THE USE AND APPLICATION OF PERFORMANCE METRICS WITH REGIONAL CLIMATE MODELS

A thesis submitted to the
School of Environmental Sciences
of the
University of East Anglia
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

By
Christopher May
June 2016

Abstract

This thesis aims to assess and develop objective and robust approaches to evaluate regional climate model (RCM) historical skill using performance metrics and to provide guidance to relevant groups as to how best utilise these metrics. Performance metrics are quantitative, scalar measures of the numerical distance, or 'error', between historical model simulations and observations. Model evaluation practice tends to involve ad hoc approaches with little consideration to the underlying sensitivity of the method to small changes in approach. The main questions that arise are to what degree are the outputs, and subsequent applications, of these performance metrics robust?

ENSEMBLES and CORDEX RCMs covering Europe are used with E-OBS observational data to assess historical and future simulation characteristics using a range of performance metrics. Metric sensitivity is found in some cases to be low, such as differences between variable types, with extreme indices often producing redundant information. In other cases sensitivity is large, particularly for temporal statistics, but not for spatial pattern statistics. Assessments made over a single decade are found to be robust with respect to the full 40-year time period.

Two applications of metrics are considered: metric combinations and exploration of the stationarity of historical RCM bias characteristics. The sensitivity of metric combination procedure is found to be low with respect to the combination method and potentially high for the type of metric included, but remains uncertain for the number of metrics included. Stationarity of biases appears to be highly dependent on the potential for underlying causes of model bias to change substantially in the future, such as the case of surface albedo in the Alps.

It is concluded that performance metrics and their applications can and should be considered more systematically using a range of redundancy and stationarity tests as indicators of historical and future robustness.

Acknowledgements

There are a number of people to thank in making this thesis possible. I am greatly indebted to my primary supervisor Dr Clare Goodess for her expertise, guidance and patience since I began at the University of East Anglia in 2011. Her rigour has given me plenty to think about over the years! I would like to thank my secondary supervisors, Dr Manoj Joshi and Professor Phil Jones for many engaging and helpful conversations and feedback.

I would like to also thank the UEA for providing three years funding towards this research. The work itself would not be possible without the vast array of climate model and observational data made available online by the ENSEMBLES, CORDEX and E-OBS projects. I would also like to thank the UEA's High Performance Computing department, with their extremely fast and helpful service, making the computational side of the research possible.

I would like to thank those members of the Climatic Research Unit who have helped me over the years. In particular, Dr Jonathan Barichivich's suggestions and code made the learning curve much easier to tackle!

Finally, I wish to thank my family and Mehrnaz for their support and encouragement towards completing this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Scientific evidence for anthropogenic climate change is widely considered to be overwhelming (Stocker *et al.*, 2013). The burning of fossil fuels on large industrial scales over the last 150 years has increased concentrations of atmospheric greenhouse gases to abnormally high levels. There is high confidence that over 50% of the average global temperature rise as observed over the last century can be attributed to these changes (Bindoff *et al.*, 2013). As a result, national governments seeking an evidence-based response to the challenges posed require information on the expected regional changes. Although future projections based on collections of global climate model (GCM) simulations (commonly referred to as multi-model ensembles) are readily available through large coordinated projects (Meehl *et al.*, 2007; Taylor *et al.*, 2012), they do not have the spatial detail needed to identify likely smaller-scale impacts. Therefore there is a need for higher-resolution regional climate change projections to inform the development of regional vulnerability and impacts assessments and appropriate adaptation policies (Jacob *et al.*, 2014).

The need for better regional information has motivated the production of regional climate model (RCM) ensembles (Van der Linden and Mitchell, 2009; Giorgi *et al.*, 2009) which provide the basis for the construction of regional climate future projections. Assessing these models based on how well they simulate climate processes and dynamics is of vital importance, not only so one can trust the reliability of future simulations, but also so that any systematic errors can be identified for model development purposes. Performance metrics, adopted after their successful use in weather forecasting, are quantitative measures used to evaluate the skill of RCMs in regard to a specific simulation aspect, such as climatological means or the simulation of extreme events, assessed against suitable observational data. Individual metric assessments may then be combined into a single generalised

score which may provide a more comprehensive indicator of overall RCM skill.

The output from metrics or general performance indicators can be utilised in constructing future projections, either through weighting of a multi-model ensemble or by the elimination/selection of subsets of models. The overall aim is to better characterise projection uncertainty based on the inferred reliability of the RCMs. However, the use of metrics in three distinct applications remains problematic. First, performance metrics have been used somewhat without scrutiny in the past (Tebaldi and Knutti, 2007), and thus the degree of sensitivity to the assessment method is unknown. Secondly, the formulation of combination procedures has been to a certain extent ad hoc and output similarly could be largely dependent on approach (Christensen *et al.*, 2010). Thirdly, the use of metric assessed RCM errors for the construction of future projections assumes that these error characteristics retain their underlying behaviour from hindcasts into the future (Räisänen and Ylhäisi, 2011), which likewise has been examined little. Therefore, the use of some performance metrics may be inappropriate as skill measures, the combination of these constituent components into more generalised performance indicators too subjective and the assumed justification for application of these skill measures in future projections perhaps tenuous. These potentially detrimental attributes of performance metrics, if not examined and understood, may place a limitation on the ability to provide robust and objective RCM performance evaluations due to the potential of overly sensitive skill assessments, and also undermine the basis for applying those assessments in methods to inform future regional climate projections.

1.1 Regional Climate Models and The Need for Performance Evaluation

Our understanding about the consequences of climate change on global scales is ever increasing, and the likely impacts affect a wide range of natural and human systems (Field *et al.*, 2014). Socio-economic impacts assessed on global scales are expected in areas such as human health (Patz *et al.*, 2005), food and water security (Hanjra and Qureshi, 2010), coastal zones due to sea-level rise (Nicholls and Cazenave, 2010) and national security concerns (Barnett and Adger, 2007). Ecological effects occur in domains such as biodiversity loss (Bellard *et al.*, 2012), desertification (Stringer *et al.*, 2009) and ocean acidification as a result of CO_2

emissions (Doney *et al.*, 2009). However, many of these impacts, such as droughts or floods, occur on relatively small regional to local scales. Therefore to assess the level of vulnerability, and design appropriate adaptation measures, regional climate models are required to provide a higher level of detail than is available from coarse resolution GCMs (Wang *et al.*, 2004). With the objective of providing more regionally relevant high-resolution projections, RCMs have been developed to produce dynamically downscaled climate simulations of use to a variety of climate change information users. Before a discussion of why the evaluation of RCM performance is important for this task, some background on the types, constructions, errors and uncertainties involved in climate modelling are important to consider.

GCMs are invaluable tools for assessing large-scale anthropogenic impacts on the climate system. They play a central role in climate change attribution studies (Rosenzweig *et al.*, 2008) and provide the essential component for producing future climate change projections. GCMs contribute to our understanding of large-scale climatic phenomena such as El Niño Southern Oscillation (e.g. Stevenson *et al.*, 2012), Atlantic Meridional Overturning Circulation (e.g. Weaver *et al.*, 2012), future sea level rise (e.g. Yin, 2012) and Arctic sea ice extent (e.g. Stroeve *et al.*, 2012). Additionally, information on the direct consequences on global temperatures of greenhouse gas (GHG) emissions through the estimation of climate sensitivity can be produced (e.g. Knutti and Hegerl, 2008; Andrews *et al.*, 2012). Typically, coupled atmosphere-ocean GCMs have been used to assess planetary-scale climatic behaviour, although more recently Earth System Models (ESMs) (such as the HadGEM2 model (Collins *et al.*, 2008)) have been developed to include more components which describe the ecosystem, carbon cycle, ocean biology and atmospheric chemistry. GCMs and ESMs are typically run at a spatial resolution of 100-200km due to computational limitations, which leads to the logical development of approaches to 'downscale' this large-scale information to the regional scale.

Downscaling can either be done dynamically, by using nested RCMs, or statistically, whereby a relationship between large-scale 'predictor' variables and observed station data is exploited by mapping future low-resolution GCM projections to produce local variable timeseries (Wilby *et al.*, 1998). RCMs focus on much smaller continental size domains and operate with considerably higher-spatial resolution than GCMs. 25-50km simulations predominated previous multi-model ensemble projects such as PRUDENCE (Christensen *et al.*, 2002) and ENSEMBLES (Van der Linden and Mitchell, 2009), although ultra-high resolution 11km RCM European

simulations are now being developed and used in the recent CORDEX collaboration (Giorgi *et al.*, 2009). The increased spatial resolution enables a more precise representation of the local orography and land surface characteristics, which in turn is hoped to lead to an improved simulation of local climate phenomena.

Both RCMs and GCMs are highly complex computer simulation programs which simulate the physical processes governing climate. They use systems of mathematical equations to represent the behaviour and interactions of the key components of the climate system: the atmosphere, hydrosphere, cryosphere, and land surface (Stocker, 2011). The simulation domain is discretised in both time and space to varying degrees of resolution, depending on the model application, and each grid-box is resolved using numerical approximation schemes. Advances in computing power have lead to a rapid increase in both the resolution and complexity of climate model design, with the expectation of improvements in simulation quality. Even so, many small-scale processes, such as cloud convection or eddy currents are unable to be resolved explicitly with the latest models (Stan *et al.*, 2010). The reliability of climate models is gauged from several factors: their being based on the established laws of physics (e.g. laws of thermodynamics), assessed simulations of known past states of climate and the level of agreement with present observed climate (Knutti, 2008). However, there remain important sources of uncertainty in the modelling of climate that are crucial to developing the understanding of the most probable future changes.

1.1.1 Sources of RCM Uncertainty and Error

Inherent within climate models are four main sources of uncertainty in the descriptions of the underlying processes and dynamics of the climate system: choice of initial conditions, boundary conditions, parameterisations and model structural considerations (Tebaldi and Knutti, 2007). Modelling uncertainty can be classified either as aleatory (a stochastic or 'random' process, e.g. El Niño) or epistemic (originating from a lack of knowledge, e.g. future GHG concentrations) (Dessai and Hulme, 2004). The total future projection uncertainty is often characterised by the range of model simulations over all available model configurations (Collins *et al.*, 2013). Model error on the other hand is quantified as the difference of a single model simulation against an observational dataset. Uncertainty relates to the degree of certainty one can have of the application of modelling to a problem whereas error relates to the quality of an individual solution. Understanding both

aspects is crucial to improve the performance of a single model and to increase the confidence in future projection results.

Initial conditions in the case of a GCM are usually taken from reanalysis or observations, and for RCMs this would also include GCM forcings as one further type. The climate model is then allowed to run for some period of time, referred to as the 'spin-up' time, to allow the atmospheric circulation to develop. Since errors in initial conditions do not have as large an impact on the long-term evolution of climate statistics as in numerical weather prediction, assessing the consequences of climate change is more commonly defined as a boundary value problem (Collins, 2002). In terms of climate change projections with GCMs, the form of boundary conditions used has changed from CMIP3 to CMIP5. The SRES emissions scenarios (Nakicenovic *et al.*, 2000) used in the earlier ensemble projections specified the amount of GHG gases to be emitted under different assumptions, whereas the newer 'representative concentration pathway' (RCP) approach specifies GHG concentrations instead (Moss *et al.*, 2010). The two approaches although different are ultimately hypothetical in nature, covering a range of plausible futures and therefore contributing to RCM and GCM projection uncertainty.

The additional boundary conditions necessary to force an RCM (sea surface temperatures and atmospheric pressure levels) can be provided either from reanalysis data for simulating hindcasts or from a parent GCM for producing future projections. In the hindcast setup for RCMs the aim is to produce accurate simulations of known past observations by utilising observationally derived reanalysis products. Despite reanalysis forcing being commonly referred to as 'perfect' boundary conditions (Prömmel *et al.*, 2010), regional biases in the climatic fields remain a possibility (e.g. Jaeger *et al.*, 2008). The presence of boundary condition biases has the potential to cloud judgements of RCM skill when we are to assume that the reanalysis forcing does not contribute additional error to simulations. The benefit of using reanalysis forcing is that the large-scale climatic characteristics of the local domain are replicated as close as possible so that RCM output can be evaluated directly against observations. This contrasts to GCM-forced simulations which are likely to contribute additional error to any simulations covering observed time periods. Therefore reanalysis-forced simulations are the least problematic approach to evaluate the skill of RCMs, since the boundary condition error is minimised. This is the approach taken in the relevant analysis in Chapters 4 and 5 covering evaluation of RCM hindcasts. GCM-forced simulations on the other hand

provide the basis for future downscaled climate change projections, and as such the quality of GCM boundary conditions is essential to producing reliable downscaled information.

Parameterisation uncertainty originates from the inclusion (or exclusion) of numerous modelling modules which provide an indirect means of simulating processes that either occur on time or spatial scales smaller than can be resolved by a climate model explicitly (Jakob, 2010) or are too complicated to be represented otherwise (Flato *et al.*, 2013). Examples of parameterisations include the representation of clouds (e.g. Klein *et al.*, 2013), land surface (e.g. Rosolem *et al.*, 2013), convection (e.g. Hourdin *et al.*, 2013) and the atmospheric boundary layer (e.g. Baklanov *et al.*, 2011). Each parameterisation can be evaluated against observations either individually or in the output of the full climate model. The overall uncertainty relating to this third aspect arises due to the number of possible configurations in which a climate model can be constructed. Perturbed Physics Ensembles are one way to assess a single climate model's parametric sensitivity by running a large number of repeated simulations in which each ensemble member has a unique parametrisation setup. Stainforth *et al.* (2005), for example, used this approach when investigating the range of climate sensitivity simulated by 2,578 model variations and found that parametric choice had a high degree of influence.

Finally, although parameterisations can have a substantial effect on simulated quantities of interest, model structural choices are also a large source of uncertainty (Sanderson, 2011). Climate model simulations are computed on discretised three dimensional grids over the spherical planetary surface, ocean and atmosphere. There is no known optimal method to do this. The speed of state of the art supercomputers is the one major constraint on the resolutions that can be employed, and therefore this aspect has increased much in line with microprocessor development. The structural choices following this are the numerical approximation schemes used, and the sensitivity to these components cannot be evaluated through a Perturbed Physics Ensemble, due to structural choices remaining constant over all simulations, but through a multi-model ensemble using model output from several modelling centres (Tebaldi and Knutti, 2007). One structural choice specific to RCMs is the technique used to force simulations, or 'nesting' approach, which consists of a vertical atmospheric component with additional sea-surface temperature data which are of lower resolution than the RCM itself (e.g. Køltzow *et al.*, 2011). Structural uncertainty is assessed through the use of multi-model ensembles

spanning a range of alternative approaches to modelling climate.

The development of RCMs and GCMs to provide both a consistent representation of the climate system and reliable future climate change projections rests on a process of continual model assessment. Model improvements require both simulation error detection and the subsequent diagnosis of the causes thereof. This procedure is very much analogous to that used in operational weather and seasonal forecast evaluation, and accordingly it is helpful to consider the approaches taken within that more mature discipline.

1.2 General Approaches to Climate Model Evaluation

1.2.1 Weather and Seasonal Forecasting Evaluation

The origins of climate modelling lie in numerical weather prediction (NWP), which demonstrated that mathematical modelling is competent to represent the chaotic multi-dimensional dynamics involved in short-term forecasting with a high degree of skill. It is standard practice that NWP centres operate several types of ensemble forecasting system (Zhang and Pu, 2010) producing simulations focussed on near-term 24-36 hour, 1-2 week medium-term out to seasonal/decadal time-frames. Short-term forecasts are of higher spatial resolution; the short-term UK Met Office UKV model for example runs with a variable resolution with 1.5km gridbox size in the domain centre (Pocock *et al.*, 2012). The ensemble members span various model structural formulations and perturbed initial conditions to test the sensitivity of probabilistic forecasts to the model setup. The introduction of more comprehensive spatially-detailed forecasts and further developments to process simulation among other things have led to considerable improvements in forecast quality (Magnusson and Källén, 2013).

The success of the application of mathematical modelling to the problem of weather prediction has relied on the repeated comparison of forecast results against observations. Verification is the test of whether a model is a reasonable representation of the chosen target system, whereas validation is a test of whether the accuracy of the simulation meets some predefined standard (Sargent, 1998). The

continuous nature of producing short to seasonal timescale forecasts lends itself to these examinations of model skill, as models can be assessed over a wide range of statistics. The assessment of model errors is the basis for improving forecasting skill by identifying model deficiencies, in addition to providing information to end users as to which forecasting system is most appropriate to their needs (Gleckler *et al.*, 2008). Furthermore this continual testing has enabled modelling centres to measure changes in forecast skill (Goddard *et al.*, 2013), leading to a quantitative and more objective approach to demonstrate improvements in performance. Rank histograms, root mean squared error (RMSE), ensemble reliability, sharpness and the Brier Skill Score are commonly applied metrics which evaluate the ability of an ensemble to reproduce the statistical characteristics of the observed weather conditions, or to assess the consistency of successive forecasts (Buizza, 2008). The benefits of these standardised metrics in NWP evaluation has lead to the question of whether such an approach could be used to assess climate model performance (Gleckler *et al.*, 2008).

An intermediate step from short-term weather forecasting to climate projections are seasonal to decadal forecasts. These types of forecast are essentially an attempt to predict the likely evolution of those external and internal variability factors which determine the course of medium-term climate (Smith *et al.*, 2012). Although the short term trajectory of weather is chaotic and thus precludes accurate forecasting for timescales beyond a few weeks, there are aspects of climate variability that are to a certain extent predictable on monthly to seasonal timescales (Hansen, 2006), although this underlying quality differs depending on the season and region of interest (Graham et al., 2005). There are two methods (sometimes used in combination) used to produce seasonal forecasts: dynamical (utilising GCMs or RCMs) and statistical (e.g. Lim *et al.*, 2011). To evaluate seasonal or decadal forecast skill the World Meteorological Organisation (WMO) developed standardised evaluation procedures which include several metrics for quantifying the error of probabilistic or deterministic forecasts (WMO, 1992). These include simple RMSE type metrics and the 'Receiver Operating Characteristic' (ROC), adopted from its widespread application in medical science, as a measure of a model's predictive 'hit rate' (or probability of detection) against the false alarm rate of an event (Marzban, 2003). The skill of seasonal forecasting as quantified by these measures has improved due to improvements in the understanding and simulation of the different components of variability (Kirtman *et al.*, 2013).

1.2.2 Climate Model Evaluation Methods

Climate models require comparison with real world observations to ensure consistency with known climatic processes and dynamics and to increase confidence in model reliability when constructing future projections (Reifen and Toumi, 2009). There are a number of approaches to achieve this, varying in complexity and specialisation depending on the purpose of the evaluation. These methods can be placed in one of two groups: statistical evaluations of model errors or assessments of the representation of climate processes. Reasons for evaluating climate models are for the identification of systematic biases aiding model development, or for informing, either through ensemble weighting or model selection/elimination, future climate change projections.

Performance metrics provide a quantitative method to evaluate model error characteristics. Despite their now standardised use in NWP development, the applicability of metrics to the assessment of climate model skill is unfortunately problematic. The most clear difference between the two problems is that climate simulations attempt to model a system evolving over decadal to centennial timescales. This presents difficulties for model verification since the observations required to test climate projections are not available (Tebaldi and Knutti, 2007). Gleckler et al. (2008) notes that a small set of variables may not be enough to assess climate model skill despite being adequate in the case of NWP, and that there is little consensus as to whether an optimal set of metrics could be defined for climate model evaluation. Finally, the agreement of climate simulations with past observations only constitutes a 'necessary but not sufficient condition' for confidence in the reliability of future projections (Xu *et al.*, 2010).

Despite these complications, some of the strengths of NWP methodology have been exploited by the climate modelling community. The multi-model ensemble approach developed in NWP is now commonplace (e.g. Taylor *et al.*, 2012), with the aspiration of a more objective framework for the production of future projections. Additionally, the in-depth evaluation and improvement to all climate model components has lead to increases in model complexity and realism (Flato *et al.*, 2013). In a similar fashion to NWP and seasonal forecast evaluation, the benefits of using quantitative metrics are numerous:

- A standardised assessment technique across the climate modelling commu-

nity leading to a more transparent and objective model development regime.

- The ability to measure the differences in skill within and between successive generations of model ensemble, either for single variables or more comprehensively.
- To provide a useful visual summary (e.g. Taylor diagrams) of absolute and relative model performance for users of climate model information.

The secondary use of metrics outside of the model development or user information provision spheres is in the application of model scores to the construction (by ensemble weighting or model elimination) of future ensemble projections. Qualitative assessment (e.g. expert judgement) may be able to achieve this to a certain extent, but may lack objectivity.

The second group of model assessment techniques involve the detailed inspection of climate process simulation. This can generally either be done by running individual model components and subsequently evaluating full model simulations with the integrated updated component, or by use of a 'regime-based' approach (Flato *et al.*, 2013). The benefits of the former approach are that the mathematical formulae representing distinct processes can be directly tested against observations (if suitable datasets exist), and therefore upgrades can be made to the descriptions of the underlying physics if required (Boone *et al.*, 2009). The latter method enables the evaluation of model processes that occur 'within specific categories that describe physically distinct regimes of the system' (Flato *et al.*, 2013), such as in different 'dynamic and/or thermodynamic' states in the case of cloud process studies (e.g. Williams and Webb, 2009). This approach therefore acknowledges the fact that many processes do not lend themselves to evaluation on strictly seasonal or annual time domains, and thus regime-based methods can be used to investigate climate dynamics as they occur in correspondence with reality.

Specific to RCMs is the possibility of direct comparison to high-quality (depending on domain) gridded observational datasets. Since the current practice has been to force RCMs with reanalysis (and GCMs), it is a logical step to use performance metrics for the purpose of RCM error assessment (Sánchez *et al.*, 2009; Christensen *et al.*, 2010; Coppola *et al.*, 2010; Kjellström *et al.*, 2010; Holtanová *et al.*, 2012; Giorgi *et al.*, 2012; Kim *et al.*, 2014; Evans *et al.*, 2014; Kotlarski *et al.*, 2014). This approach resolves the difficulty of using GCMs, which are likely

to be a substantial contributor to simulation error, as the provider of boundary conditions. However, the sensitivity of performance metrics to the underlying assessment methodology has received little attention thus far, with many studies using ad hoc types of skill measure without consideration of the robustness of the method. Therefore the use of such quantitative performance metrics to evaluate RCM performance, given the importance of robust assessment, is the focus of this thesis. The reason for focussing on performance metrics is due to the need for more objective quantitative approaches to climate model evaluation, in addition to the increasing trend towards the use of multi-model ensembles, which are ideally suited to this application.

1.3 The Use of Performance Metrics

1.3.1 Performance Metric Sensitivity

What is a performance metric? As a working definition it is beneficial to describe what constitutes a performance metric and what the preferable characteristics of such a measure might be.

Definition: A **Performance Metric** is a quantitative measure of climate model error composed of four distinct elements: variable, statistic, spatio-temporal domain and reference dataset.

Desirable characteristics for a particular metric might be:

- Gives output which is robust/insensitive to small changes in evaluation methodology.
- Is able to produce information on general or specific model simulation aspects.
- Can provide a quantitative measure of the magnitude or direction of model errors.

