Complex Networks. In *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*, SensorKDD '09, pages 23–31, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-668-7. doi: 10.1145/1601966.1601973. URL http://doi.acm.org/10.1145/1601966.1601973.

T. Stocker, Q. Dahe, G. Plattner, M. Tignor, and P. Midgley. IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections. *Atmospheric Research*, 2010.

T.F. Stocker, G.K.C. Clarke, H. Le Treut, R.S. Lindzen, V.P. Meleshko, R.K. Mugara, T.N. Palmer, R.T. Pierrehumbert, P.J. Sellers, K.E. Trenberth, and J. Willebrand. Physical Climate Processes and Feedbacks. Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.*, 2001.

J.V. Stone. *Independent Component Analysis: a Tutorial Introduction.* MIT press, 2004.

R. Suppiah, K.J. Hennessy, P.H. Whetton, K. McInnes, I. Macadam, J. Bathols, J. Ricketts, and C.M. Page. Australian Climate Change Projections Derived from Simulations Performed for the IPCC 4th Assessment Report. *Australian Meteorological Magazine*, 56(3):131–152, 2007.

P. Sura, M. Newman, C. Penland, and P. Sardeshmukh. Multiplicative Noise and Non-Gaussianity: A Paradigm for Atmospheric Regimes? *Journal of the Atmospheric Sciences*, 62(5):1391–1409, may 2005. ISSN 0022-4928. doi: 10.1175/JAS3408.1. URL http://journals.ametsoc.org/doi/abs/10.1175/JAS3408.1.

M.B. Sylla, F. Giorgi, E. Coppola, and L. Mariotti. Uncertainties in Daily Rainfall over Africa: Assessment of Gridded Observation Products and Evaluation of a Regional Climate Model Simulation. *International Journal of Climatology*, 2012.

P. Tan, M. Steinbach, V. Kumar, C. Potter, S. Klooster, and A. Torregrosa. Finding Spatio-Temporal Patterns in Earth Science Data. In *KDD 2001 Workshop on Temporal Data Mining*, volume 19, 2001.

S. Tang and S. Dessai. Usable Science? The UK Climate Projections 2009 and Decision Support for Adaptation Planning. *Weather, Climate, and Society*, 4 (4):300–313, 2012.

K.E. Taylor, R.J. Stouffer, and G.A. Meehl. An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, 93(4): 485, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https:// www.R-project.org.

C. Tebaldi and R. Knutti. The Use of the Multi-Model Ensemble in Probabilistic Climate Projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1857):2053–2075, 2007. doi: 10.1098/rsta.2007.2076.

C. Tebaldi, R.L. Smith, D. Nychka, and L.O. Mearns. Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles. *Journal of Climate*, 18(10):1524–1540, 2005.

H. Tomita, H. Miura, S. Iga, T. Nasuno, and M. Satoh. A Global Cloud-Resolving Simulation: Preliminary Results from an Aqua Planet Experiment. *Geophysical Research Letters*, 32(8), 2005. doi: 10.1029/2005GL022459.

K.E Trenberth. Observational Needs for Climate Prediction and Adaptation. *Bulletin of the World Meteorological Organization*, 57(1):17–21, 2008.

K.E. Trenberth and D.P. Stepaniak. Indices of El Niño Evolution. *Journal of Climate*, 14(8):1697–1701, 2001.

S.M. Uppala, P.W. Kållberg, A.J. Simmons, U. Andrae, V. Bechtold, M. Fiorino, J.K. Gibson, J. Haseler, A. Hernandez, G.A. Kelly, et al. The ERA-40 Re-Analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612):2961–3012, 2005.

A. Voldoire, E. Sanchez-Gomez, D. Salas y Mélia, B. Decharme, C. Cassou, S. Sénési, S. Valcke, I. Beau, A. Alias, M. Chevallier, et al. The CNRM-CM5. 1 Global Climate Model: Description and Basic Evaluation. *Climate Dynamics*, 40(9-10):2091–2121, 2013.

J.M. Wallace and D.S. Gutzler. Teleconnections in the Geopotential Height Field during the Northern Hemisphere Winter. *Monthly Weather Review*, 109(4): 784–812, 1981.

W.M. Washington, L. Buja, and A. Craig. The Computational Future for Climate and Earth System Models: on the Path to Petaflop and Beyond. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1890):833–846, 2009.

A.P. Weigel, R. Knutti, M.A. Liniger, and C. Appenzeller. Risks of Model Weighting in Multimodel Climate Projections. *Journal of Climate*, 23(15): 4175–4191, aug 2010. ISSN 0894-8755. doi: 10.1175/2010JCLI3594.1. URL http://journals.ametsoc.org/doi/abs/10.1175/2010JCLI3594.1.

F. Westad and M. Kermit. Cross Validation and Uncertainty Estimates in Independent Component Analysis. *Analytica Chimica Acta*, 490:341–354, 2003. doi: 10.1016/S0003-2670(03)00090-4.

S. Westra, C. Brown, U. Lall, and A. Sharma. Interpreting Variability in Global SST Data Using Independent Component Analysis and Principal Component Analysis. *International Journal of Climatology*, 346(March 2009):333–346, 2010. doi: 10.1002/joc.1888.

X. Xiaoge, W. Tongwen, and Z. Jie. Introduction of CMIP5 Experiments Carried out by BCC Climate System Model. *Adv. Climate Change Res*, 8:378–382, 2012.

R. Xu and D.C.W. Ii. Recent Advances in Cluster Analysis. *International Journal*, pages 484–508, 2008. doi: 10.1108/17563780810919087.

N. Yussouf, D.J. Stensrud, and S. Lakshmivarahan. Cluster Analysis of Multimodel Ensemble Data over New England. *Monthly Weather Review*, 132(10): 2452–2462, 2004.