The use of performance metrics with RCMs has in the past been somewhat arbitrary in terms of metric construction (for further details see Section 2.1), without regard for the effect these underlying choices have on output sensitivity. RCM evaluation using performance metrics relies on the premise that they are robust; that they will not produce inconsistent or overly volatile results if subjected to minor

changes in application. The consequences of this not being the case are problematic. Assessment of RCM skill in relation to the simulation of a given process may be misrepresented, potentially masking or exaggerating an issue, distracting or undermining efforts to improve overall model performance. Additionally, if a RCM ensemble is used for the production of future climate change projections through methods such as ensemble weighting using output from sensitive metrics, the future changes and uncertainty information may be misleading and unfounded.

Although some desirable characteristics of performance metrics may be implicitly assumed in RCM evaluation, without further investigation it remains an open question whether the use of metric output based on a single set of construction choices is fully justifiable.

1.3.2 Metric Combination Approaches

Comprehensive measures of RCM performance, encapsulating a range of simulation aspects, are one approach toward a more holistic and rigorous model evaluation process. Through the consideration of more than one performance metric, it may be possible to arrive at a single overall indicator of model performance.

Definition: A **Generalised Performance Indicator (GPI)** is an amalgamation of several chosen performance metrics, combined in such a way as to produce a scalar value quantifying overall or specific targeted model performance.

Alternatives to commonly used methodological approaches with GPIs have, in a similar vein to the use of single performance metrics, been under-sampled (see Chapter 2.2). Most studies utilise identical combination methods, mainly geometric multiplication of metric output, which is not the only possibility. Additionally, the number and type of metric included in combination is also important; if a pair of included metrics are giving the same information, then the overall indicator score may be biased in some respect.

Therefore more wide ranging sampling of combination possibilities in tandem with a more systematic and objective approach to choosing which metrics or performance aspects to include is needed.

1.3.3 Stationarity of Model Errors in Future Projections

Attempting to reduce uncertainty in regional future projections through the use of metric evaluations, by adjusting, constraining or elimination of model ensemble output, has been a central theme for improving projected climate changes (Chapter 2.3). In the simplest case of bias correction for example, the calculation of mean error would be used and removed from future projected changes, thereby arriving at the 'true' (or most plausible under the emissions scenario assumptions) climate. However, the direct application of model errors in hindcasts to future projections relies on a major assumption: that the distribution of model ensemble errors remains stationary in time; that whatever emissions scenario and resultant lateral boundary condition forcing is applied, the error characteristics of each RCM will respond uniformly. Given the presence of potential 'tipping points', nonlinearities within the climate system, such behaviour in projections is not guaranteed. Furthermore, if GPI output is used to 'weight' model ensembles in future projections the model error characteristics measured by each individual metric component have to be stationary.

Therefore, although not directly verifiable, this assumption requires further scrutiny as it is the basis for many efforts to reduce or constrain uncertainty in future regional downscaling methods.

1.4 Aims and Objectives

This thesis investigates the use and application of performance metrics with RCMs. The 'use' of metrics relates to how to choose a metric, or set of metrics, in their capacity as an indicator of model performance for that specific variable, region and domain of interest. 'Application' of performance metrics relates to the use of metric output in several other capacities; either collectively, generating measures of overall model performance, or as indicators of future projection reliability. Performance metrics, as measures of model error, have a degree of subjectivity in their selection, commonly giving rise to ad hoc use with little consideration for underlying sensitivity to small changes in method. This has implications for how performance metrics are used in the validation, overall assessment, selection/elimination and future projection weighting of RCMs. To better understand and identify appropriate approaches using and applying performance metrics, the thesis has three aims:

- To assess and develop objective approaches for the assessment of RCMs using performance metrics.
- To investigate and develop criteria and analysis methods more likely to provide robust outcomes from the application of performance metrics.
- To provide guidance and recommendations to relevant groups on the use and application of performance metrics.

To achieve these aims, three objectives are identified which the analysis in Chapters 4, 5 and 6 respectively explore in detail:

- To investigate the sensitivity of performance metrics to small changes in methodological approach.
- To evaluate the use of GPIs as robust measures of overall RCM performance.
- To assess the plausibility of the stationarity assumption in climate change projections.

The first two aims have a specific focus on assessing and developing more objective methods of RCM evaluation. Relevant to this is the investigation and development of model evaluation criteria more likely to provide robust outcomes: what factors should be considered? The third aim is to provide guidance to relevant parties (e.g. model developers, impacts modellers and policy makers) as to how best to utilise performance metrics in their areas of work. To further investigate these three aims, various types of analysis method are applied tailored for each specific topic. Further methodological information is provided in each relevant analysis chapter, and a brief outline provided in the next section.

1.5 Structure of Thesis

Chapter 1 provides a brief description of climate change, the production of regional climate change projections, sources of RCM uncertainty and error, and the need and development of RCM robust and objective evaluation methods. Performance metrics and GPIs are introduced, in addition to the concept of non-stationary simulation biases.

The Literature Review (Chapter 2) surveys the current understanding of three

distinct but related topics: the current practice of approaches to performance metrics, the development of metric combinations methods, and of their use in constructing future climate change projections, reliant on the stationarity assumption.

Chapter 3 details desirable specifications of RCM and observational data for use in model evaluation studies. It provides detailed examination of the reanalysis and GCM-forced ENSEMBLES and CORDEX RCM and E-OBS gridded observational data used in the three analysis chapters. The format of the model and observational data in addition to pre-processing and standardisation methods are also discussed.

The first analysis Chapter 4 investigates the sensitivity of performance metrics in their use in assessing reanalysis forced RCMs in hindcasts. The aim is to identify more robust model evaluation procedures which have an objective basis for their application, and also to reduce the level of metric redundancy. 'Redundancy' is defined as the degree to which information provided by a metric can be obtained using another metric, thereby making the use of both unnecessary.

Chapter 5 Metric Combinations explores methods of producing generalised performance indicators (GPIs), used to give an overall comprehensive test of model skill, to understand the sensitivity of these overall skill measures to changes to their inputs (different choices of metric) or the combination approach (multiplicative/additive) used. The number and type of metrics included in the GPI is varied as to observe to what extent such overall indicators of model performance are influenced by such changes in assessment criteria. A final expert set of metrics is given for combination in an overall score of RCM performance.

The Stationarity Assumption is explored in the final analysis in Chapter 6, assessing whether performance metric assessments relative to other ensemble members or the 'true' future climate may remain constant in time, or behave in a non-stationary way. This indirectly tests whether and in what circumstances the stationarity premise, relied upon by bias-correction and by methods that directly apply quantitative RCM assessments, is reasonable in future climate change projections.

The final Chapter 7, Conclusions, Recommendations and Outlook, provides an overview of the results from the analysis of Chapters 4, 5 and 6, detailing key findings and points. The implications and recommendations drawn from these

results for relevant groups is discussed. Recommendations for particular analysis methods, approaches and criteria for metric and GPI construction are provided, with potential development of these methodologies and further analysis suggested.

Chapter 2

Literature Review

2.1 Performance Metrics: Sensitivity and Robustness

The evaluation of climate model skill through the use of performance metrics plays a central role in measuring progress towards more realistic simulation of climate (Räisänen, 1997; Moise and Delage, 2011). Performance metrics give a concise quantitative measure of a model's ability, or skill, to accurately reproduce climatic observations over a given validation period. By utilising such statistical measures, one may identify the strengths and weaknesses of a given model and potentially diagnose which processes are represented poorly (Phillips *et al.*, 2004). In addition, one may simply rank the models by order of skill, or compare the performance of different generations of climate model. A common finding is that no single model consistently outperforms all other ensemble members when evaluated over all variables in both GCM and RCM ensembles (Lambert and Boer, 2001; Gleckler *et al.*, 2008; Christensen *et al.*, 2010), inducing a limitation on absolute judgements of which model is 'best' overall. This is due to the fact that different sets of metrics could give different answers. Moreover, it is unclear which processes or variables are most relevant for determining how well a model reproduces climate, and at what scales these should be measured, be it temporal or spatial. Thus the task of constructing and choosing metrics for model performance evaluation remains subjective, leading to disagreement as to what the most justifiable approach might be (Tebaldi and Knutti, 2007).

The usefulness of performance metrics is reliant on their providing robust, relevant information on the absolute and relative model performance in historical

simulations, and giving an indication of model reliability in future projections. A 'robust' assessment by a metric would require it to be insensitive to small changes in the construction of the test. 'Relevant' information would be assessments which assist in identifying areas of model deficiency, in terms of the direction and magnitude of errors or the presence of systematic model biases across an ensemble. Absolute performance is related specifically to the magnitude of model errors against observations, whereas relative performance pertains to the distribution of model performance within a multimodel ensemble, the secondary conclusions from direct comparison with observations (Radic and Clarke, 2011). Thus far there has been some effort made to explore different metric constructions with a wide array of variables or statistic choices, in particular for GCMs but less so for RCMs. Comprehensive approaches assessing GCMs such as Murphy *et al.* (2004) and Reichler and Kim (2008), utilising a single statistic applied over a wide array of variables, have attempted to produce generalised performance indicators. RCM approaches (Coppola *et al.*, 2010; Eum *et al.*, 2012; Fowler and Ekström, 2009; Holtanová *et al.*, 2012; Kjellström *et al.*, 2010; Lenderink, 2010; Lorenz and Jacob, 2010; Perkins *et al.*, 2007; Pierce *et al.*, 2009; Sánchez *et al.*, 2009) have been more focussed in their scope, aiming to capture model performance for a single or multiple variables with more than one corresponding statistic. Some studies have taken a further step in using information obtained from performance metrics to construct future climate change projections, either through model ensemble weighting techniques or by eliminating those models judged to be unrealistic and thus inappropriate for use (e.g. Holtanová *et al.*, 2012). However, a thorough examination of the inherent uncertainties involved in metric construction is presently lacking for such uses, particularly in the case of RCMs, due to somewhat arbitrary methodological choices without regard in many cases to the underlying sensitivity of results. This has the potential to undermine conclusions drawn from performance metrics if they are not adequately scrutinised.

The subjective aspects involved in constructing metrics originate from four specific methodological choices (Pincus *et al.*, 2008);

- Variable
- Spatio-temporal domain
- Statistic
- Observational reference dataset

One of the first decisions in evaluating the performance of a climate model is to determine which variables should the simulation quality be assessed with respect to. Applications of performance metrics fall into two main groups: either specific process-based assessments, or more general 'overviews' of model performance covering a wider range of variables (Ma *et al.*, 2013). Each type has certain requirements for which variables are needed for the evaluation to meet its purpose. Process-based studies investigating the representation of large scale modes of variability such as ENSO or the MJO would be interested in variables that relate to the underlying physics of the system such as those governing feedback processes or specific patterns of spatio-temporal variability. For example, in the case of El-Niño Southern Oscillation this approach has been taken by evaluating the reproduction of the Bjerknes feedback in the CMIP5 ensemble (Bellenger *et al.*, 2014), whereas the latter approach was applied to metric assessments of Madden-Julian Oscillation representation wherein an Empirical Orthogonal Function analysis of Outgoing Longwave Radiation and zonal winds was used (Kim *et al.*, 2009). In the case of RCMs, the large scale circulation is provided by either reanalysis or GCM boundary conditions and as such the reproduction of large scale modes of variability not usually the focus of performance assessment. However, there are examples of studies investigating the effect of these systems on RCM simulation quality, such as the influence of ENSO on South African RCM simulations (Boulard *et al.*, 2013). More common to RCM process-based assessments are studies which tend to focus more on the representation of climate feedbacks such as the soil moisture-precipitation feedback (e.g. Jaeger and Seneviratne, 2011).

In studies exploring performance metrics, this role of variable choice has been explored most, with the previously mentioned GCM studies of Murphy *et al.* (2004) and Reichler and Kim (2008) being the notable examples. Within their Climate Prediction Index (CPI), Murphy *et al.* (2004) calculated a range of metrics spanning a broad range of 32 atmosphere and surface variables. These included the most commonly assessed 1.5m temperature and precipitation variables, but also less frequently investigated aspects such as top-of-atmosphere outgoing short and longwave radiation and sensible heat fluxes. Reichler and Kim (2008) similarly evaluated GCMs over a set of 14 variable types, chosen primarily due to the availability of model data. Both studies have shortcomings with respect to the remaining three metric aspects (spatio-temporal domain, statistic and observational dataset). The choice of spatio-temporal domain in both studies involved an evaluation of model performance over the total global domain, and the temporal domains were

limited to mean climatologies. Each study employed a single statistic involving a variant of RMSE for all variables, and finally, only one observed dataset is used as reference for each variable (some are reanalysis products, which may have additional inherent error). As a result, each metric within both studies has several limitations. As an evaluation of GCMs, use of the total global domain is not surprising, but ignores spatial variation in model performance. Similarly, evaluation only of mean climatologies does not sample higher order moments of variability. Furthermore, the statistic used ignores whether such a choice can have an effect on the inferred level of performance. The consequence of these limitations is that for each of those evaluated metrics, one cannot know how reliant each assessment is on those specific metric aspects selected.

RCM metric studies thus far have not been as wide ranging with respect to how variable choice affects model evaluations as those extensive GCM multiple-metric analyses referred to. Several studies have investigated ENSEMBLES RCMs for a range of different variables. Kjellström *et al.* (2010) assessed the simulation of daily temperature and precipitation over the European and sub-domains, going further than assessment of the mean climatology by considering the percentile distribution of simulations. They found that for both variables, evaluations of the mean climatology were not fully representative of the whole ensemble performance; model errors tended to increase in magnitude when extremes were considered, but the general distribution of model performance (i.e. 'good' or 'bad' models remaining so across the percentile range) remained relatively constant. One drawback of their study was the limitation to only two metric types (Cumulative Distribution Function and Probability Density Function) and use of a single observational dataset to evaluate the models. Therefore the degree to which the model evaluations would be similar in comparison to other alternative datasets, particularly at the extremes, is an unknown factor. Coppola *et al.* (2010) assessed RCMs in their representation of seasonal temperature and precipitation mesoscale behaviour, with the aim of testing the 'added value' of dynamical downscaling for Europe. Model performance in wind simulation was investigated by Donat *et al.* (2010), who found that model skill was influenced heavily by the representation of highly variable orography. Lenderink (2010) studied precipitation extremes relative to the E-OBS gridded observational dataset, and found that there was little change in model performance when assessed over different sub-domains or seasons, but found a high level of sensitivity to the choice of metric. Lorenz and Jacob (2010) evaluated long term historical simulation temperature trends in ENSEMBLES RCMs relative to E-OBS,

and found a systematic underestimation of the observed trend.

More recently, the CORDEX regional climate modelling project (Giorgi *et al.*, 2009) has provided a range of RCM historical simulations and future projections for several global regions. Kotlarski *et al.* (2014) evaluated the EURO-CORDEX ensemble (Jacob *et al.*, 2014) in its representation of seasonal and monthly temperature and precipitation relative to E-OBS. They found that systematic model errors identified in previous RCM ensembles, such as the prevalence of 'drizzling' in most RCMs, were also found in the latest RCMs. Vautard *et al.* (2013) assessed the EURO-CORDEX RCMs for simulation of temperature extremes, specifically heat waves, and found that models had a high degree of spatial variability in model performance. Northern regions tended to simulate too cold temperatures, whereas southern areas were too warm in reproducing the observed 90th-percentile of mean temperature. Interannual variability of heat waves is found to be simulated with a reasonable degree of accuracy, although it is suggested that this behaviour is likely more a product of the boundary forcing rather than inherent model skill. Africa has been of particular focus of dynamical downscaling with RCMs within the CORDEX framework. Nikulin *et al.* (2012) evaluated the representation of precipitation for 50km resolution RCMs, including the West African Monsoon and seasonal climatologies, against several observational products. They found that models were reasonably skilful in temporal and spatial precipitation variability, although some models presented large systematic errors. RCM skill in simulating mean, minimum and maximum temperature, precipitation and cloudiness with CORDEX-Africa was investigated by Kim *et al.* (2014), who found that model performance, although generally of good quality, varies with the region in question, posing difficulties for further evaluation and use of RCMs information for end users.

There has been some exploration of the role of spatial variability in RCM metric evaluations; the European 'Rockel' regional sub-domains proposed by Rockel and Woth (2007) being one illustration of the possibilities in this area. Additionally, some statistics have been tested which explore beyond the standard error statistics (e.g. RMSE) used in earlier GCM studies. Evaluations using CDF and PDF (e.g Sánchez *et al.*, 2009), linear trend (e.g. Eum *et al.*, 2012) and spatial correlation (e.g. Xu *et al.*, 2010) methods provide a more in-depth consideration of other model simulation aspects. However, these studies although spanning a range of alternative methodologies, do so in isolation from one another, precluding a comparison between RCM evaluations. Therefore there remains a question as to how these

different methodologies compare in their assessment of model performance.

The effect of observational choice also has the potential to have a substantial influence on model performance evaluations. Thorne *et al.* (2005) describes how methodological decisions and limitations on station data availability in creating observational datasets result in a level of unavoidable structural uncertainty, giving rise to a degree of variation between products. Therefore, the differences between two gridded observational or reanalysis data used to quantify model simulation quality may give rise to dissimilar model error evaluations. Although many studies assessing RCM or GCM performance do not focus on this potentially important aspect of metric construction in favour of spanning a number of variables and statistics, there are some studies directly addressing this issue.

Sillmann *et al.* (2013) evaluated the CMIP3 and CMIP5 ensembles in their reproduction of a set of extreme indices proposed by the Expert Team on Climate Change Detection and Indices (ETCCDI). As reference, they chose four reanalysis datasets with which to evaluate models: NCEP-DOE, NCEP/NCAR, ERA-40 and ERA-Interim. The magnitude of differences in some regions between the various reanalysis data was found to be considerable, and as a consequence some indices of extremes were heavily influenced by the choice of reanalysis. One finding was that the common outperforming characteristic of the multimodel mean versus the best performing model was insensitive to the choice of reanalysis, even though the relative individual model performance within both ensembles was not. This result adds weight to the use of the multi-model mean given its apparent robustness relative to changes in observations. Sillmann *et al.* (2013) also found that the evaluation of extreme climate statistics specifically using reanalysis data was problematic due to differences between products, thus leading to the recommendation that gridded observationally derived datasets be used in future studies specifically targeting climate extremes.

Such a study was undertaken by Casanueva *et al.* (2013), who assessed the performance of five ENSEMBLES RCMs in addition to five statistical downscaling methods over the Iberian peninsula for maximum and minimum temperature 5th and 95th percentiles. Two observational based gridded datasets were used, the 25km E-OBS and 20km resolution Spain02 (Herrera *et al.*, 2012), with some discrepancy found between the two, particularly for the winter 5-th percentile of minimum temperature. Although no specific quantitative assessment was made

of the sensitivity of metric evaluations to the choice of observational dataset, they suggest that there is a potential, given the differences between the two datasets used, for such decisions to influence metric assessments regarding model performance. This specific question was investigated by Gómez-Navarro *et al.* (2012) who similarly tested ENSEMBLES RCMs over Spain for precipitation and both minimum and maximum temperature variables. They considered the ranking of models based primarily on the spatial correlation statistic against three reference datasets; E-OBS, Spain02 and the Spanish Meteorological Centre's AEMET dataset and found that model rankings were highly dependent on observational reference; in some cases models were gauged to be better or worse performing subject to the chosen observations. However, it could be argued that such behaviour for extreme values is not unexpected, given that structural differences in observational dataset construction, such as interpolation or smoothing methods, can lead to substantial differences between different datasets especially in the tails of the distribution (Hofstra *et al.*, 2010), whilst being less apparent when assessing models for the mean climatology (Sylla *et al.*, 2013).

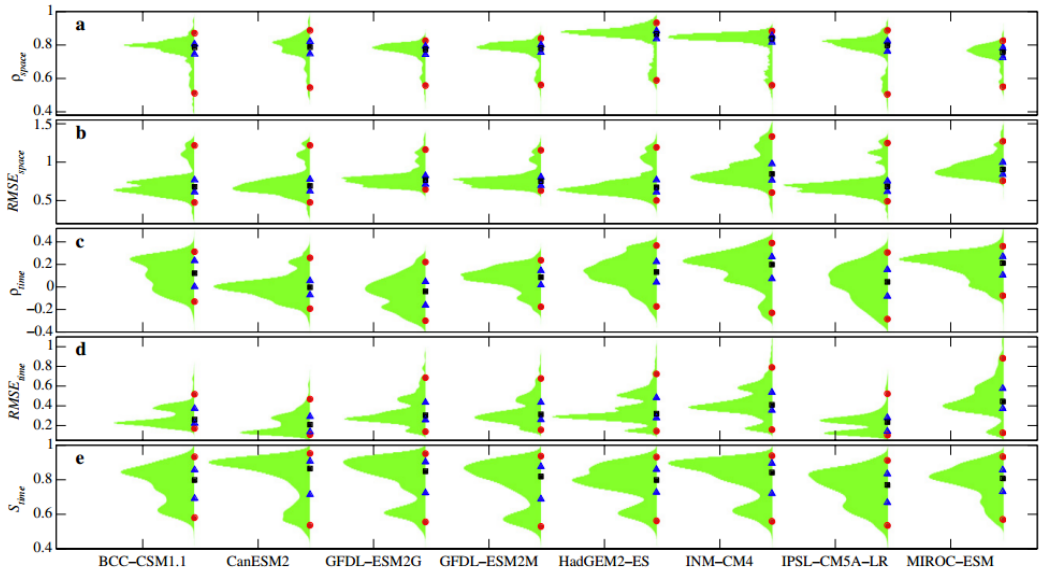

Figure 2.1: GCM metric scores for CMIP5 GCMs assessing their performance in simulating evapotranspiration. PDFs represent the range of scores generated in assessing GCMs for (top to bottom) spatial correlation, spatial RMSE, temporal correlation, temporal RMSE and distributional similarity from using six observational datasets. Modified from Schwalm *et al.* (2013).

A more in-depth investigation of metric construction sensitivity regarding the three choices of observational data, spatio-temporal domain and statistic was

undertaken by Schwalm *et al.* (2013). They concentrated on one variable, evapotranspiration, for which they used six global datasets as reference. The eight models assessed were from the CMIP5 ensemble, evaluated over five statistics; spatial and temporal RMSE, spatial and temporal correlations and the PDF statistic of Perkins *et al.* (2007). For sampling on the temporal scale, models were assessed for each 10-year period covering the full observational time length. In all, for each model and metric construction combination, 68,700 evaluations are calculated to provide a range of metric output for each statistic. They found that each metric assessment had a high degree of sensitivity, especially with respect to the choice of observational dataset, which was found to be the most influential of all metric construction aspects. One result from their work is shown in Figure 2.1, in which eight GCM performance assessments are found to span a wide range of statistic values in response to changes in reference dataset. Furthermore, model rankings were found to be not of a single value, but a range of plausible ranking values depending on the choices underlying the chosen metric. They suggest that their results give reason to be hesitant of definitive conclusions based on single metric constructions not spanning a range of choices, and consequently that metric assessments of model relative performance should not be considered as definitive in character.

Whether or not the conclusions of Schwalm *et al.* (2013) are as relevant for more commonly considered variables such as mean temperature or mean climatologies however, has yet to be shown. Furthermore, one cannot expect GCMs to display highly correlating temporal skill over different short time periods due to the inherent internal variability. In the case of reanalysis forced RCMs, this may be a fairer test given that one should expect the RCMs to be temporally synchronised with the observations. Nevertheless, the results of Schwalm *et al.* (2013) in addition to Sillmann *et al.* (2013), Kjellström *et al.* (2010), Murphy *et al.* (2004) and Reichler and Kim (2008) would give cause to investigate further the effects of metric construction choice on apparent model performance. In particular, if metric assessments are to be used as a basis to construct future projections by, for example weighting ensembles, a clearer understanding of how sensitive different types of metric are with respect to their underlying construction is vital, otherwise such uses of metric output might be considered overconfident.

2.2 Metric Combination Approaches

A single performance metric, although able to give a concise quantitative indication of model skill, is limited in its assessment to the specific variable aspect in question. A broader and more comprehensive assessment of model performance across a range of simulation aspects is desirable to produce generalised performance indicators, or GPIs, so that overall climate model skill can be summarised in as compact a format as possible. The potential benefits of doing so would be to evaluate the improvements of multimodel ensembles over different successive generations, and to provide a useful summary benchmark for users of RCM or GCM data on which they can quickly assess, for example, whether they wish to use those models. Metric combination approaches are a logical solution to this problem, due to their ability to take into account a range of factors relevant to whatever is considered 'good' performance. A good example of this is provided by the I^2 GPI of Reichler and Kim (2008) (Figure 2.2), where the performance of individual GCMs, in addition to that of several generations of model, can be assessed quickly and intuitively. However, there remains much scope in what such combination methods might entail, and whether it is possible to reach a well defined, optimal and justifiable method has yet to be demonstrated, with most approaches again using ad hoc constructions.

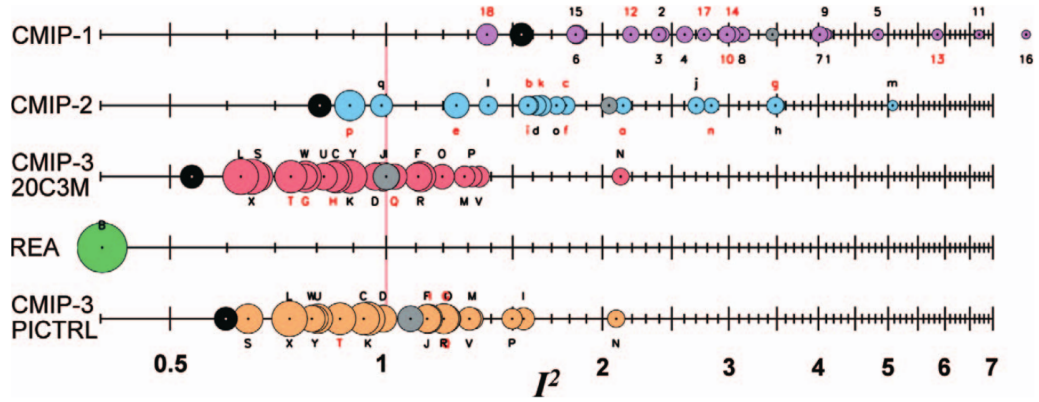

Figure 2.2: I^2 metric combination scores for CMIP1, CMIP2, CMIP3, and CMIP3 pre-industrial GCMs providing a measure of overall performance covering 14 mean climatological variables, including sea-level pressure, surface heat flux and snow fraction. Coloured dots represent individual GCMs, back dots multimodel means, whereas grey dots represent the average I^2 value for the ensemble. The REA green dot represents the I^2 value of the NCEP/NCAR reanalysis. Smaller values of I^2 indicate better overall performance when assessing GCMs relative to gridded observations or reanalysis (depending on the variable in question). Modified from Reichler and Kim (2008)

Giorgi and Mearns (2002) considered two distinct aspects of model simulation into one combined model assessment framework, named the 'Reliability Ensemble Averaging' (REA) method. It aims to provide an indicator for the reliability of projections for a single variable, in addition to the average change from historical to future scenario and uncertainty estimates. Their approach involves the calculation of two factors: $R_{B,i}$ and $R_{D,i}$, and combining into one 'Reliability Factor' R_i according to

$$R_i = \sqrt[mn]{(R_{B,i})^m \cdot (R_{D,i})^n} \tag{2.2.1}$$

where

$$R_{B,i} = \frac{\epsilon_T}{ABS(B_{T,i})} \tag{2.2.2}$$

and

$$R_{D,i} = \frac{\epsilon_T}{ABS(B_{D,i})} \tag{2.2.3}$$

The factors of ϵ_T refer to the long term average of observed precipitation or temperature variability. This is calculated first by detrending an observed timeseries, then producing a 30-year moving average timeseries; for each sliding window of 30 years duration, a single value is produced. Next to estimate the natural variability, two alternative options are available: calculation either of differences between predefined percentiles or the difference between maximum/minimum values of the 30-year averages. Giorgi and Mearns (2002) recommend the use of the latter approach for shorter observed timeseries, given that ϵ_T is probably an underestimation of natural variability in this case. The skill of model i is given by $R_{B,i}$, the inverse of the absolute model bias relative to observations; the smaller the bias, the higher the skill score. $B_{T,i}$ refers to the model error relative to observed mean climatology covering the period 1961-1990, $B_{D,i}$ is defined as the difference between the average projected future temperature change (mean temperature 2071-2100 - mean temperature 1961-1990) of GCM i and the ensemble's REA weighted average change. This is necessarily an iterative process; the first value of the REA weighted projected change is simply the standard multimodel mean change, $B_{D,i}$ are defined as relative to this, and thus the REA is recalculated and so on. $R_{D,i}$, the degree of convergence, calculates the 'distance' of the projected average temperature or precipitation of model i to the ensemble mean change. The further away a model's projected change is from the average of all models, the lower the score assigned. The exponents m and n in R_i can be varied to emphasise one factor over the other. Although not truly a metric combination approach, given

that the second component is not strictly a performance metric and more precisely a parameter assessing overall ensemble behaviour, the REA method was an early attempt at considering more than one simulation aspect in its evaluations.

Despite the novel approach in including more than one criterion in the evaluation procedure, two main criticisms to the stated components can be raised. First, it is unlikely that a metric relying on a single variable for information can be fully representative of the overall model performance (Gleckler *et al.*, 2008). Second, the convergence factor assumes that the mean projection is derived from a randomly sampled ensemble producing independent members, which is unlikely (Tebaldi and Knutti, 2007). This model independence is an issue for the convergence factor because if some models are more closely related, and therefore showing a level of dependence, they will be more likely to produce similar projected temperature changes. For example, going to the extreme case, if an ensemble of three models has two identical members, then these two will be judged more 'reliable', by the fact that their projected changes will be closer to the (biased) mean change. The point is that there is nothing intrinsically about these two simulations which will make them more reliable, other than the fact that they are identical. Therefore, models which produce projections closer to the ensemble average cannot conclusively be said to be more reliable, due to the fact that the similarity may be as a result of common assumptions or systematic biases. Furthermore, by penalising those models whose projections lie towards the tails of the distribution when weighting projections, such a condition will lead to an unjustified narrowing of the range of changes due to the under-sampling of anomalous results (Knutti, 2010). Despite the questionable assumptions the REA method relies upon in the choice of its constituent elements, this early study has initiated further research into more comprehensive combined metrics incorporating different and an increased number of components.

More recent studies have aimed to provide metric combinations spanning multiple variables and variability moments. A further development of the approach of Giorgi and Mearns (2002) was done by Xu *et al.* (2010), who proposed an updated version of the REA method motivated by the criticisms directed toward the original configuration. They assessed 18 GCMs from the CMIP3 ensemble simulating the East Asian domain against gridded observations. The problematic measure of convergence was removed in favour of evaluating model skill strictly on the reproduction of observations alone, and incorporated several additional tests of performance spanning three variables simultaneously: temperature, precipitation and sea-level

pressure, addressing the second main criticism. The metric R described is a combination of five 'sub-metrics':

$$f_1(\bar{T}) = \frac{\epsilon_t}{ABS(T)} \tag{2.2.4}$$

$$f_2(T_{var}) = \frac{\epsilon_{\sigma_t}}{|\sigma_m - \sigma_o|} \tag{2.2.5}$$

$$f_3(\bar{P}) = \frac{\epsilon_p}{ABS(P)} \tag{2.2.6}$$

$$f_4(P_{var}) = \frac{\epsilon_{\sigma_p}}{|\sigma_m - \sigma_o|} \tag{2.2.7}$$

$$f_5(SLP_{corr}) = corr(SLP_m, SLP_o) \tag{2.2.8}$$

Here, σ_m and σ_o are defined as the interannual standard deviation of temperature and precipitation in f_2 and f_3 respectively. $corr(SLP_m, SLP_o)$ is the spatial correlation of sea level pressure between observed and simulated fields. These are then composed into a final score R based on their product:

$$R = \prod_{j=1}^{5} f_j^{m_j} = f_1^{m_1} \cdot f_2^{m_2} \cdot f_3^{m_3} \cdot f_4^{m_4} \cdot f_5^{m_5} \tag{2.2.9}$$

The factors of f_1 and f_3 are of the same form as $R_{B,i}$ in Giorgi and Mearns (2002), but take into account performance of both temperature (T) and precipitation (P) in the final score. f_2 and f_4 are measures of a model's skill to simulate interannual variability of T and P. Finally, sea level pressure (SLP) performance is included with sub-metric f_5 as a measure of spatial correlation skill. The factors of ϵ in $f_1 \ldots f_4$ remain measures of natural variability, but are now calculated for each specific variable tested. The exponents m_j of each sub-metric can be altered following on the original approach of the REA metric.

To analyse the sensitivity of model scores to the metric construction choice they varied the values of $m_1 \ldots m_5$ over six permutations, including combinations compromising T or P only, and an 'unweighted' ensemble mean from $m_j = 0 \ \forall \ j \in [1, \ldots, 5]$. By evaluating models from the CMIP3 ensemble taken over several East Asian domains, they found that the total metric scores R for each GCM varied most with the inclusion or exclusion of the precipitation focussed sub-metrics,

which indicated that the range of ensemble simulations for this specific variable was greater than that for sea level pressure and temperature. One possible criticism of the study could be due to the fact that the multiplicative (or geometric) combination scheme itself was held constant, which potentially could be a source of sensitivity if an alternative approach were used.

Murphy *et al.* (2004) developed the concept of combined metrics by introducing the 'Climate Prediction Index' (CPI) with the aim of improving projections of climate sensitivity with a perturbed physics GCM ensemble. Their metric employs 32 climatic variables in one combined assessment of model skill, and is calculated by first taking the normalised root mean squared errors (RMSE) of model and observational temporal means for variable k and season j:

$$CPI_{jk} = \sqrt{\frac{1}{n\sigma^2_{ANN}} \sum_{i=1}^{n} (M_i - O_i)^2} \tag{2.2.10}$$

Here, σ^2_{ANN} is found first by calculating the simulated interannual variance for each grid point, then averaging spatially to produce the final quantity, n gridpoints and M_i, O_i are 20-year mean values for model simulations and observations respectively at gridpoint i respectively. The final score CPI^2 for each model is given by first taking each CPI_k averaged over all seasons,

$$CPI_k = \frac{1}{4} \sum_{j=1}^{4} CPI_{jk} \tag{2.2.11}$$

and then taking the product of the squares of each resultant CPI_k.

$$CPI^2 = \prod_{k=1}^{32} CPI_k^{2m_k} \tag{2.2.12}$$

The weights m_k given to each variable are equally set to 1 except in the case of cloud height optical thickness, which are given $1/3$ weighting due to the acknowledged interdependencies of the variable. 'Interdependencies' refers to the fact that the nine cloud cover varables are closely related; there are three optical thickness and height categories, and so 1/3 weighting is applied to each to normalise this variable category. The potential for sub-metric weighting has been explored by Xu *et al.* (2010), but definitive conclusions as to whether this is necessary, or can be done objectively remains an open question. Reichler and Kim (2008) adopted a similar method in their I^2 model performance index in their study investigating

improvements in different generations of GCM ensembles. Their proposed metric comprised 14 variables, combined as in the CPI but with a small scaling factor to take into account error differences in each generation of model ensemble. They found that the I^2 metric is relatively insensitive to changes in the variables included, with skill scores converging as the number of such parameters is increased. As a result, it was concluded that I^2 successfully differentiates different ensemble generations with respect to their ability to replicate mean climate statistics.

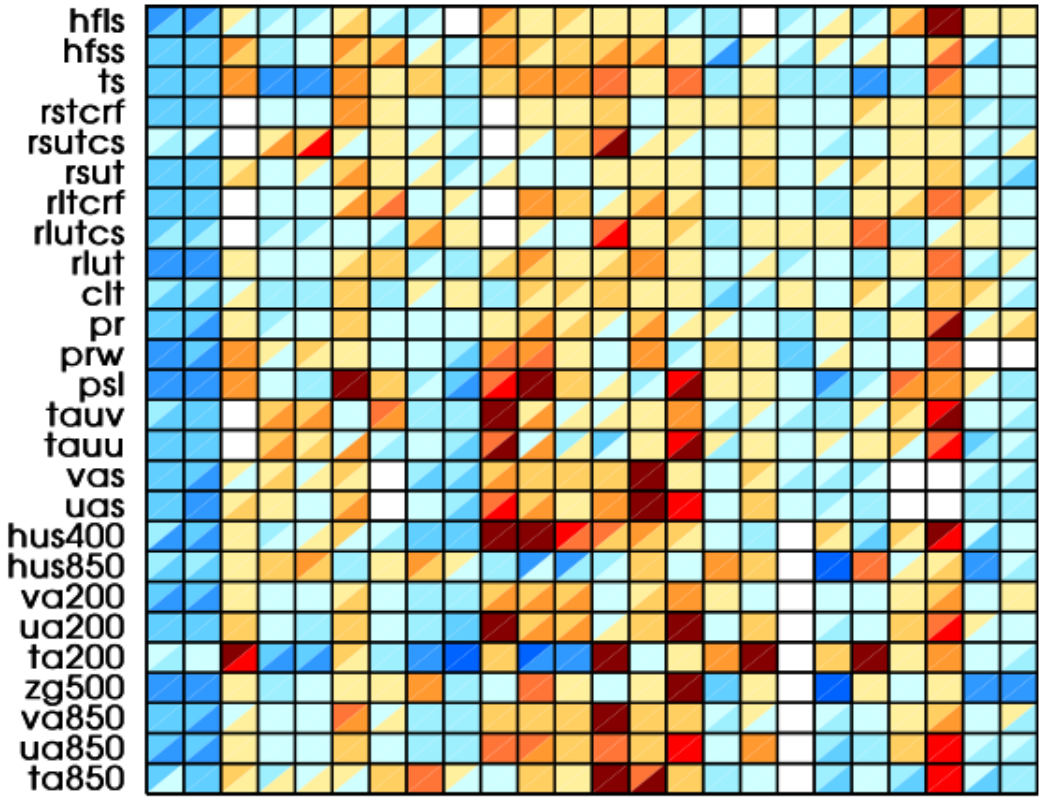

Figure 2.3: Relative performance of CMIP3 GCMs for 26 variables, including temperature, zonal wind speed, heat fluxes and radiative forcing. Split boxes indicate relative scores calculated from two separate observational datasets. Modified from Gleckler *et al.* (2008)

By taking into account a higher number of variables in their evaluation, Murphy *et al.* (2004) and Reichler and Kim (2008) were able to take a wider ranging approach in their performance metric formulation. However, as noted by the latter study, sampling first order moments of variability may not be the only aspect of climate simulation relevant for the evaluation of model skill. Gleckler *et al.* (2008) acknowledged this issue in their study evaluating not only the mean state of CMIP3 ensemble members over 26 variables independently but also considered the relationship and possibility for metric redundancy between them. Assessing a wide

range of variables and providing a simple, qualitative visual reference was shown to be one of the key benefits of metric evaluations (Figure 2.3). The presence of redundant metrics entails the 'double counting' of information which had already been obtained by another metric, which (aside from practical considerations of the efficiency of model assessment) may undermine the goal of producing a balanced assessment of model performance, without over emphasising certain characteristics of simulations. Additionally, the higher variability moments were assessed separately with a 'Model Variability Index' (MVI)

$$MVI_{mr} = \sum_{f=1}^{F} \left(\beta_{mrf} - \frac{1}{\beta_{mrf}} \right)^2 \tag{2.2.13}$$

where m, f and r correspond to the model, variable and observational dataset respectively. β^2_{mrf} is defined as the ratio $\frac{v_s}{v_o}$, where v_s and v_o are the simulated and observed variance for each the spatial average timeseries and variable respectively. Thus $MVI_{mr} > 0$ with values closer to 0 signifying closer agreement of the variance of model m with the observed dataset r. They then investigated to what extent a relationship exists between the model mean state error and MVI_{mr} variability metrics and found mixed results. In the tropics they found very little correlation between the two measures whereas in northern extra-tropical regions, although it appears to be weak, they argue there is some relationship. In concluding, it was cautioned that metrics based on mean climate statistics alone risk giving incomplete and deficient assessments of overall model performance.

The approach of Xu *et al.* (2010) in utilising several variables and moments of variability to assess GCM performance is similar to several other metric proposals that focus specifically on evaluating RCMs. Sánchez *et al.* (2009) formulated a metric combination of five sub-components each assessing the reproduction of precipitation cumulative distribution functions for model m and observations o,

$$f_1 = 1 - \sqrt{\frac{|A_m - A_o|}{2A_o}} \tag{2.2.14}$$

$$f_2 = 1 - \sqrt{\frac{|A^+_m - A^+_o|}{2A^+_o}} \tag{2.2.15}$$

$$f_3 = 1 - \sqrt{\frac{|A^-_m - A^-_o|}{2A^-_o}} \tag{2.2.16}$$

$$f_4 = 1 - \sqrt{\frac{|\overline{P_m} - \overline{P_o}|}{2\overline{P_o}}} \tag{2.2.17}$$

$$f_5 = 1 - \sqrt{\frac{|\sigma_m - \sigma_o|}{2\sigma_o}} \tag{2.2.18}$$

where A represents the total area below the CDF curve for models (RCM) and observations (CRU) respectively, A^- the area below the 50$^{\text{th}}$ percentile, and A^+ the area above the 50$^{\text{th}}$ percentile. These CDFs are produced from regional spatial average timeseries. $\overline{P}_*$ is defined as the spatio-temporal average and σ_* the PDF standard deviation. Coppola *et al.* (2010) proposed a combined metric evaluating the mesoscale signal of both precipitation and temperature signals concurrently in European RCM simulations with five sub-metrics. These cover spatial correlation patterns (g_1, g_2), normalised RMSE (g_3, g_4) and inter-correlations between precipitation and temperature patterns (g_5).

$$g_1 = R(P_m, P_o) \tag{2.2.19}$$

$$g_2 = R(T_m, T_o) \tag{2.2.20}$$

$$g_3 = \frac{\sigma(P_o)}{\text{RMSE}(P_m)} \tag{2.2.21}$$

$$g_4 = \frac{\sigma(T_o)}{\text{RMSE}(T_m)} \tag{2.2.22}$$

$$g_5 = 1 - \frac{|R(P_o, T_o) - R(P_m, T_m)|}{2} \tag{2.2.23}$$

Here, T and p are defined as the seasonal average mesoscale fields (RCM simulation - large scale $\sim$250km component) for temperature and precipitation respectively. $\sigma(x)$ is calculated by finding the interannual standard deviation of seasonal values, and then spatially averaging for a given region of interest. Following in this approach, Eum *et al.* (2012) incorporated additional aspects of climate simulation into their metric combination by including metrics assessing the reproduction of linear trends and extremes.

Only a limited sampling of the possibilities for combination methods has been attempted in the current literature. The predominant method involves taking the

product over all included sub-metrics, given by

$$\prod_{k=1}^{n} f_k^{\alpha_k} = f_1^{\alpha_1} \cdot f_2^{\alpha_2} \cdots f_n^{\alpha_n} \tag{2.2.24}$$

where the f_k are individual metrics and each α_k can be varied to 'weight' one or more f_k. This provides a demanding appraisal of skill by requiring that models perform well over all variables and parameters evaluated to merit a high score (Xu *et al.*, 2010). However, Eum *et al.* (2012) argue that this approach has the potential to give too much emphasis to particularly low scores, thereby producing a misleading assessment of overall model skill. They investigate by comparing the metric product combination approach with an additive method (the numerical average of all metric scores), which would not be as sensitive to anomalous sub-metric scores, but found that the former multiplicative approach when used in model ensemble weighting gives a lower error than the simple multimodel mean in the reproduction of mean climate. However, it is unclear to what extent this finding justifies the view that a metric product approach, over an additive approach say, is superior and gives a robust assessment of model performance.

Christensen *et al.* (2010) investigated the sensitivity of combination procedure by comparing three different methods, W_{PROD}, W_{RANK} and W_{REDU}. W_{PROD} is the standard approach given by (Equation 2.2.24) with $n = 5$ and $\alpha_k = 1$; W_{RANK} is a method which ranks models by their respective scores under each f_k and transforms this into individual scores based on each model's relative performance with other ensemble members; W_{REDU} takes the 'standard' method of W_{PROD} and transforms the output scores so that the ratio of best to worst model is altered by a factor of 1.2 by manipulating the values of α_k. It was found that W_{PROD} gives a higher variation in model scores but that the overall ranking of models is largely insensitive to the choice of combination approach within their proposed methods. They acknowledge that W_{RANK} and W_{REDU} do not provide unique model scores since they are calculated by relative performance differences within the ensemble i.e. inclusion or exclusion of a single model would change scores for the other ensemble members. Therefore, these approaches are of limited use for comparing inter-generation improvements in climate model performance, as they do not provide an absolute performance measure.

To analyse the uncertainty specifically within the multiplicative combination procedure Coppola *et al.* (2010) tested several alternative combinations of g_k, given

by:

a)	$g_1 \cdot g_2 \cdot g_3 \cdot g_4 \cdot g_5$
b)	$g_1 \cdot g_3$
c)	$g_2 \cdot g_4$
d)	$g_1 \cdot g_2 \cdot g_3 \cdot g_4$
e)	$(g_1)^2 \cdot (g_2)^2 \cdot g_3 \cdot g_4$

Table 2.1: Metric combination approaches proposed by Coppola *et al.* (2010). g_n represents individual metrics.

The standard method (Equation 2.2.24) is given by (a); (b) and (c) take into account either precipitation or temperature respectively; (d) eliminates the cross correlation sub-metric g_5; approach (e) gives higher emphasis to correlation sub-metrics g_1 and g_2 with increased values of α_1 and α_2. They found that the precipitation sub-metrics g_1 and g_3 had a greater influence on overall model scores due to the higher range of values produced. The effect of this on final scores was that the overall metric will be evaluating more in terms of precipitation performance than temperature; multiplicative combinations may place too much emphasis on sub-metrics which discriminate between models more than others. Under this approach, if two models A and B score similarly in one aspect, but A scores twice as high than B in another, the overall score of A will be double that of B under this framework. In addition, the results of Coppola *et al.* (2010) indicated that the performance of weighted ensemble means, when evaluated against observations, were for the most part insensitive to combination method. This offers support to the multiplicative approach due to the apparent robust assessments it provides with respect to choice of components g_k used.

The second fundamental source of subjectivity in constructing combined metrics lies with which sub-metrics to include or exclude. The early literature somewhat ignores this issue. The approach taken by both Murphy *et al.* (2004) and Reichler and Kim (2008) are comprehensive in the sense that they span a large range of variables. However, Gleckler *et al.* (2008) noted that when combining a variety of metrics it is uncertain to what extent they contribute independent assessments of model performance. This has the potential, if strong relationships are found, to bias the combined metric through a 'double-counting' of particular model characteristics.

To assess the similarity in metric results over a large range of variables, Gleckler *et al.* (2008) used two tests to investigate the level of metric redundancy i.e. the amount of repeated information given by different metrics. First, they proposed a simple test whereby model rankings under each of the 20 variables are compared for all combinations, amounting to 190 unique pairs. The idea being that the lower the change in average ranking, the more redundant information is present. It was found that the change in ranking over all metrics pairs and 24 models varied from 3 to 6, with a mode value of 4. The significance of this result is difficult to assess, however the average ranking value change is such that models would for the most part appear to be consistently performing across most variables. Models highly scoring in one variable tend to do so for other variables on average. This suggests a reasonable degree of metric redundancy, and the second test of Gleckler *et al.* (2008) may assist in identifying such relationships. This test involves calculating the correlations between model scores based on two different metrics. They found that in some cases high correlations exist, such as between outgoing longwave radiation and precipitation rates, and also cases where there is no perceived relationship. Waugh *et al.* (2008) analysed chemistry-climate atmospheric models with a process based metric over 16 variables. They too looked at the correlations between all 120 metric pairs. Their results showed that most pairs had weak to no correlation, but high correlations were present, mostly between metrics targeted at similar variables or processes. Correlation analyses such as these could be exploited as a basis for removing one of the variables from an overall set of metrics, although it would be for expert judgement to determine which variable to keep, since the statistic used does not provide any additional information to assist in this decision.

Pierce *et al.* (2009) recognised this redundancy issue and proposed an objective method to reduce the effective number of sub-metrics in combined metrics. They utilised Empirical Orthogonal Function (EOF) analysis to find the largest modes of variability within the metric space. Thus, by taking the first few components representing a majority of the variance between metric scores, a set of EOFs with which to assess model performance can be produced. They acknowledge a fundamental issue with this method is that an EOF analysis will find orthogonal basis functions which most optimally describe the variability of model scores, but will not necessarily produce EOFs which 'point' directly towards improved skill. The implication of this is that the EOF basis functions would not assist in identifying the metrics that are most important to characterise model performance, only those metrics which best describe the metric output data. Additionally, there is no

criterion for choosing how many components to take, although this overall method might be considered more objective than the arbitrary weighting of individual sub-metrics as proffered in studies mentioned above (see Equation 2.2.24).

Nishii *et al.* (2012) further investigated this matter by testing whether a chosen combined metric is sensitive to the redundant information with a variety of multi-variate analysis methods. The Climate Prediction Index (CPI) of Murphy *et al.* (2004) was used as the baseline metric, with the 24 models of the CMIP3 GCM ensemble being utilised. The CPI was calculated using 22 climatic variables and then decomposed using three methods in turn: Principle Component Analysis (PCA), Cluster Analysis and a Non-negative Matrix Factorisation (NMF) method of Lee and Seung (1999). It was found that the overall model performance index, when constructed using each of the stated methods, produces similar results when a large enough sample of variables (in this case 22) is considered. These conclusions are limited to mean climate statistics used in the CPI however, which as stated earlier are unlikely to be fully representative of the overall model performance. These results indicate that there may be a potential for reducing the number of metrics within a combination.

Overall, a wide range of metric combination approaches have been investigated, with some common construction approaches being shared among them. Multiplicative methods are widely adopted, with less attention being paid to alternatives; additive methods for example. A range of metrics which are included in combination spanning different variables and measures of variability is a typical feature, however the impact of different possible choices in this area has not been explored in detail. Finally, the level of metric redundancy within combination approaches has begun to be investigated, which could potentially lead to more objective methods. Metric Combination approaches as used to generate generalised performance indicators thus far have yet to reach any standard methodological basis, nevertheless they are one of the most promising avenues by which to reach such comprehensive model performance assessments.

2.3 Constructing Climate Change Projections: Approaches and Assumptions

With the emergence of large coordinated global and regional multi-model ensemble projects (Meehl *et al.*, 2007; Taylor *et al.*, 2012; Christensen *et al.*, 2002; Van der Linden and Mitchell, 2009; Giorgi *et al.*, 2009) the development of approaches to construct climate change projections has become an area of active research (Tebaldi and Knutti, 2007). More specifically, consideration has been given to methods aimed at producing, characterising and potentially reducing the uncertainty in projections (Daron and Stainforth, 2013), which are of vital importance to provide robust climate change information both to impacts, adaptation and vulnerability communities and decision makers. There are several approaches commonly used to go about these different tasks: the use of simple "one-model-one-vote" (Knutti, 2010) constructions of multimodel means and ensemble simulation spreads, probabilistic statistical methods for quantifying uncertainty and a range of "post-processing" bias correction techniques for 'correcting' model systematic errors for impacts applications (Dosio *et al.*, 2012). A common theme throughout is the concept of stationarity; that model characteristics identified in historical simulations will continue to hold into future projections. Based on this assumption model historical-projection simulation differences are considered reliable estimates of the climate change signal (Bellprat *et al.*, 2013), model ensemble projections are weighted based on historical performance and model biases are removed in projections. To ensure the robustness of these applications and the findings generated the assumption of stationarity should be considered in more detail.

2.3.1 Standard Projection Constructions

The current basic toolbox for producing information on climate change projections constitutes a range of approaches: multimodel mean projections as a 'best guess' of the future climatic state, the spread of ensemble simulations as an indicator of projection uncertainty and assessments of model consensus as a measure of projection robustness. Each of these approaches, although commonplace, have weaknesses or assumptions not always acknowledged which may have an effect on how robust conclusions drawn from projections can be.

The rationale for using multimodel means (MMM) for ensemble projections in large part is based on the the frequent finding that the MMM often has a lower error

relative to observations than the best ensemble member in historical simulations (e.g. Hagedorn *et al.*, 2005; Gleckler *et al.*, 2008; Reichler and Kim, 2008; Pierce *et al.*, 2009). The common explanation for this phenomenon is the cancellation of errors (Weigel *et al.*, 2008) which would occur with a set models which are assumed to give largely independent results. However Annan and Hargreaves (2011) suggest that the reason may be due to the statistical characteristics of multimodel ensembles. An over dispersive ensemble, in which models are sampled from a wider distribution than the observations (i.e. observations rarely exhibiting behaviour that the model ensemble does not capture), will give a situation where the MMM is much more likely to be closer (lower overall error) to the observations than would be expected, meaning that this apparent 'skill' of the MMM may be somewhat unremarkable. Therefore whether the projected MMM can be relied upon as being in some sense 'close to the truth' in future projections is uncertain. Furthermore, the interpretation of the projection spread of multimodel ensembles (e.g. Knutti and Sedláček, 2013) is limited by the fact that they do not correspond in any statistical sense to likely 'true' uncertainty range (Collins *et al.*, 2013). Indeed, information as to the skill of individual models is ignored in these constructions, potentially leaving in unrealistic outlier models which may exaggerate the 'likely' projection range if model biases grow under climate change (Christensen and Boberg, 2012). Such issues have been demonstrated to be an issue for many regions around the globe (Figure 2.4), where systematic temperature dependent model biases are found. Finally, methods to measure the apparent consensus of ensemble simulations in certain aspects such as the future sign of change in precipitation levels often assume the independence of systematic error biases between ensemble members (Power *et al.*, 2012), which has in recent times been raised as a questionable premise (Abramowitz and Gupta, 2008).

To improve on these standard approaches to producing future climate change projections, utilising information gained about model characteristics and performance in historical runs to construct projections has been suggested. Climate model skill is determined from the ability to replicate past observations over a range of climatic variables (Tebaldi and Knutti, 2007). As Räisänen and Ylhäisi (2011) note, it is uncertain to what extent present and past skill can be relied upon as indicators of future performance. Xu *et al.* (2010) agree that due to the nature of climate model development, 'good' performance should be considered at present a necessary but insufficient condition for future reliability. Oreskes *et al.* (1994) discussed the issue of the non-uniqueness of model solutions to observational

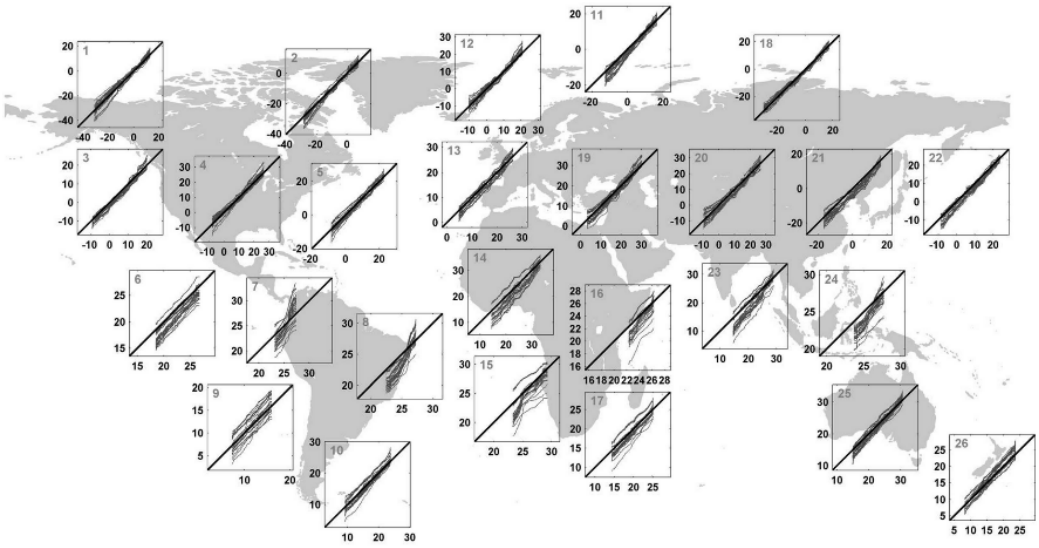

Figure 2.4: CMIP5 GCM bias analysis for various regions; horizontal axes are observations, vertical GCM temperatures. GCM biases are seen to be strongly dependent on temperature in most cases, suggesting that mean bias correction methods may be unable to remedy such discrepancies. Modified from Christensen and Boberg (2012).

data, which implies that one can only conclude that a model is consistent with the observations, and not infer from this that it correctly and completely describes the dynamical system. This difficulty is emphasised due to the clear lack of alternative validation methods for long-term multi-decadal projections, which thus demands that skill be defined for the most part on the reproduction of past observations and important climatic processes alone (Knutti, 2008).

One difficulty inherent in evaluating climate models based on their replication of past observations is highlighted by Parker (2011) who notes that the construction and development of GCMs, and to a lesser extent RCMs, requires a degree of model calibration, or 'tuning', to form simulations that are consistent with known fundamental climate aspects (Flato *et al.*, 2013). One common example of this is to correct for TOA radiative errors by varying cloud parametrisations (Mauritsen *et al.*, 2012). If a model's parameterisations are altered in this way to improve simulations relative to a set of chosen observed characteristics, it is important to know what specific aspects have been tuned for, since evaluating a model for a variable which has been tuned for risks a degree of circular logic. Guilyardi *et al.* (2013) discuss the recent improvement in transparency towards this and more general forms of climate model 'metadata'; technical details relating to model structural developments, numerical approximation schemes, parametrisation choices and experiment plans. They note that although much information of this type has been released for

the latest CMIP5 ensemble, details regarding the model calibration stages remain undocumented. However, even if total knowledge behind model developments and calibration datasets were in the public domain, an understanding of how such information could be utilised in model evaluations is currently lacking (Abramowitz and Gupta, 2008). Such information could be relevant in choosing independent datasets to those used in climate model development stages, although it may be impossible to completely avoid the overall issue.

2.3.2 Probabilistic Climate Change Projections

Efforts to reduce uncertainty with probabilistic projections have taken either a Bayesian or frequentist statistical perspective in recent times (Räisänen and Ylhäisi, 2011). Both approaches have a degree of subjectivity in their application; the Bayesian perspective requires the construction of prior distributions from which the final 'posterior' distributions are built (Tebaldi and Knutti, 2007), whereas the frequentist model ensemble weighting methods necessitate the selection of a number of performance metrics to provide the weights. Tebaldi *et al.* (2005) suggested a Bayesian framework to incorporate known model performance and observational information to producing probability density functions (PDFs) of regional changes in temperature with GCMs. They showed that probabilistic projections could be constructed whilst including different assumptions concerning the characteristics of the model ensemble and model performance weightings, in this case by the factors of ensemble convergence and mean bias in line with the REA method of Giorgi and Mearns (2002). Furrer *et al.* (2007) produced temperature and precipitation probabilistic projections on a grid point level, developing from the larger regional sized projections previously considered. The single variable approach was generalised to include multiple variables by Smith *et al.* (2009), taking into account results from 22 regions at once rather than separately. Kang *et al.* (2012) produced projections of northern winter temperatures assessed from NARCCAP RCM simulations within a Bayesian framework. Although becoming more widespread, the choice of prior distributions has come under criticism by Knutti *et al.* (2010), who argue that because of the somewhat random sampling for multimodel ensembles, in contrast to that of perturbed physics ensembles, the choice of an appropriate prior is complicated and may not be justifiable. The stationarity assumption also arises where any model performance information is included to construct the posterior distributions, such as that used in Tebaldi *et al.* (2005), since it is assumed that this performance information is time invariant. Buser *et al.* (2009) suggest an approach

to account for such time varying model biases, by incorporating this specific aspect into their hierarchical Bayesian framework. They found that RCM Alpine seasonal temperature projections were sensitive to this assumption, for both a "constant bias" and "constant relation" bias behaviour, and recommend further research in this area.

The frequentist approach to constructing climate change projections involves the application of weights to individual models within an ensemble, on the basis of previous assessed performance under a given metric or set of metrics. This weight could be effectively a zero weight; in practice removing a model from the constructed projection. Elimination of weaker performing ensemble members based on predefined benchmarks may provide a practical way of accounting for demonstrably unrealistic simulations. One example of this method is in producing estimates for when the Arctic sea will be ice free in summer (Stroeve *et al.*, 2007; Wang and Overland, 2009; Zhang, 2010). However, Mahlstein and Knutti (2012) criticise the subjective nature of such 'model elimination' approaches, since as they point out criteria could be generated to eliminate any model if desired. Furthermore, the restriction of multi-model ensembles to small 'suitable' subgroups also ignores the issue of model independence, since by only using models passing imposed benchmarks may lead to an artificial narrowing of projections leading to overconfidence (Knutti *et al.*, 2010).

One problem in determining whether to weight model ensembles is that there is no consensus on the best way of doing so (Weigel *et al.*, 2010). Although multimodel means have been shown in several cases to produce more lower error simulation fields than the best ensemble member (Tebaldi and Knutti, 2007), Doblas-Reyes *et al.* (2005) argue that this potentially can be improved upon given that the assumptions of uniform model skill and independence are rarely applicable in current ensembles. Thus it is suggested that performance related weighting may be introduced to produce more robust projections, but this may simply be adding a new level of uncertainty (Christensen *et al.*, 2010). Utilising weighting metrics based on model performance has been shown to be an improvement over equal weighting in short-term seasonal and weather forecasting (Hagedorn *et al.*, 2005). However, this does not guarantee that the method will be applicable to long-term decadal projections as there are several key differences in the two applications. Climate projections do not have data to verify the simulations or the application of weighting schemes, unlike for short-term forecasts. Additionally, the structure (issues of for example, model independence) of multimodel ensembles used in

short-term and weather forecasting are better understood than those 'ensembles of opportunity' generated for climate simulations.

The main assumption of model ensemble weighting is that model skill assessed by a performance metric in historical simulations will be stationary into future projections (Räisänen and Ylhäisi, 2011). Put simply, stationarity implies that 'good' performance in the past, however defined, justifies an emphasis on that model in future relative to weaker models. Whetton *et al.* (2007) analysed the relationship between spatial patterns of past mean climatologies and the corresponding future changes with 17 CMIP3 GCMs. They utilised a pseudo-reality method in which the seasonal 'M' metric of Watterson (1996) is calculated between all 136 pairs of GCMs for three variables; mean temperature, sea level pressure and precipitation. They found that in higher latitudes a notable relationship between those models simulating similarly in historical runs to the future changes in temperature, with weaker relationships found for the other variables. In the tropics with the variables assessed only weak connections were present. It is suggested that such an approach could be used to inform ensemble weighting approaches. Abe *et al.* (2009) similarly analysed the pseudo-reality correspondence between GCM pairs of past and future spatial patterns for CMIP3. They included three separate metrics; the similar 'M' statistic, the Climate Predication Index of Murphy *et al.* (2004) and a centred spatial correlation measure, R for temperature, sea-level pressure and precipitation. The metrics were applied to each pair combination of GCM for historical (1981-2000) and future projection (2081-2100) periods, and correlation coefficients were produced to measure the similarity of simulation. These assessments were made for three variables; surface air temperature, sea level pressure and precipitation, and the analysis was done seasonally over global and four regional domains. Their results concur with those of Whetton *et al.* (2007) in that they found some moderately significant correlations for temperature in high latitudes, but decreasing correlations in lower latitudes. Precipitation correlations were generally less reliable than for temperature, although over the tropics stronger relationships were found which they suggest may have some use for weighting projections. In concluding, they find a weak case for the use of such spatial correlations for multimodel ensemble weighting due to the generally low level of correlation between past and future assessments. The main conclusion from this is that alternative metrics should be investigated to determine if stronger historical-future projection relationships can be identified, with which the stationarity assumption would hold.

2.3.3 The Stationarity Assumption

The reliability of regional and global climate model future projections is often inferred from their respective performance in replicating observed historical conditions (Tebaldi and Knutti, 2007). This assessment involves a quantitative or qualitative appraisal of model errors over a range of variables, which in turn can provide information on whether or not various simulated climate processes are sufficiently realistic. Evaluation of the behaviour of these systematic model errors beyond the timeframe of observations, on the other hand, is much more problematic due to the lack of verification data. Understanding whether these model characteristics change relative to the unknowable future 'truth' is crucial for the robustness of a number of applications: statistical downscaling, bias correction and ensemble projection weighting. The basic reason is that these methods all assume that historical bias characteristics remain constant over time, or time-invariant. The second problem is that the scientific basis for this assumption may be questionable since emergent model properties and behaviour, such as land surface characteristics, which are the cause of model bias are also implicitly considered to be time-invariant (Maraun, 2012).

The use of bias correction to reduce the systematic errors of regional and global climate model data for use in impacts studies is becoming more commonplace (Vannitsem, 2011). Such 'post-processing' of spatio-temporal data has been seen, particularly for users of downscaled climate information, as desirable due to its simplicity of use (Johnson and Sharma, 2012). A variety of methods have been proposed, from simple additive methods (e.g. Déqué, 2007), linear regression techniques (e.g. Hay and Clark, 2003) or more complex CDF matching (e.g. Piani *et al.*, 2010; Heinrich and Gobiet, 2012); and gamma-gamma transformation approaches (e.g. Sharma *et al.*, 2007). However, such methods have been questioned due in part to their lack of physical rationale (Ehret *et al.*, 2012) and more fundamentally their assumption that the correction function will remain stationary or time invariant, that is model biases remaining stationary from historical simulations to future projections (Teutschbein and Seibert, 2012). Since the potential rate of change of the future temperatures is simulated to be higher than in the observed past, it is questionable whether this assumption holds. Indeed, long-term projections may become "skewed" due to inaccuracies in the model processes or parametrisations, such as carbon-cycle feedbacks (Frame *et al.*, 2007).

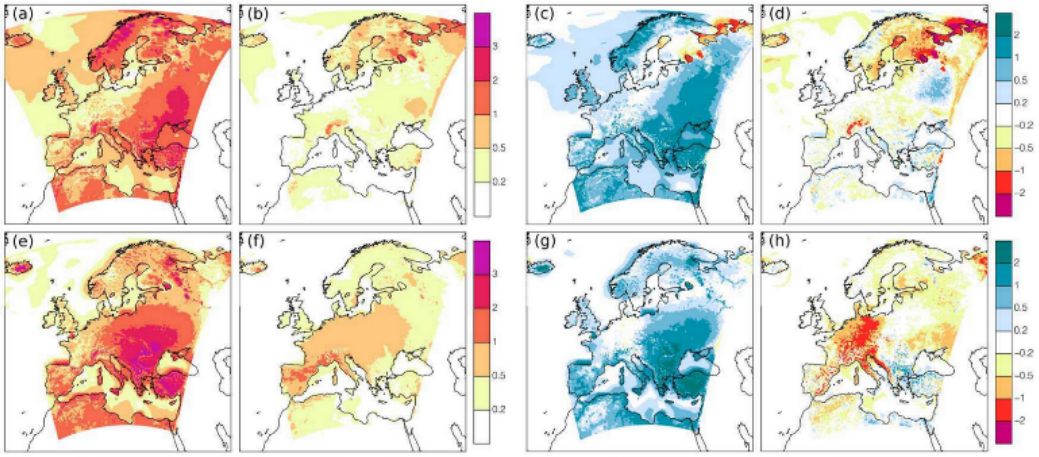

Figure 2.5: Seasonal (DJF top row, JJA bottom row) mean temperature bias and bias corrections assessed and applied to ENSEMBLES RCMs. First column indicates average 1970-1999 mean bias across all RCMs, second column average change from 1970-1999 to 2070-2099, third column average projection 'error' reduction, fourth column minimum projection 'error' reduction. Modified from Maraun (2012).

Maraun (2012) investigated the stationarity of mean seasonal temperature and precipitation biases in future projections in RCMs, using models from the ENSEMBLES project. They utilised a pseudo-reality method in which the unknowable future real world observations or 'truth' is replaced by members from the GCM/RCM ensemble in turn in future projections. Historical RCM biases were assessed from simulations forced by reanalysis data. Using this approach it is possible to assess the potential for such biases to change in character over long multi-decadal timescales, given the premise that each model GCM/RCM combination will behave similarly to, and thus be exchangeable with, the 'true' climate. They found that the relative model biases for both mean temperature and precipitation were generally consistent from 1970-1999 to 2070-2099, as can be seen in Figure 2.5. This bias change was found to be particularly strong in regions such as the white and baltic seas where sea ice is common. These results contrast with the findings of Christensen *et al.* (2008) who evaluated reanalysis forced RCMs over Europe and suggested a potential for biases in temperature and precipitation to be amplified with increases in temperature. This finding was echoed by Christensen and Boberg (2012) who assessed CMIP5 GCMs for their regional warming response and found that models with a warm bias tended to project warmer temperatures, although there was some regional variation in this behaviour. They suggest that this general characteristic of GCMs should warn against an overdependence on interpreting the spread of future projected temperature changes without taking such non-stationary biases into account. Although the results of Maraun (2012) suggest that bias

correction may be justified for mean seasonal errors, the conclusions are based in pseudo-reality meaning that the evaluated bias changes are only relative to other potentially similarly constructed models. Christensen *et al.* (2008) and Christensen and Boberg (2012) provide reasons to question whether the behaviour of climate model ensembles for such a pseudo-reality application can be relied upon to give a definitive answer.

2.4 Implications

This literature review covers three general themes: performance metrics and their underlying sensitivity to changes in methodology, approaches to combine metrics for more general overviews of model performance and the use of model performance information to construct future climate change projections. The merits of considering these issues is to identify where the use of performance metrics in assessing RCMs may be improved, either in their inherent robustness and redundancy, or in the application of model assessments either in combination or to improve understanding of future climate change projections. Several findings are apparent from the work that has occurred so far in these areas.

First, the evaluation of RCMs with performance metrics has yet to be done within a fully comprehensive framework, assessing a range of variables, over several spatial and temporal domains, with a range of statistics and observational datasets. Although these tasks may have been done in isolation, for an assessment of the sensitivity of performance metrics to these differences in methodology it is required that a wider assessment should be done to fulfil this goal. In particular, an effort to evaluate models in their temporal behaviour and performance in relation to extreme events, going further than mean climatological evaluation, would be of benefit to produce a more thorough assessment of overall performance. Spatial variation in RCM performance is of importance, particularly to the climate change impacts community, who may wish to select certain models, and require robust RCM evaluation frameworks for doing so.

Second, there are several possible approaches to combine performance metric output to produce a single quantitative measure of overall performance (GPI). The literature generally shares similar methodologies in this regard, and in particular to RCM assessment do not incorporate a large number of variables or

statistics. Two issues arise in this case. Investigation into both the number and type of performance metric to include and the sensitivity of GPI output based on these decisions. Additionally, the construction approach itself is lacking in scrutiny, as a similar method is used across all studies. The ability to rely on a robust, comprehensive and as far as is possible objective overall evaluation framework for RCMs will be of benefit both to assess progress in simulation quality and to provide a coherent, simplified digest for users of climate change information.

Third, the use of performance metrics in approaches to constrain multimodel ensemble climate change projections relies on a stationarity assumption whereby model performance assessments are assumed to remain constant into future projections. This is of relevance to both ensemble weighting methods and bias correction applications. Little has been done as far as RCM dynamical downscaling on this issue and is the basis for some methods attempting to go beyond the standard equal model approach. Assessment of this assumption, be it only in an indirect manner, will be of use to test the justification for the use of these methods in applications of model performance evaluations.

Chapter 3

Regional Climate Model and Observational Data

3.1 Overview of RCM projects and Observational Data Options

Coordinated regional downscaling programmes are a relatively recent development, following on from the precedent set in the GCM community with the AMIP, CMIP and the later CMIP3 and CMIP5 projects (Gates, 1992; Meehl *et al.*, 2000, 2007; Taylor *et al.*, 2012). These earlier GCM efforts established a framework in regard to coordinating modelling groups and their experiments, thereby ensuring analogous simulation protocols. By employing this collaborative scheme, global models can be evaluated alongside one another in a more objective and standardised manner. The RCM modelling community has been developing on an analogous path, producing multi-model ensembles covering a variety of domains (see Table 3.1). The Co-ordinated Regional Downscaling Experiment (CORDEX) belongs to the latest generation of these projects, although it could be considered distinctive due to its far larger scope. It encompasses thirteen, sometimes overlapping, domains spanning the majority of the Earth's land surface: Africa, the Arctic and Antarctic, Europe, the Mediterranean, Northern, Central and South America, the Middle East and Northern Africa, South, Central and East Asia and Australasia. Most simulations are run at 0.44° (50km grids) resolution, with higher 0.11° (12.5km grids) resolution simulations additionally produced for Europe. These more spatially detailed historical and future projection runs enable regional assessments as to the likely impacts with higher spatial fidelity than previous downscaling projects (Jacob *et al.*, 2014). Unfortunately, due to time constraints, RCM data from CORDEX is

not used for the first or second analysis, however the third analysis utilises a subset of historical and future projection runs from dataset. At the time of writing most of the simulations for CORDEX have been completed and are available online for download.

When selecting an RCM ensemble for a particular analysis, a number of factors should be taken into consideration, such as:

- Length and period of simulation timeframe
- Availability of high quality observational datasets
- Ensemble size
- Type of lateral boundary condition forcing implemented
- Spatial and timeframe (e.g. daily/monthly means) resolution of output

Naturally, all of the mentioned past and current generation RCM projects vary in these aspects, and for different scientific questions or applications one may prefer one RCM ensemble over another. For the purposes of investigating the use of performance metrics with RCMs however, most important is the requirement of high quality observations with which to evaluate RCM simulation quality. If observational datasets are used that are of lower resolution, or of fewer underlying station number, then the benefits of using high resolution dynamical downscaling with RCMs are not as well exploited. Lower resolution observations would not be of use in determining model performance over complex orography for example. Secondly, those ensembles utilising reanalysis forcing are able to provide simulations without the potential influence of GCM boundary condition biases (with the generally reasonable assumption that reanalysis forcing biases are negligible) thus enabling as far as it is feasible the evaluation of the RCM simulation alone. Thirdly, long multi-decadal simulation runs will be preferred as to ensure robust climate statistics, especially in the case of extreme indices. Fourthly, ideally the size of the ensemble should be large (this study uses nine RCMs), ensuring that intermodel comparisons of absolute and relative performance are meaningful. Finally, high spatial resolution output will preferred, to make sure that the analysis findings are relevant to studies utilising more detailed model output.

Of the available project data, the European ENSEMBLES dataset meets these requirements best. Unlike some earlier projects which rely on GCM forced runs

Region	Available RCM Simulations	Observational Datasets
Africa	2009 ENSEMBLES: ERA-Interim reanalysis driven RCM runs covering 1989-2007 over west Africa at 50km resolution. 3 GCMs (HadCM3Q0, ECHAM5-r3 and ECHAM5) used to drive 11 separate RCMs covering roughly 1950-2050 at 50km resolution.	TARCAT v2.0 decadal, monthly and seasonal rainfall estimates from satellite data. 10 year climatologies and anomalies against the 10 year climatology covering 1983-2012.
Europe	2005 PRUDENCE: 50km GCM forced runs covering 1961-1990 and 2071-2090, at mostly 50km resolution. 2009 ENSEMBLES: 25km and 50km ERA-40 forced 1961-2000 simulations and CMIP3 GCM 1950-2050+ projections	25km and 50km E-OBS gridded tas, tasmax, tasmin, pr, slp data
North America	2007 NARCCAP: NCEP/DOE II Reanalysis driven runs at 50km resolution covering 1979-2004. GCM nested RCM runs at 50km resolution covering 1971-2000 and 2041-2070. GCMs forced with A2 emissions scenario	JISAO Precipitation Grids available in 0.5x0.5 and 1.0x1.0 degree resolution. Covers 1900-2008
South America	2008 CLARIS-LPB runs: 50km resolution ERA-Interim driven RCM runs for 1990-2008. CMIP5 GCM driven RCM runs for 1960-1990, 2010-2040 and 2070-2100. A1B emissions scenario utilised.	0.5x0.5 gridded minimum and maximum temperature for 1961-2000
East Asia	2005 RMIP Project: 1989-1998 decadal RCM simulation (60km resolution), and future climate change scenario both forced by a GCM	Station and gridded temperature and precipitation observations

Table 3.1: RCM Projects by domain

alone (e.g. PRUDENCE), ENSEMBLES includes ERA-40 reanalysis forced historical simulations covering 1958-2001. This is of benefit when investigating the skill of the RCMs, and not so much the lateral boundary condition driving component. Furthermore, the reanalysis forced historical runs and in particular the GCM forced projections are longer in their timeframe duration and higher in spatial resolution than pursued in previous downscaling projects. The ENSEMBLES (and CORDEX) GCM forced simulations are transient (i.e. continuous), whereas in PRUDENCE a time-slice approach was used whereby two 30-year timeframes (1970-1999 and 2070-2099) are used (Christensen *et al.*, 2002). Time-slice simulations although less computationally expensive do not give information on the decades immediately following the historical time period which can used to assess impacts for example on water systems (Burton *et al.*, 2010). Most importantly, the high resolution E-OBS dataset (Haylock *et al.*, 2008; Van den Besselaar *et al.*, 2011) was assembled specifically for comparison with those RCMs produced, and is output on identical grids to the models. This allows model-observation comparison without the concern of regridding observational datasets with unknown interpolation error introduced. This contrasts with other simulated regions such as Africa (e.g. Kim *et al.*, 2014), where the quality and spatial coverage of observations is comparably low, and the use of re-interpolated observations or non-station based reanalysis has to be considered, such as the TRMM satellite product (Huffman *et al.*, 2007) as a proxy for precipitation over the tropics. Overall, the ENSEMBLES dataset currently remains the best option for investigating performance metrics with RCMs, when model/observational comparison is required. For consideration of the stationarity of metric assessments however, since the method does not use observations (obviously none are available for validation of future projections) the high resolution CORDEX historical and future projection simulations are the best current option for investigations in this area.

3.2 ENSEMBLES RCM Data

The RCM data which is utilised in the following analysis is taken from the European ENSEMBLES project. The aims of the project were threefold: to develop dynamical downscaling methods within a multi-model ensemble framework, to improve our simulation and understanding of climate feedbacks and to enhance networking between the RCM community and users of such downscaled information (Van der Linden and Mitchell, 2009). Several regional modelling centres from

across Europe, in addition to one Canadian group, contributed to this 'ensemble of opportunity', producing two types of simulation; ERA-40 reanalysis forced historical and GCM forced future projections. The purpose of the reanalysis forced runs was predominantly to provide simulations with 'perfect' boundary condition, giving an opportunity for RCM validation exercises with as little external bias introduced as possible. The GCM forced simulations on the other hand enabled the construction of long time scale future projections across Europe for use in the vulnerability, impacts and adaptation user communities.

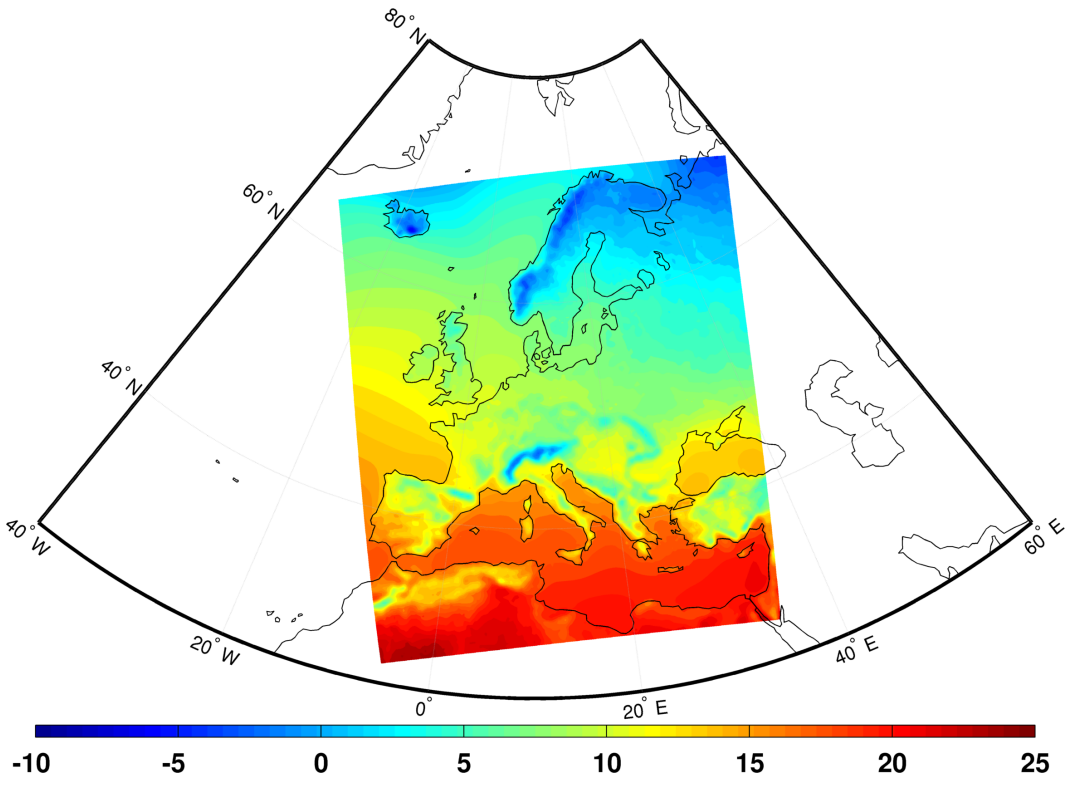

Figure 3.1: Standard minimum domain rotated-pole projection for ENSEMBLES RCMs. Displayed is the 2-metre temperature (°C) annual mean climatology for KNMI-RACMO2 covering 1961-2000

The experimental setup for ENSEMBLES required the use of a standard spatial domain and timeframe for both simulation types. Two different resolutions of RCM were supplied by modeling groups for ENSEMBLES; 0.22° (approx. 25km) and 0.44 (approx. 50km) data. For the higher resolution RCM output, which is utilised in the analyses, models are expected to cover an equal area rotated-pole projection with a minimum 190x170 latitude-longitude grid mesh. This region covers Morocco and Northern Algeria in the South-West, Turkey and Cyprus in the

South-East, Finland and Eastern Bloc nations in the North-East and Iceland in the North-West (Figure 3.1). Most models exceed this area in simulation domain (see Table 3.2 for details for each RCM) with a 'buffer zone' of 8 or more gridboxes in the surrounding region to allow the RCM to incorporate the lateral boundary conditions adequately and produce a consistent internal mesoscale simulation (Liang *et al.*, 2001). The models are constructed with different, although in some cases closely related, physics packages (several use modified versions of ECHAM-4 or ECMWF), various Soil-Vegetation-Atmosphere (SVAT) model components, orography physiographical datasets, surface characteristics, solar constant, GHG and aerosol concentrations. The RCMs are configured on spatial domains of different sizes and internal time step lengths. The time domains covered by historical and future projection runs for ENSEMBLES are different; reanalysis forced runs naturally correspond closely to the time domain of ERA-40 (September 1957 to August 2002), with models covering at least the 40-year period 1961-2000. The GCM forced projections on the other hand cover at minimum 1950-2050 with some RCM-GCM combinations reaching 2100. RCM output covers a large set of climatic variables (80+) over various timescales (daily/monthly) depending on the variable type. Of the RCMs submitted to the ENSEMBLES database only those whose output is given on the same rotated-pole grid, identical to E-OBS, are included in the two analyses in Chapters 4 and 5. This is due to two reasons: the fact that a regridding process would contribute some interpolation error which could adversely affect results and also computational difficulties in regridding the curvilinear grids on which the RCMs are based. Notwithstanding this drawback however, the size of the ensemble used is considered large enough for the purposes of the analysis.

Table 3.2: ENSEMBLES RCM details for those used in Chapter 4 and 5 analysis on Metric Sensitivity and Metric Combinations

Institute	**RCM**	**Version**	**Grid Dimensions**	**Vertical Levels**	**Time Step Duration**	**Physics Package**	**Model Components**
C4I (Community Climate Change Consortium for Ireland)	RCA	3	206x206	31	30 minutes	Semi-implicit, semi-lagrangian core	SVAT models: no name; developed at Rossby Centre.
DMI (Danish Meteorological Institute)	HIRHAM	2	213x198	31	5 minutes	Semi lagrangian dynamics coupled with physics of ECHAM4	Vegetation effects and run-off scheme (Dimenil and Todini 1992). Variables for maximum soil water holding capacity of the soil (Clausen et al. 1994; Roeckner et al. 1996).
ETHZ (Swiss Federal Institute of Technology in Zurich)	CLM	2.4.6	230x210	32	6 minutes	CLM is based on the COSMO weather forecast model.	SVAT model:BATS, soil model:TERRA3D
KNMI (Royal Netherlands Meteorological Institute)	RACMO2	2.1	206x224	40	15 minutes	Based on ECMWF model cycle 23 release 4 (similar to that used in ERA40). Updated by KNMI.	SVAT model:TESSEL

Continued on next page

Table 3.2 – *Continued from previous page*

Institute	RCM	Version	Grid Dimensions	Vertical Levels	Time Step Duration	Physics Package	Model Components
METO-HC (Met Office Hadley Centre)	HadRM3Q0	n/a	214x220	19	5 minutes	HadRM3 is based on a modified version of the HadCM3 atmospheric component.	MOSES land surface scheme, US Navy 10° orography dataset, solar constant=1365W/m^2
METNO (Norwegian Meteorological Institute)	HIRHAM	2	213x198	31	3.75 minutes	Semi lagrangian dynamics coupled with physics of ECHAM4	Vegetation effects and run-off scheme (Dimenil and Todini 1992). Variables for maximum soil water holding capacity of the soil (Clausen et al. 1994; Roeckner et al. 1996).
MPI (Max Planck Institute)	REMO	5.7	218x242	27	4 minutes	REMO is based on the ECHAM-4 GCM physics with the Europa-Modell NWP model as its dynamical core.	Land surface scheme extension of ECHAM4 parametrisations, REMO contains MPI ocean model, the Hydrological Discharge Model (HD Model) and the REMO atmosphere model.
RPN (Recherche en Prévision Numérique)	GEMLAM	1	n/a	56	12 minutes	Semi-lagrangian, semi-implicit numerical scheme, CMC/RPN physics	SVAT model: ISBA

Continued on next page

Table 3.2 – *Continued from previous page*

Institute	RCM	Version	Grid Dimensions	Vertical Levels	Time Step Duration	Physics Package	Model Components
SMHI (Swedish Meteorological Hydrological Institute)	RCA3.0	3	204x222	24	15 minutes	Semi-implicit, semi-lagrangian core	SVAT model: RCA land surface model, updated from that used in RCA2.0

The ENSEMBLES RCMs in their representation of summer (JJA) and winter (DJF) mean temperature climatologies (see Figures 3.2, 3.3, 3.4) produce similar bias characteristics. Seven out of the nine RCMs in summer months have a systematic 2-4°C warm bias in the Mediterranean and Balkan regions, spreading to central Europe in some cases, most notably RPN-GEMLAM. A majority of simulations in northern Scandinavian regions have a systematic 0 to -3°C cold bias. SMHI-RCA is clearly the best performing RCM overall for this season with low mean difference produced across the whole European domain. In winter months, RCMs systematically are too warm in northern Scandinavia whilst being too cold in the south, particularly over the Alps.

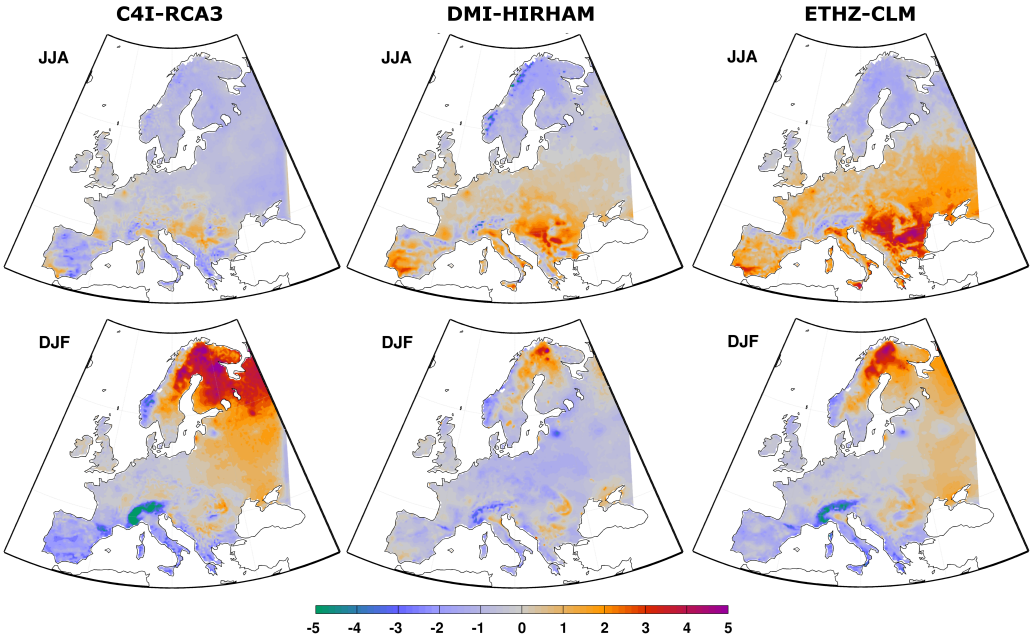

Figure 3.2: C4I-RCA3, DMI-HIRHAM and ETHZ-CLM - EOBS summer (JJA) and winter (DJF) mean 2-metre temperature (°C) differences over 1961-2000

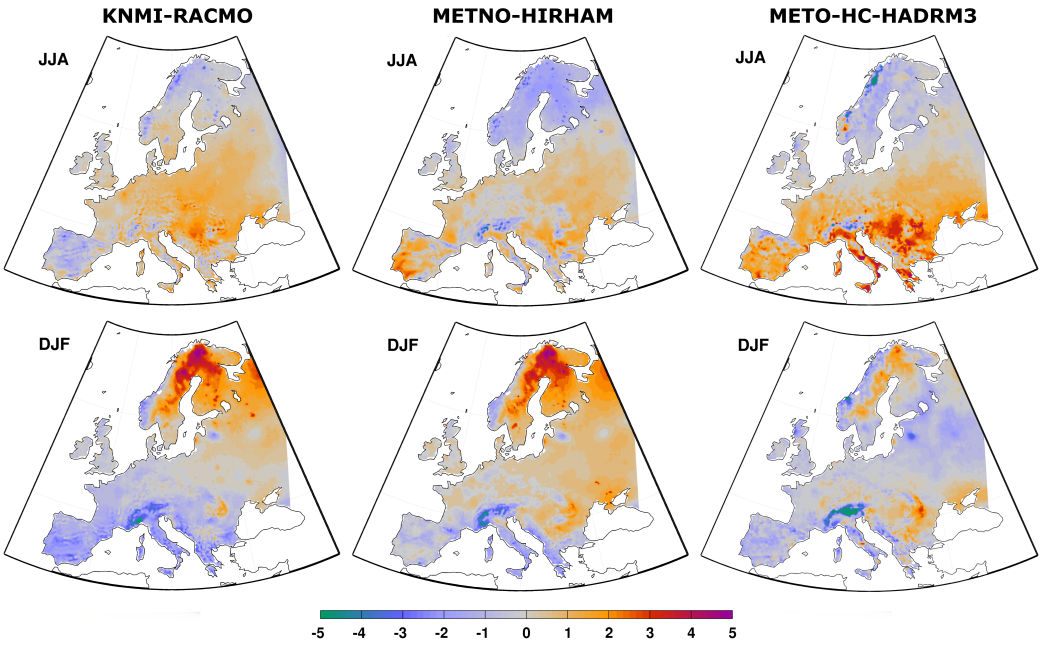

Figure 3.3: KNMI-RACMO, METNO-HIRHAM and METO-HC-HADRM3 - EOBS summer (JJA) and winter (DJF) mean 2-metre temperature (°C) differences over 1961-2000

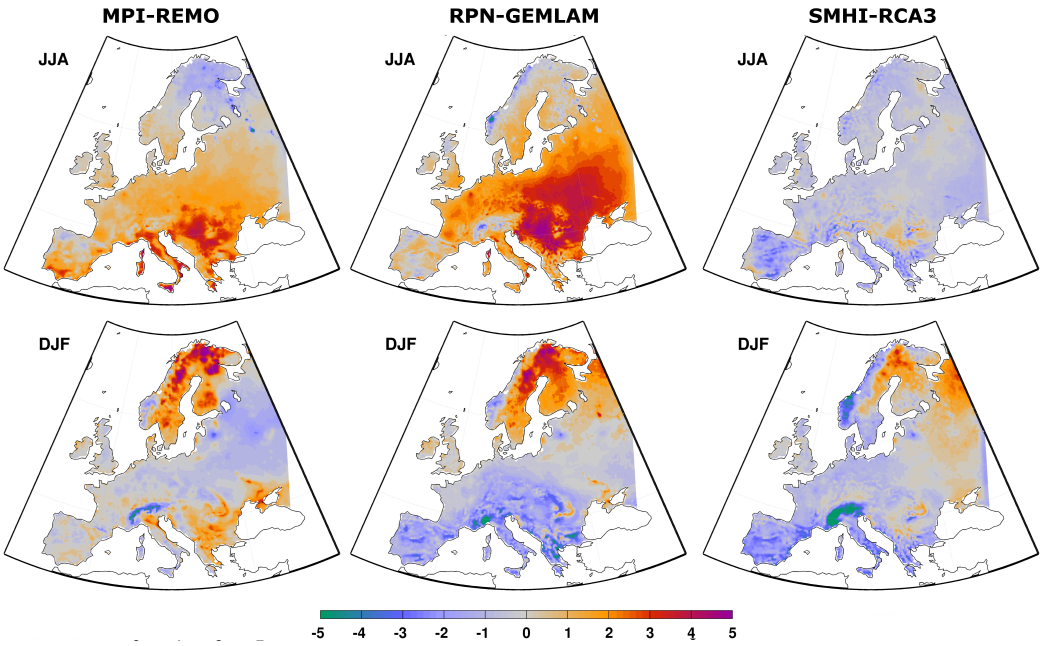

Figure 3.4: MPI-REMO, RPN-GEMLAM and SMHI-RCA - EOBS summer (JJA) and winter (DJF) mean 2-metre temperature (°C) differences over 1961-2000

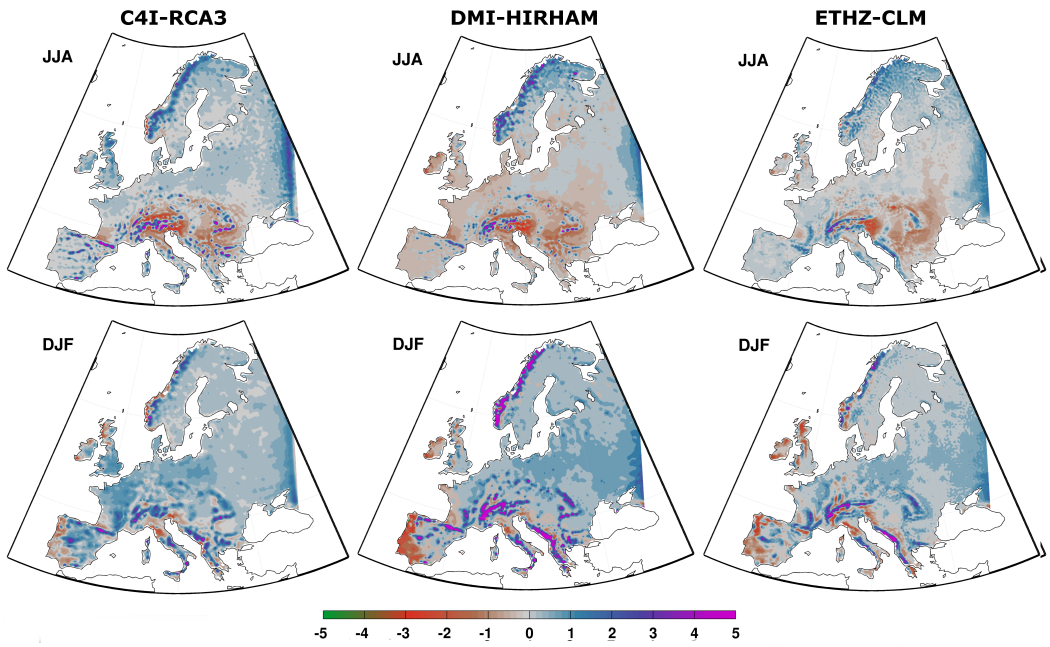

Figure 3.5: C4I-RCA3, DMI-HIRHAM and ETHZ-CLM - EOBS summer (JJA) and winter (DJF) mean precipitation (mm/month) differences over 1961-2000

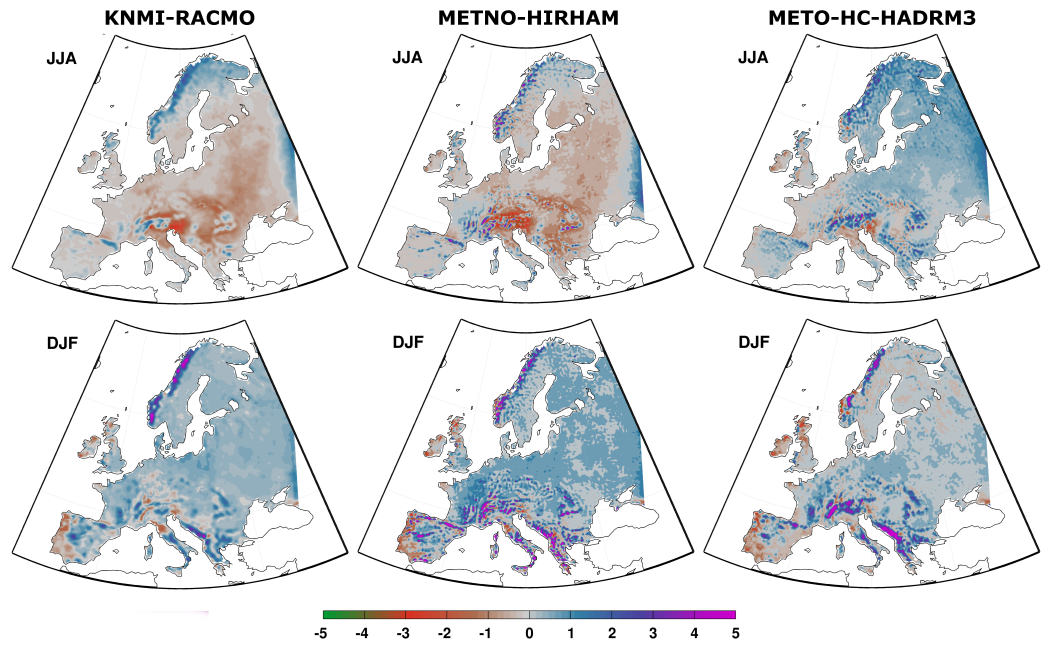

Figure 3.6: KNMI-RACMO, METNO-HIRHAM and METO-HC-HADRM3 - EOBS summer (JJA) and winter (DJF) mean precipitation (mm/month) differences over 1961-2000

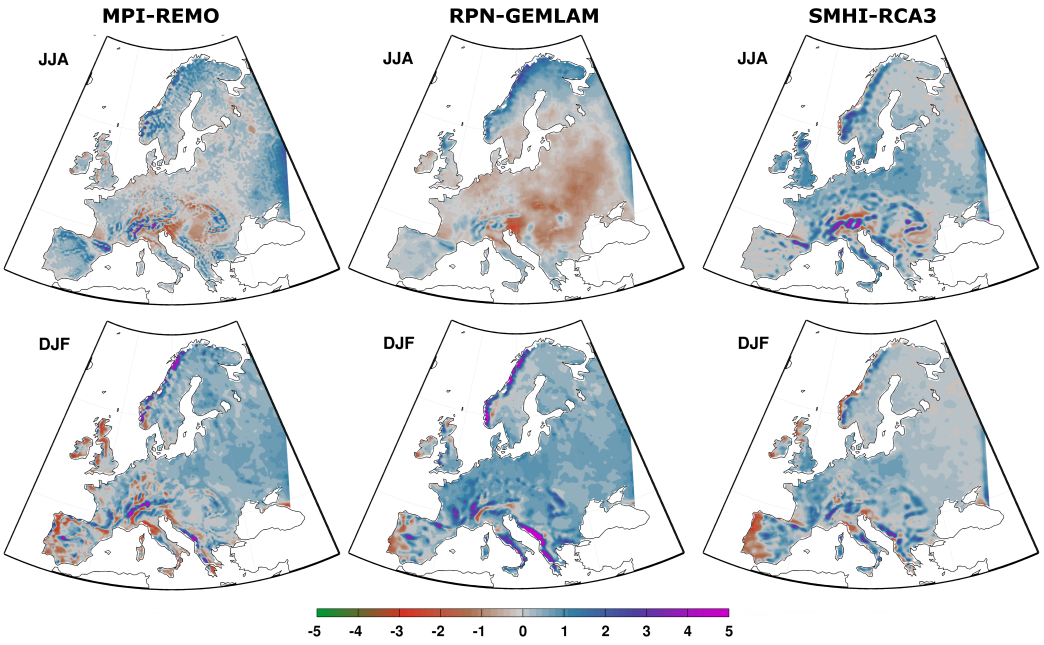

Figure 3.7: MPI-REMO, RPN-GEMLAM and SMHI-RCA - EOBS summer (JJA) and winter (DJF) mean precipitation (mm/month) differences over 1961-2000

In contrast to temperature biases, there are fewer shared systematic precipitation biases in summer, with the RCMs producing a both drier and wetter simulations over central and eastern Europe relative to E-OBS. In some regions however the RCMs are similar in their bias pattern, such as a wet bias in the higher altitude regions (e.g. Norwegian coast, Alps, Pyrenees). Winter months are generally wetter than summer overall, with the RCMs producing more similar bias patterns than in summer months. High altitude regions exhibit the highest precipitation bias whereas in eastern Europe RCMs perform generally better with low error.

3.2.1 Forcing Details

The European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-40 global reanalysis product, covering from late 1957 to 2002, is a hybrid of in-situ station based and (since 1970) satellite observations assimilated with the Integrated Forecast System (IFS) atmospheric model (Uppala *et al.*, 2005). It has 60 vertical levels of varying height, ranging from 0.5-1km in the troposphere to 1.5km in the stratosphere up to 0.1hPa, and a horizontal resolution of 1.121° (∼125km). The timestep used in the reanalysis is every six hours (Kallberg *et al.*, 2004), publicly available from the ECMWF online database. Although ERA-40 is based upon real world observations, and is in practice treated as such in many applications,

Institute-RCM	Greenhouse Gases	Solar Forcing	Aerosols
C4I-RCA3	'Effective' CO2 approx. linear increase (1.5ppm$_v$/yr)	1365W/m^2 (variable)	Included in CO2 forcing
DMI-HIRHAM	1960-1990 observed values, 1990 onwards supplied from MPI ECHAM4	1376W/m^2	ECHAM4 tabulated constant
ETHZ-CLM	360 ppm constant	1368W/m^2	constant
KNMI-RACMO2	SRES-A1B defined concentrations 1950-2100	1370W/m^2	Four aerosols;Tanré climatology
METO-HC-HADRM3	HadCM3 setup	1365W/m^2 constant	SO2 and Dimethyl Sulfide HadGEM setup
METNO-HIRHAM	1960-1990 observed values, 1990 onwards supplied from MPI ECHAM4	1376W/m^2	ECHAM4 tabulated constant values
MPI-REMO	Constant	Constant, undefined	Tanré climatology
RPN-GEMLAM	Constant	1367W/m^2	n/a
SMHI-RCA3	CO2 linear increase (1.5ppm$_v$/yr), other GHGs not accounted for	1370W/m^2	Constant

Table 3.3: External Forcing for included ENSEMBLES RCMs

it is not perfect for three reasons: forecast model biases, observational errors and the choice of data assimilation method (Bao and Zhang, 2013). Intercomparison between different reanalysis products and gridded observations shows how these differences in approach affect general biases, trends and variability across both the globe and throughout the 45 year time period (Simmons *et al.*, 2004). Generally these differences are small however errors do arise in specific regions and times, particularly in the earlier years, often due to a lack of observational coverage (Mooney *et al.*, 2011).

It is often not straightforward to locate detailed information for each RCM in regard to how they are forced, either by reanalysis or GCM, however, for those that did have published documentation some similarities are identified. The majority of the ERA-40 vertical atmospheric levels are used, interpolated to some three dimensional common grid. For SMHI-RCA3, vertical levels 13-60 (descending to

the surface level) were used interpolated to the identical spatial resolution as the RCM (Kjellström, 2005). These atmospheric fields are used as lateral boundary conditions and, in addition to sea-surface temperatures (SSTs) and sea-ice data for some RCMs such as MPI-REMO, are updated every 6 hours. In addition to lateral boundary conditions, other external forcings are included such as greenhouse gases (GHGs), solar and aerosols in the simulations, and these are detailed in Table 3.3. The ENSEMBLES RCMs are one-way forced; they only receive information from the lateral boundary conditions (LBC) and external forcings and do not feed back information. ETHZ-CLM, unlike the other models in the ensemble includes spectral nudging whereby the LBCs can influence the internal simulation to strengthen consistency with the large scale circulation (Separovic *et al.*, 2012). The models have varying levels of 'spin-up' time, where the RCM is given a period (5-10 years) to reach a consistent internal flow. This is done to account for processes that require a long period to produce behaviour consistent with reality (Laprise, 2008).

3.3 CORDEX RCM Data

The RCM data used for the Chapter 6 analysis focussing on the stationarity assumption of metric assessments is taken from the CORDEX ensemble (Giorgi *et al.*, 2009). The pseudo-reality method used in this analysis requires a set of RCMs with identical boundary conditions, so that the intercomparison between how the models respond to a changing climate is as fair as possible. Since GCMs have different climate sensitivities (Andrews *et al.*, 2012), the resulting forced RCM projections will inherit the climate change signal of the GCM and thus be dissimilar in their simulations characteristics. Testing the changes in bias distribution between two RCMs with different GCM forcings would be in effect testing how different the GCMs behave and not the RCMs. Identically forced RCM projections on the other hand provide the exact experimental setup which allows such a determination to be made, since any change in relative bias must be due to RCM features alone. From the CORDEX ensemble, three RCMs are identified which satisfy the requirement of the same GCM forcing, in this case HADGEM2-ES: CCLM4-8-17, RACMO22E and RCA4 (Table 3.4). Again, the simulations cover an historical (1971-2005) and future projection period (2006-2100) . The analysis only uses a total of two subsets from these simulated timeframe to assess the changes between the RCMs: 1971-2000 as an 'evaluation' period and a projection period 2071-2100. These GCM forced simulations follow the representative concentration pathway (RCP)

8.5 (Van Vuuren *et al.*, 2011), the highest radiative forcing increase. This forcing level is used to provide the most stringent test of bias stationarity, as this will maximise the climate change signal between evaluation and projection time-periods, ensuring that any non-stationary bias changes will be identified. To further assist in setting the most exacting test of RCM bias stationarity, the furthest apart time periods (1971-2000 and 2071-2100) are chosen, with the same aim of maximising the simulated signal.

Table 3.4: CORDEX RCMs forced by HADGEM2-ES details for those used in Chapter 6 analysis on the stationarity of metric assessments

Institute	**RCM**	**Country**	**RCP**	**Atmosphere Model**	**Land surface/Soil**	**Ocean/Sea Ice**
CLM (CLM Community with contributions by BTU, DWD, ETHZ, UCD,WEGC)	CLMcom-CCLM4-8-17	Germany	RCP8.5	ECHAM5	CLandM, Veg3D, Terra soil model	CICE ice model, NEMO ocean
KNMI (Royal Netherlands Meteorological Institute)	RACMO22E	Netherlands	RCP8.5	ECMWF	-	-
SMHI (Swedish Meteorological and Hydrological Institute)	RCA4	Sweden	RCP8.5	-	-	-

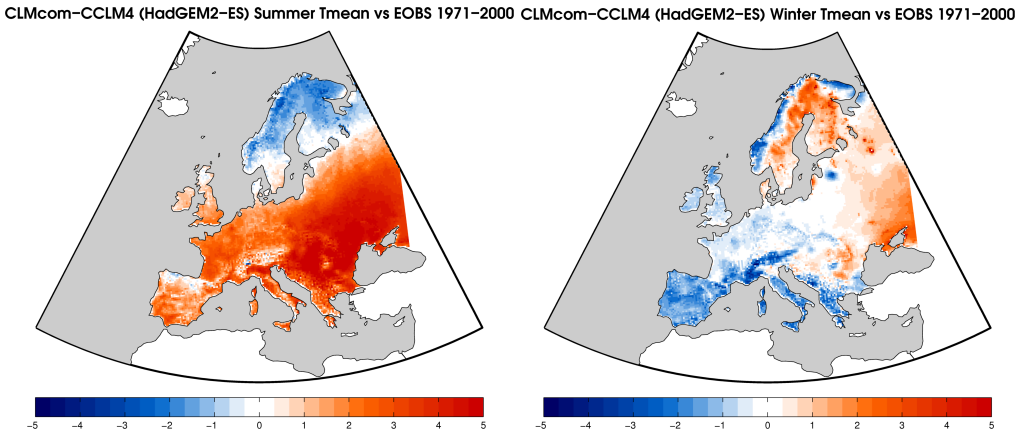

Figure 3.8: CLMcom-CCLM4 forced by HADGEM2-ES mean temperature bias against E-OBS (°C) for 1971-2000.

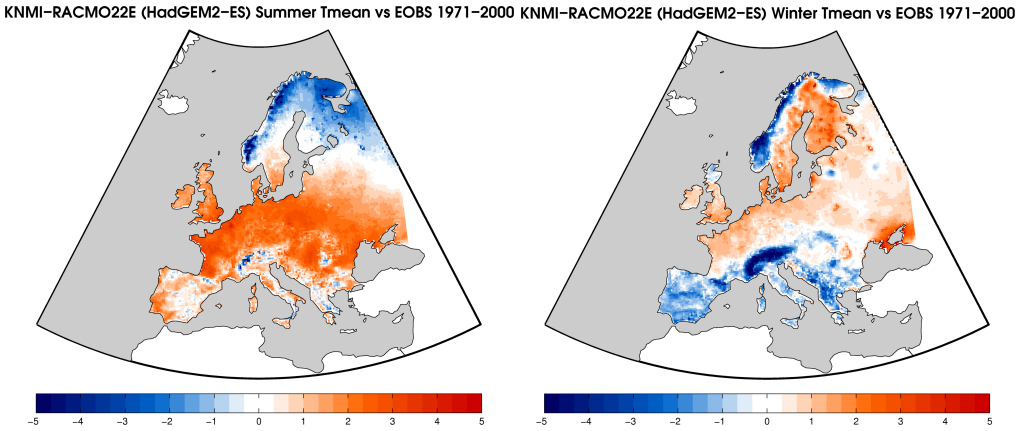

Figure 3.9: KNMI-RACMO22E forced by HADGEM2-ES mean temperature bias against E-OBS (°C) for 1971-2000.

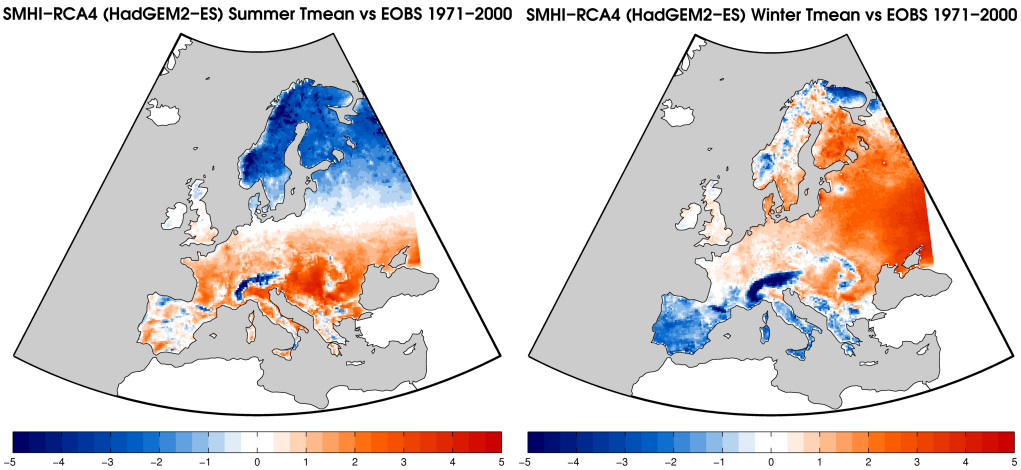

Figure 3.10: SMHI-RCA4 forced by HADGEM2-ES mean temperature bias against E-OBS (°C) for 1971-2000.

For the HADGEM2-ES forced RCMs CLMcom-CCLM4 (Figure 3.8), KNMI-RACMO22E(Figure 3.9) and SMHI-RCA4 (Figure 3.10), the historical temperature bias characteristics are similar both between the models and to ERA-40 forced ENSEMBLES simulations. Summer months see a north/south bias split, with central and southern Europe up to 5°C too warm whereas northern regions are systematically cold. In winter, southern areas are too cold, with the strongest biases occuring in the Alps. Central regions are simulated reasonably well with low average error, however in the north the RCMs are too warm except in some cases for the higher altitude Norwegian mountains.

3.4 Observational Datasets

To evaluate the quality of ENSEMBLES RCMs over the full European spatial domain, as opposed to a more localised study, gridded observational products are preferred for two main reasons. First, although RCMs are of a comparably high spatial resolution compared to GCMs, the output variables are still in effect smoothed. This means that comparison directly to raw station data would not be a consistent test of model errors over regions of highly variable orography for example. Gridded observational products necessarily require a smoothing process, and therefore provide a fairer test for models over larger spatial scales, since it is comparing 'like with like'. Furthermore, station data can be of shorter time lengths, or have missing values which can be remedied through considering information from other local stations through a gridded product. However, it should be noted that due to the low station density in E-OBS for example, significant errors can arise when compared to higher density localised products, which may have a detrimental impact on the robustness of model evaluations (e.g. Kyselỳ and Plavcová, 2010). More localised studies, such as statistical downscaling, make better use of station data as they are interested in the model bias at that precise area. Second, more specific to the ENSEMBLES RCM data are practicalities; the E-OBS data was produced specifically for comparison with this model ensemble and therefore is a natural choice.

To assess the quality of the RCMs in historical simulations forced by ERA-40, daily mean, minimum and maximum temperature, precipitation sums, and mean sea-level pressure data are available from the E-OBS database. They are output at both 25km and 50km resolutions on identical grids to the majority of ENSEMBLES

RCMs. The temperature and precipitation datasets were produced by Haylock *et al.* (2008) and cover the time period 1950-2006. They cover land surface regions, and draw together meteorological station data from 2316 locations, although this is not a constant value through time in the final data. The density of stations is not uniform (see Figure 3.11); the Netherlands, Ireland, Switzerland and the UK having the greatest spatial coverage. Other regions such as Iceland, Turkey, Northern Africa, and Northern Scandinavia have considerably fewer stations per km^2. Mainland countries such as France, Spain, Portugal, Germany and Poland have moderate station density. Quality control (QC) measures were implemented first to filter out any unrealistic values from each station's timeseries (e.g. maximum temperature<minimum temperature, precipitation<0). Station data passing these QC processes was then interpolated to the rotate-pole equal area grid, with interpolation error estimates calculated, output on separate files. These could be used in further work investigating the impact of observational uncertainty on RCM metric assessments, but for the following analysis are not used. As regards the station density it is unlikely that a dataset can be constructed such that each gridbox contains at least one real world meteorological station; for this dataset, each of the $625km^2$ gridboxes has on average approximately a 15% chance that it contains a station; for the more station sparse regions this likelihood will decrease.

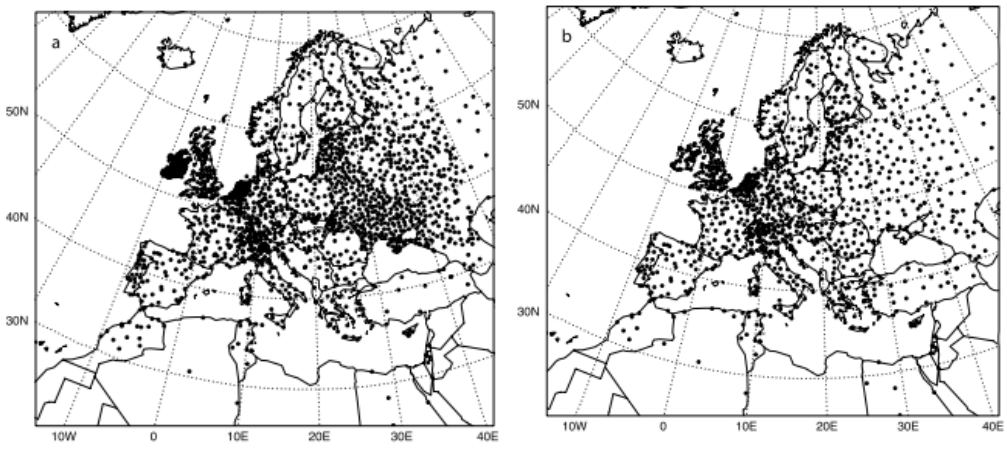

Figure 3.11: Station locations for E-OBS precipitation (left) and temperature (right) gridded datasets. Variable dataset density is clear, with the highest spatial coverage in central regions, whereas regions such as Iceland and Turkey have sparse coverage. Modified from Haylock *et al.* (2008).

E-OBS also has a gridded land-surface-only sea-level pressure dataset constructed by Van den Besselaar *et al.* (2011) which is also output on the ENSEM-

BLES RCM grids. The number of stations included in the dataset varies throughout the time period 1950-2010. Initially, data from approximately 100 stations was available, although this increases roughly linearly to 416 in 2010. Van den Besselaar *et al.* (2011) note that it is of no disadvantage to use fewer stations than in temperature datasets, as the latter are expected to be more heterogeneous over the spatial domain, and as such require higher station density. However, one drawback of this dataset is the lack of stations in some of the Eastern and Northern European countries, such as Ukraine, Romania, Belarus, Lithuania, Finland and Sweden. QC procedures implemented in the construction of this dataset included a removal of concurring extremely low or high values but the lack of a homogeneity test for the stations data is one of the weaknesses of the original dataset, but is said to have been rectified in later versions (Van den Besselaar *et al.*, 2011).

3.5 Data Format and Preprocessing

RCM data was obtained from the ENSEMBLES and CORDEX online databases, both of which provide data in NetCDF file format. This type of file delivers a structured array configuration, ideal for large gridded climate model output. For each variable, the data is given as a three-dimensional matrix; two spatial dimensions and one in time. Additionally, the generated NetCDF file specification includes grids for latitude and longitude coordinates, number of time steps in addition, variable units, to more general simulation attributes such as model specification, forcing and the mapping projection type. The majority of the ENSEMBLES RCM simulations are run on an identical rotated-pole grid projection of dimensions 190x170 grid points, which similarly has been used in the E-OBS dataset. The CORDEX RCMs also are on rotated-pole grids, and all cover an identical grid size of 412x424. This projection type produces approximately equal area grid boxes, and is the unmodified and uninterpolated output from each RCM. Some of the ENSEMBLES RCMs are given on atypical grids, which are equal area projections, but do not match with the observed datasets.

The different types of RCM data used in the analyses were subject to a range of processing stages prior to use in any evaluation. Firstly, decadal or otherwise incomplete data was concatenated to assemble the full timeseries for evaluation. Next, units were standardised; for example precipitation data is provided in units of $\mathrm{kgm^{-2}s^{-1}}$ for RCMs and in millimetres for E-OBS, temperature data in K for

models and °C for E-OBS. A land-sea mask is applied to only leave land surface regions, as gridded observations are only available in these locations. This land-sea mask is different for CORDEX and ENSEMBLES RCMs given their different spatial resolution, and is also computed in different ways. For ENSEMBLES, since it is to be directly compared to the land-only E-OBS dataset, only those grid points for which E-OBS has any data are used. Observational gridboxes of which timeseries are incomplete are also removed from the ENSEMBLES analysis as are land surface regions of North Africa, Iceland and South-Eastern regions such as Turkey. The former is done to ensure that extreme indices (which are calculated in many cases on an annual basis) are correctly represented, the latter to exclude regions with little in the way of observational station data spatial coverage in the E-OBS datasets. For CORDEX, grid cells with $<10\%$ land area are excluded (calculated from the percentage land-fraction dataset). Leap years are taken into account by removing February 29^{th} days from the timeseries, to leave (for reanalysis forced historical simulations) the number of time steps at 14600=365*40. This is done to simplify considerably the programming of evaluation code without loss of significant information. Although for the majority of analyses presented in this thesis the omission of 1 leap year day's data is unlikely to make substantial difference, for the calculation of cold day persistence extreme indices this factor may become much more relevant in more detailed examinations. The observations and ENSEMBLES RCM data matrices are then cropped to the standard domain size of 190x170 grid points, ready to be used for calculations.

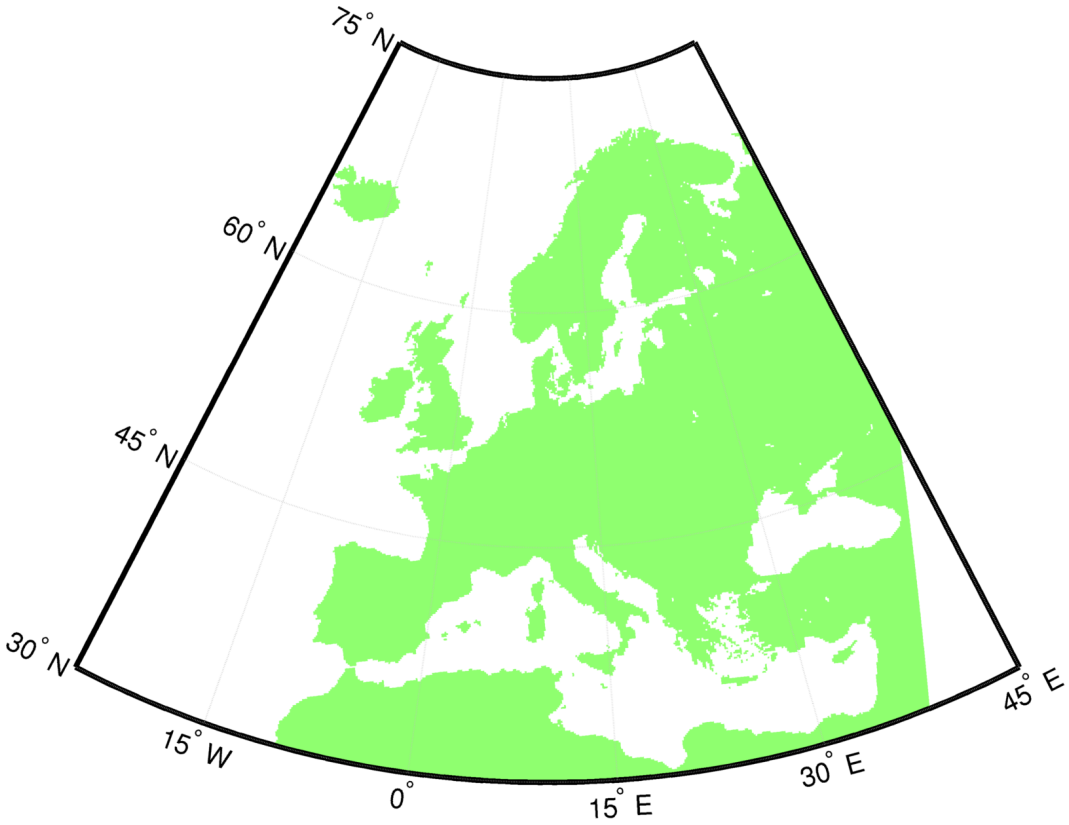

Figure 3.12: Land-sea mask used for CORDEX RCM analysis

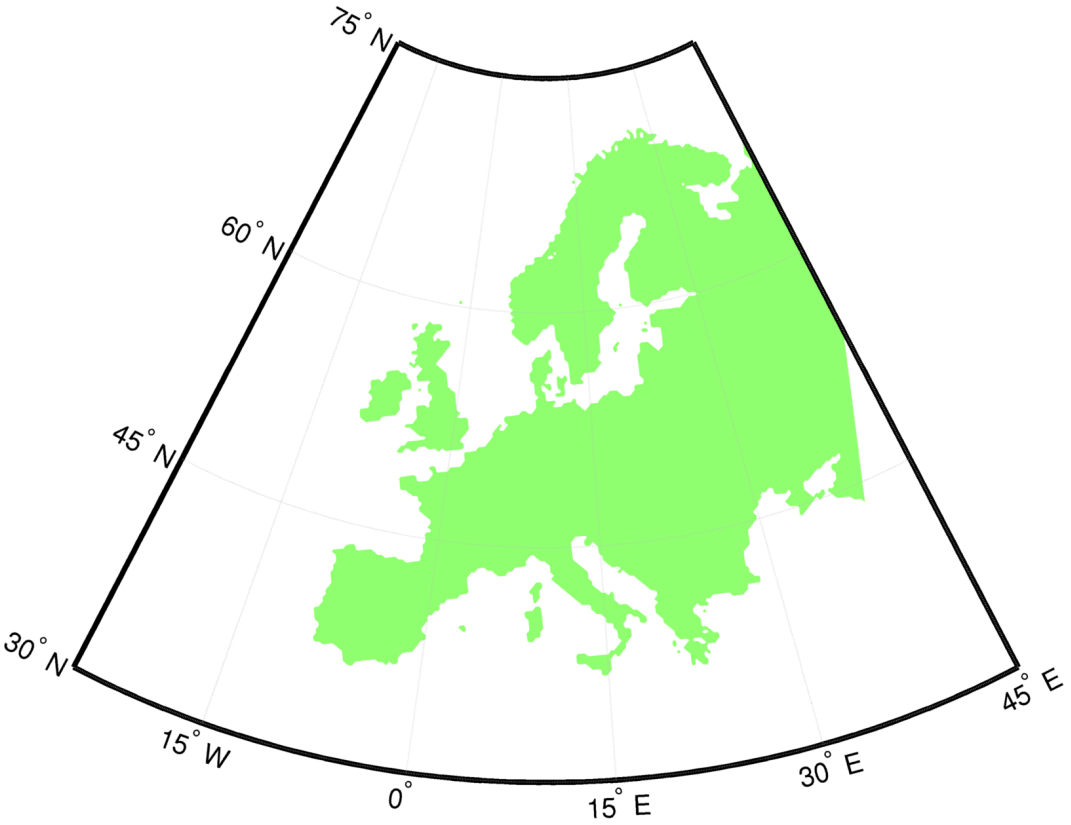

Figure 3.13: Land-sea mask used for ENSEMBLES RCM analysis

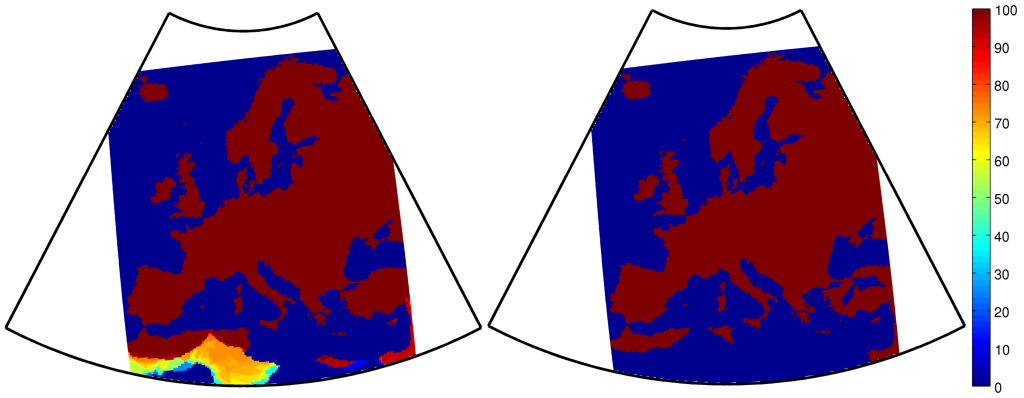

Figure 3.14: Percentage timeseries with no missing values for Temperature (Left) and Precipitation (Right) E-OBS datasets

Chapter 4

Metric Sensitivity

4.1 Introduction and Methodology

When evaluating an RCM with a performance metric, there are four distinct metric component choices to be made: variable, spatial-temporal domain, statistic and observational dataset. Quantitative RCM evaluations will to different degrees be influenced by these choices, and as such it is of central importance to understand how the overall metric is sensitive to changes in each component. The following analysis focuses on assessing the sensitivity to the first three metric elements by evaluating ENSEMBLES RCMs over a range of variables, temporal and spatial domains and statistics against gridded observational data products. The purpose of this investigation is to first determine the degree of robustness of different metric types to small changes in methodological approach, and secondly to provide guidance on how to approach the evaluation of RCMs with performance metrics. Evaluation of RCMs and the causes of model error are not per se the objective of the study, although the aims of the investigation require this step to be carried out.
The three different sources of metric sensitivity tested in this chapter utilise RCM assessments using temperature (Tmax, Tx90p, Tmin, Tn10p, DTR), precipitation and sea level pressure and associated extreme indices, and this set is reduced to a final set of 16 for redundancy analysis (Section 4.5). These indices (Tables 4.1, 4.2 and 4.3) are selected from the set proposed by the Expert Team on Climate Change Detection and Indices (ETCCDI) by three criteria: they must be appropriate for model-observation comparison, that the final group should capture as much information as possible and must evenly cover the three main elements of extreme events : frequency, magnitude and persistence.

The statistics utilised in the analysis fall into four separate categories: standard error, temporal variability, spatial pattern and event frequency statistics, each selected to assess different moments of variability. The simplest and most common form of statistic is the standard error type. These are generic in construction and can be applied in evaluating many different aspects of model simulation performance. Examples of such statistics are Root Mean Squared Error or Standard Deviation (Table 4.4). Some of these are used more than once in the following analysis, both in the evaluation of spatial patterns and for other gridpoint type output. The second type of statistic are Temporal Variability statistics (Table 4.1), which are more specific in construction and purpose, aimed at evaluating RCM simulations in their representation of linear trends or interannual variability. The third type of statistic used covers Spatial Pattern statistics (Table 4.5), and finally 'event frequency' statistics (Table 4.7), used to assess RCMs over the distribution of model errors.

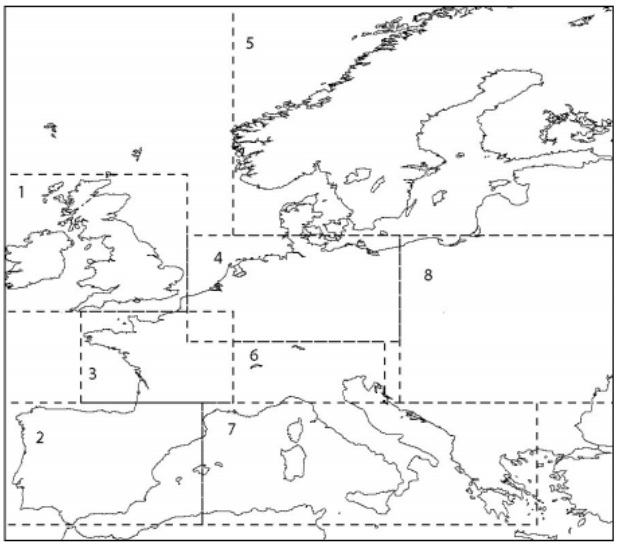

Figure 4.1: The eight sub-European 'Rockel' regions as first suggested by Burkhardt Rockel and Katja Woth for use in analysis of regions of homogeneous character. Modified from Rockel and Woth (2007)

To evaluate the degree of sensitivity of these metric types to minor changes in their component choices, RCM performance is quantified for each variable, index, statistic and spatial domain. These domains cover the 'Rockel' regions (Figure 4.1), leading to nine domains overall: the whole European domain and eight sub-domains of varying area sizes. Each metric component requires specific comparisons to judge the level of inherent sensitivity due to the specific contribution each makes to the overall RCM assessment process. For the statistics, relative metric output will

be produced and transformed such that metric output increases with model performance (several metrics such as RMSE decrease in value as performance increases) to aid visual appraisal of how metric output varies with the underlying statistic. Transforming involves a procedure for metrics which increase with increasing error (RMSE, MAE and IA); these metrics are transformed by calculating the inverse for each value. This transformed output is then further modified by normalising it, and thus making each statistic comparable in terms of how the overall distribution of model performance is presented by each in turn. This is carried out by removing the mean and dividing by the standard deviation, removing differences both in magnitude and in variation. Normalising therefore enables the differences between two or more statistics, which may be quite different with respect to the level of dispersion (the magnitude of how wide the range of metric output is), to be evaluated.

Table 4.1: Extreme Indices - Frequency

ETCCDI Number	Name	Description	Chosen	Reason
1.	FD	Number of frost days (Annual count)	yes	Measure of anomalies in the length of spring and autumn seasons (Tebaldi *et al.*, 2006).
2.	SU	Number of summer days (Annual count)	no	Arbitrary threshold, will not be comparable over domains
3.	ID	Number of icing days (Annual count)	yes	Measure frequency of more extreme sub-zero Tmax events
4.	TR	Number of tropical nights (Annual count)	no	Arbitrary threshold, will not be comparable over domains
10.	TN10p	Percentage of days when TN $<$ 10th percentile	no	Best used for assessing changes in exceedance rates over long time frames. Standard percentile thresholds given below will provide more meaningful information for shorter analysed time-series.
11.	TX10p	Percentage of days when TX $<$ 10th percentile	no	“
12.	TN90p	Percentage of days when TN $>$ 90th percentile	no	”
13.	TX90p	Percentage of days when TX $>$ 90th percentile	no	“
20.	R10mm	Annual count of days when PRCP $\geq$ 10mm	yes	Arbitrary absolute threshold; can be better represented by the given percentile measure in index 22.
21.	R20mm	Annual count of days when PRCP $\geq$ 20mm	yes	”
22.	Rnnmm	Annual count of days when PRCP $\geq$ nnmm, nn is a user defined threshold	no	Provides a measure of moderate and high intensity precipitation events. nn is set to the 90th and 95th percentiles of daily precipitation, calculated for each grid box from observations

Table 4.2: Extreme Indices - Magnitude. Percentile thresholds are calculated using a 5-day running window centred on each calendar day in question to estimate percentiles from a $5 \cdot 40 = 200$ day sample (1961-2000 period), as recommended by Zhang *et al.* (2005)

Number	Name	Description	Chosen	Reason
n/a	Tx90p threshold	90th percentiles of Tmax	yes	Provides spatially comparable information relating to the distribution of both moderate and severe extremes. 90th percentile chosen to give robust output from 40 year time series as well as being the most commonly adopted threshold.
n/a	Tn10p threshold	10th percentile of Tmin	yes	“
6.	TXx	Monthly maximum value of daily maximum temperature	no	Can be excessively influenced by anomalous results
7.	TNx	Monthly maximum value of daily minimum temperature	no	“
8.	TXn	Monthly minimum value of daily maximum temperature	no	”
9.	TNn	Monthly minimum value of daily minimum temperature	no	“
16.	DTR	Diurnal temperature range (Monthly)	yes	Not directly evaluated in other metrics
17.	Rx1day	Monthly maximum 1-day precipitation	yes	Provides a measure of short term intense precipitation events
18.	Rx5day	Monthly maximum consecutive 5-day precipitation	yes	Provides a measure of long term intense precipitation events
19.	SDII	Simple precipitation intensity index (Annual)	no	Not a measure of extremes

Table 4.3: Extreme Indices - Persistence

Number	Name	Description	Chosen	Reason
5.	GSL	Growing season length (Annual)	no	Not an indicator of extremes
14.	WSDI	Warm spell duration index. Annual count of days with at least 6 consecutive days when TX $>$ 90th percentile	yes	Measure of persistency of warm extremes and is spatially comparable given that it is based on a percentile threshold.
15.	CSDI	Cold spell duration index. Annual count of days with at least 6 consecutive days when TN $<$ 10th percentile	yes	Measure of persistency of cold extremes and is spatially comparable given that it is based on a percentile threshold.
23.	CDD	Maximum length of dry spell, maximum number of consecutive days with RR $<$ 1mm (Annual)	yes	Provides an indicator of sustained drought conditions
24.	CWD	Maximum length of wet spell, maximum number of consecutive days with RR $\geq$ 1mm (Annual)	yes	Provides an indicator of sustained precipitation events

Statistic	Equation	Reference
Root Mean Squared Error (RMSE)	$\text{RMSE}(M,O) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(M_k - O_k)^2}$	
Mean Absolute Error (MAE)	$\text{MAE}(M,O) = \frac{1}{n}\sum_{k=1}^{n}\lvert M_k - O_k\rvert$	
Standard Deviation (SD) (σ)	$\sigma(M) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(M_k - \bar{M})^2}$	
Index of Agreement (IA)	$\text{IA} = 1 - \left(\frac{\sum_{k=1}^{n}(M_k - O_k)^2}{\sum_{k=1}^{n}\left[\lvert M_k - \bar{O}\rvert + \lvert O_k - \bar{O}\rvert\right]^2} \right)$	Legates and McCabe Jr (1999)

Table 4.4: Standard Error Statistics. Here, M_k and O_k refer to RCM and Observed data at gridpoint (or timestep if considering single timeseries) k respectively, $\bar{O}$ is the observational mean, $\bar{M}$ the RCM mean value.

Statistic	Equations	Reference
Spatial Skill Score (SSS)	$\text{SSS}\ (M,O) = 1 - \left(\dfrac{\sum_{k=1}^{n}(M_k - O_k)^2}{\sum_{k=1}^{n}(\bar{O} - O_k)^2} \right)$	Pierce *et al.* (2009)
Correlation (R)	$\text{R}(M,O) = \dfrac{1}{n\,\sigma_m \sigma_o} \sum_{k=1}^{n}(M_k - \bar{M})(O_k - \bar{O})$	Taylor (2001)
Spatial Skill Metric (SSM)	$f_3 = \sum_{k=1}^{n} \dfrac{1}{\lvert S_{m,k} - S_{o,k} \rvert}$ where $S_k = \mu_k - \mu_l \quad \forall\ k,l\ \epsilon\ \{1,...,n\}$	Eum *et al.* (2012)

Table 4.5: Spatial Pattern Statistics. Here, σ_m and σ_o are defined as the RCM and observational standard deviation respectively, μ_k and μ_l represent mean values for grid points k and l where the total number of gridpoints is n. $S_{m,k}$ and $S_{o,k}$ are length-n vectors for RCM and observational data respectively with entries $\mu_k - \mu_1, \mu_k - \mu_2, ..., \mu_k - \mu_n$. For each gridpoint k, the difference of that gridpoint's value to all other gridpoints l is calculated, and the sum of these differences given by each S_k. Next, and the difference between the observational and simulated S_k is calculated. This produces a scalar quantity, of which the inverse of the absolute is taken. The summation over all gridpoints k is then carried out. According to Eum *et al.* (2012), this metric measures the 'spatial distribution of difference of mean values between a gridpoint and the rest of grid points within the region of interest; therefore it gives an idea about the heterogeneity/homogeneity of the spatial information that an RCM is able to reproduce with respect to the observed value'.

Statistic	Equations	Reference
Annual Cycle Skill Score (ACSS)	$\text{ACSS}\ (M,O) = 1 - \left(\dfrac{\sum_{k=1}^{12} w_k (M_k - O_k)^2}{\left(\sum_{k=1}^{n} (\bar{O} - O_k)^2 \right) \left(\sum_{k=1}^{12} w_k \right)} \right)$	Holtanová *et al.* (2012)
Annual Variability Metric (AVM)	$f_2 = \dfrac{1}{\sum_{k=1}^{n} \left(A_{m,k} - A_{o,k} \right)}$ where $A_{m,k} = \dfrac{a_{m,k} - \bar{a}_m}{\sigma_m}$	Eum *et al.* (2012)
Interannual Variability Metric (IVM)	$f_2(T_{var}) = \dfrac{\epsilon_{\sigma_t}}{\lvert\sigma_m - \sigma_o\rvert}$, $f_4(P_{var}) = \dfrac{\epsilon_{\sigma_p}}{\lvert\sigma_m - \sigma_o\rvert}$	Xu *et al.* (2010)
Linear Trends Metric (LT)	$\text{LT} = 1 - \dfrac{\lvert\beta_m - \beta_o\rvert}{\zeta - \lvert\beta_m - \beta_o\rvert}$	Lorenz and Jacob (2010)

Table 4.6: Temporal Variability Statistics. For the ACSS, w_k is the number of days in month k to weight each month equally. In the AVM, $a_{m,k}$ represents the mean annual value of RCM m for year k, $\bar{a}_m$ is the mean climatological monthly value for RCM m, $A_{o,k}$ is defined equivalently to $A_{m,k}$. $\epsilon_{\sigma t}$ and $\epsilon_{\sigma p}$ $are measures of observed natural variability for temperature and precipitation respectively, defined by the difference between maximum and minimum val$ $year moving averages. \beta_m$ and β_o are the gradient slopes for RCM and observed temperature trends, and ζ is a pre-defined constant, here set at 0.5.

Statistic	Equations	Reference
PDF Overlap Skill Score (PDF)	$\text{PDF} = \sum_{b=1}^{n} \inf \left\{ \text{PDF}_{m,b}, \text{PDF}_{o,b} \right\}$	Perkins *et al.* (2007)
CDF Metrics	$f_1 = 1 - \sqrt{\frac{\lvert A_m - A_o \rvert}{2A_o}}, \quad f_2 = 1 - \sqrt{\frac{\lvert A_m^+ - A_o^+ \rvert}{2A_o^+}}, \quad f_3 = 1 - \sqrt{\frac{\lvert A_m^- - A_o^- \rvert}{2A_o^-}}$	Sánchez *et al.* (2009)
	$f_4 = 1 - \sqrt{\frac{\lvert \overline{P_m} - \overline{P_o} \rvert}{2\overline{P_o}}}$ and $f_5 = 1 - \sqrt{\frac{\lvert \sigma_m - \sigma_o \rvert}{2\sigma_o}}$	

Table 4.7: Event Frequency Statistics. Here, $\text{PDF}_{m,b}$ and $\text{PDF}_{o,b}$ are the PDF distributions of RCM and observed timeseries respectively, the metric measures the overlap between the two areas. A_m and A_o refer to the areas under the CDF curves of RCM and observational data respectively, A^+ and A^- refer to the regions above and below the 50th percentile respectively. $\overline{P_m}$ and $\overline{P_o}$ are defined as the time and space average of simulated and observed data respectively.

These statistics (Tables 4.4 4.5 4.6 4.7) although on the surface do span a range of various alternative, there are several shared characteristics with which it is possible to make some predictions. First, there is a set of statistics (RMSE, MAE, IA and SSS) which share a central 'model-observation' component, with the remainder of the statistics often applying some normalising factor. Taking RMSE and MAE first, it is clear these two statistics are the closest in construction, however, since RMSE involves squaring errors, it will inevitably lead to an emphasis of any highly variable input data. IA on the other hand also has a 'squaring-errors' term, but also normalises by a factor called the 'potential error', which is also sensitive to extreme values. This term appears to try to offset model scores if either the observed or modelled variance is too large, assisting in generating better scores. The Spatial Skill Score too is normalised by the observed variance, and therefore for homogeneous data will behave much like RMSE or MAE, but for high variable datasets will factor in the difficulty for RCM to replicate the values, as such it is a less stringent statistic. The remaining statistics are not as closely related, in part due to their more targeted nature, such as the linear trends statistic.

To evaluate the more general behaviour and relationship of the statistics, variables and domains to one another, Principal Component Analysis (PCA) is utilised to objectively identify similarities within these metric construction choices. To apply PCA to the metric output, the 4-d metric, comprising variable, model, domain and statistic dimensions, must be transformed into a 2-d matrix (columns the variable of interest, in this example statistics, rows observations). Before applying PCA the data are standardised as each statistic produces output which is of different units, and each variable has a different range of variability.

For the example of statistic sensitivity, the 2-D matrix to which the PCA is applied to is produced by concatenating several smaller 2-D matrices each of which represents a variable with columns statistics and rows RCMs. This leads to a 144x7 matrix or 144x5 (9 RCMs * 16 variables by 7 or 5 statistics, depending on the analysis in question). The domain from which this data is taken from is held constant, and so this PCA is processed nine times for each domain (total European domain and eight sub-domains). The PCA is calculated from the covariance matrix of this 2-d input matrix after normalisation of the input variables. The reason for standardising the data is that each variable has a different range of output and it is important that one avoids one variable disproportionately skewing the total variance over all evaluations (if one variable has a large variance and the data is normalised

after concatenating, the other variables will appear to have little variance, which will adversely affect the PCA results).

In addition to this PCA approach for evaluating how absolute metric output varies with respect to statistic, variable choice or domain, the relative model performance for each configuration is analysed by looking at model ranks, the performance of RCMs relative to one another in the ensemble. These are qualitatively assessed as to how much the changing of statistic can alter the overall relative positions of RCMs within the ensemble. Issues such as whether the model ranks are concentrated or dispersed or if a clear hierarchy of models can be determined is examined. These would both be characteristics of metric output which is either robust or sensitive to the choice of statistic in the construction.

4.2 Sensitivity to Choice of Variable

In the second group of performance metric applications, the main purpose is to enable a more general characterisation of model simulation quality to be made and therefore inevitably spans a wider array of variables. This can take the form of comprehensive tests of skill over a number of commonly used variables such as temperature, precipitation commonly in terms of a basic error statistic (e.g. Murphy *et al.*, 2004; Gleckler *et al.*, 2008; Reichler and Kim, 2008). However in the past such studies have either neglected temporal aspects and the representation of climate extremes, or have not spanned a wide range of variables, potentially weakening the certainty of final model assessments. This thesis is focussed towards understanding how robust this second type of general model performance metric is, both in terms of a single metric and a set of metrics, and the following analysis aims to investigate to what degree the choice of variable can have on the assessed model performance.

Temperature Variables

The nine RCMs assessed in this study are taken from the European ENSEMBLES project, and are evaluated against the E-OBS gridded temperature observational datasets. Figure 4.2 shows the minimum, mean and maximum annual temperature timeseries produced from annual spatial averages for the RCM ensemble and E-

OBS. Only European area average RCM assessments are considered in this section. Overall, the models have good performance in simulating mean temperatures, with the average magnitude of error being approximately 0.2-0.5°C, although as Lorenz and Jacob (2010) found the models do not capture the observed warming trend. For the minimum and maximum temperatures (Tmin and Tmax respectively), a wider range of model performance is identified. For the most part, the RCMs on average systematically overestimate Tmin and underestimate Tmax leading to a smaller than observed DTR. Mean absolute errors for Tmin range from 0.2-3.7°C, and 1.6-4.8°C for Tmax.

Information gained through evaluating the RCMs only at the mean of the temperature distribution in the case of this ensemble does not capture the full characteristics of model behaviour. Some RCMs which perform poorly at the extremes cannot be identified when assessed in their representation of mean climatology, as is the case with SMHI-RCA and RPN-GEMLAM. The fact that these two RCMs are at the tails of the mean temperature timeseries ensemble distribution does not guarantee that similarly performing models will therefore be poorly performing at the extremes, as illustrated by MPI-REMO which has a similar mean temperature error as RPN-GEMLAM but has substantially lower errors particularly in its simulations of Tmax and Tx90p. Furthermore, the inadequacy of evaluating RCMs solely on their skill of replicating mean temperatures is shown by the assessment of diurnal temperature range (DTR) performance, as the highly unrealistic performance of RPN-GEMLAM would not be able to be identified using this single test of model skill.

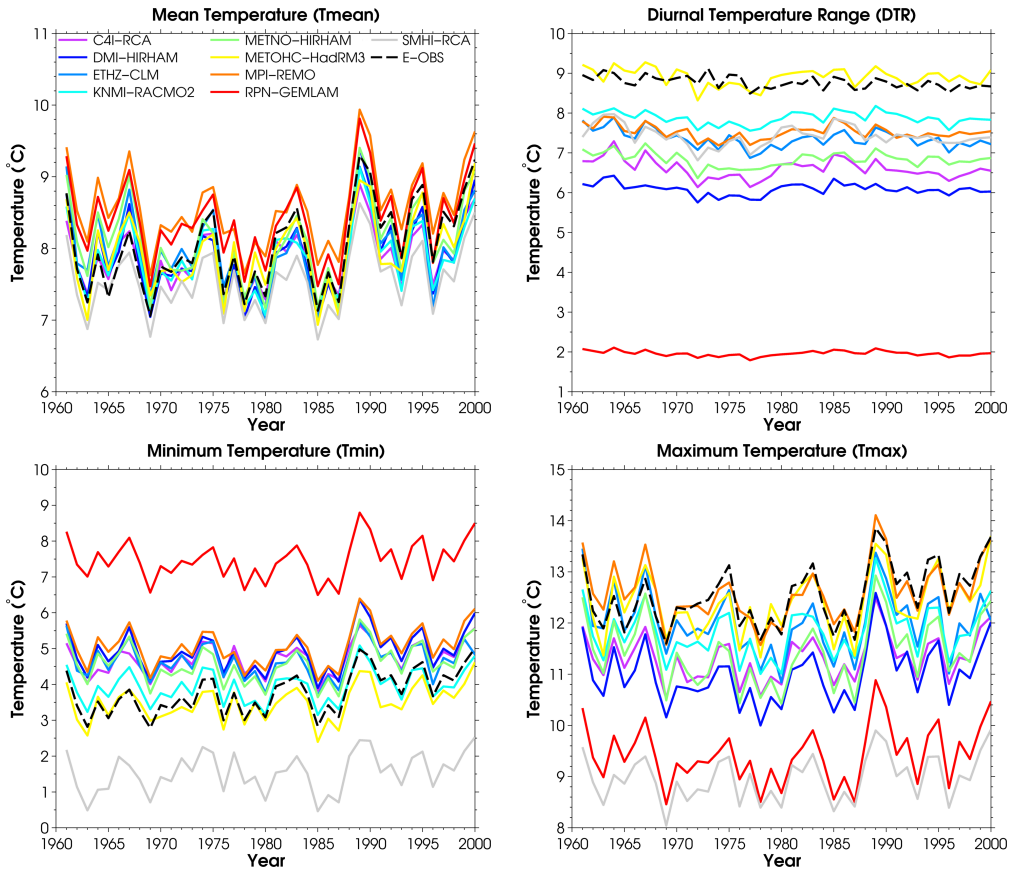

Figure 4.2: Europe spatial-average annual mean minimum, mean and maximum temperature and diurnal temperature range time series for ENSEMBLES RCMs, E-OBS observations black dotted line.

Whether it is sufficient to evaluate RCMs in their representation of Tmax and Tmin to gain information on DTR would depend on the metric considered. Commonly used absolute magnitude metrics would not take into account the direction of model error and as such no inference can be made from Tmax and Tmin scores as to the RCM skill at simulating a realistic DTR. A DTR score on its own would also not give a complete account of model characteristics since a model which scores poorly in both Tmax and Tmin may have a uniform bias which could give rise to an apparently skilful representation of DTR. This is the case with SMHI-RCA, which is the least realistic RCM with respect to Tmax and Tmin, but scores as the third best RCM in terms of DTR. It is recommended that metrics evaluating the direction of errors for Tmax and Tmin are included, or that DTR skill is additionally assessed alongside Tmax or Tmin absolute scores.

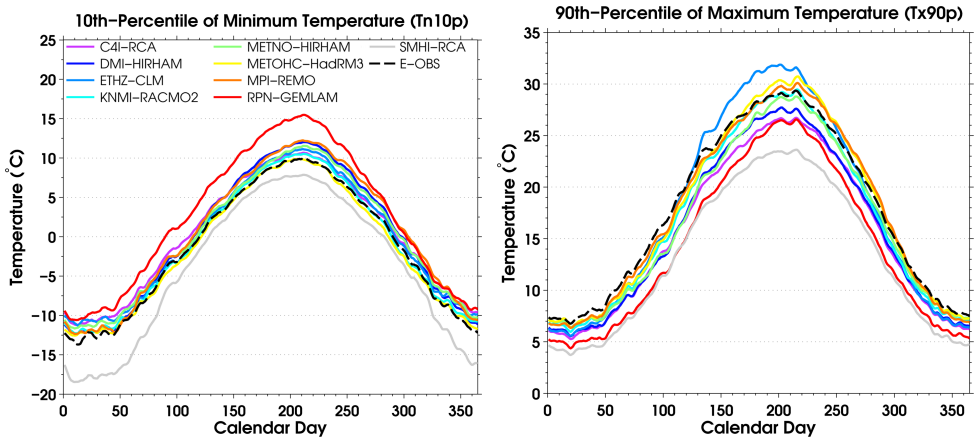

Figure 4.3: Calendar day percentiles for 10th-percentile of Tmin and 90th-percentile of Tmax averaged over the European domain for ENSEMBLES RCMs, E-OBS black dotted line.

Some of the extreme indices considered below are based on the extreme tails of the temperature distribution: the 10th-percentile of Tmin (Tn10p) and the 90th-percentile of Tmax (Tx90p). One question is whether these two percentile based indicators give further information alongside Tmax and Tmin and consequently if they should be included in any final set of metrics used to provide a general overview of model performance. Figure 4.3 shows the distribution of RCMs and the observed percentiles for each calendar day calculated over the 40-year period 1961-2000. The systematic biases identified in the simulations of Tmin and Tmax are still present in these more extreme cases, yet the magnitude of errors in the majority of cases increases. There is some seasonal variation in the magnitude of error, particularly for the summer/winter Tx90p comparison, although the direction of errors remains for the most part constant. Figure 4.5 shows the temporal RMSE for the five temperature distribution variables Tn10p, Tmin, Tmean, Tmax and Tx90p. Model errors are substantially higher at the extremes than at the mean, and deficiencies clearer for some poorly performing models. In some cases, models which perform worse at the extremes perform better than others nearer the centre of the

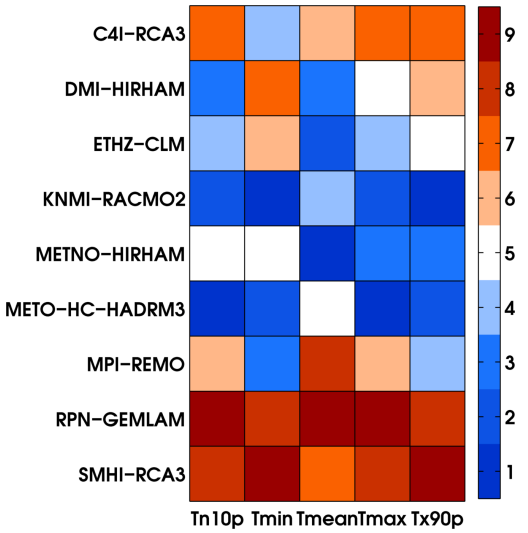

Figure 4.4: RCM skill ranking scores for representation of Tn10p, Tmin, Tmean, Tmax and Tx90p. Skill increasing with lower ranks.

distribution, for example C4I-RCA3 has a higher Tn10p error than DMI-HIRHAM, but outperforms the same model in Tmin, and similarly for RPN-GEMLAM and SMHI-RCA for Tmax and Tx90p. More generally, the RCM error characteristics are similar; most RMSE scores being less than 2°C. The last two models stand out as the worst performing with this metric although it is probably unnecessary to assess the RCMs for Tn10p and Tx90p to identify these large biases, since the RMSE scores for Tmax and Tmin are by themselves very large relative to the rest of the ensemble. However, this does not necessarily preclude assessment of the frequency and persistence of temperature extremes based on Tx90p and Tn10p as they relate to different aspects of the temperature distribution.

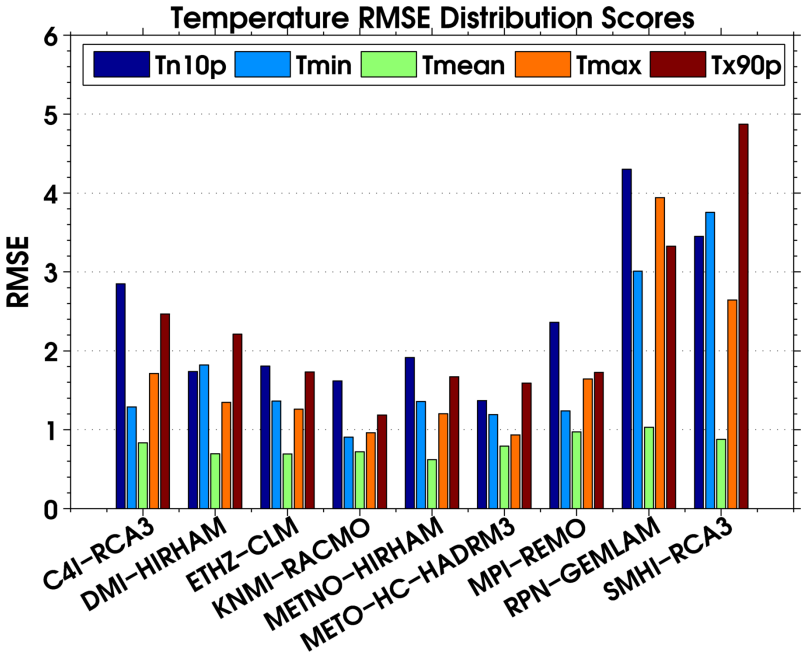

Figure 4.5: ENSEMBLES RCM temporal RMSE annual scores for Tn10p Tmin, Tmean, Tmax and Tx90p over 1961-2000

The relative model performance (Figure 4.4) based on these five temperature distribution variables can give a further indicator of the independence of information provided. Models which perform poorly or well at one extreme tend to do so also at the other tail of the distribution, but as highlighted earlier the pattern of relative model performance changes little from Tmax to Tx90p and Tmin to Tn10p. Relative performance for Tmean can be misleading, for example METO-HC-HADRM3 scores as a middle ranking model for Tmean, yet ranks in the top three in the distribution tails, outperforming all other models for Tn10p. A recommendation based on these results is to evaluate RCMs in their performance either for Tmax/Tmin or Tx90p/Tn10p, as the information gained from one is similar in character to that of

the other.

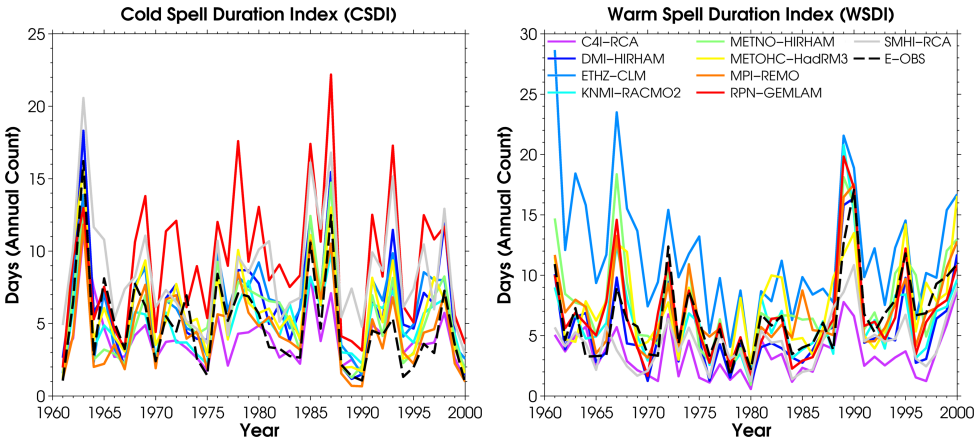

Figure 4.6: Europe spatial mean temperature CSDI and WSDI time series for ENSEMBLES RCMs, E-OBS observations black dotted line.

Persistence of extreme temperature events is assessed with the cold spell duration index (CSDI) and warm spell duration index (WSDI) (Figure 4.6). These are calculated from bias-corrected RCM data, as suggested by Sillmann *et al.* (2014), for each gridpoint using observationally derived Tn10p and Tx90p percentiles. Bias-correction is used to ensure that the indices are not simply evaluating model temperature biases relative to the observations, but are providing an analogous quantity to be compared to observed persistence rates. For example, WSDI calculates the annual count of >6 consecutive days in which the simulated maximum temperature exceeds the observed 90th-percentile of Tmax. If model biases are not removed, then if the RCM is cold biased, say, then the index will suggest that the model's persistence rates are too low, when in fact the simulation rarely exceeds the Tx90p threshold. When the bias is removed for each gridpoint and calendar day, the true persistence rates can be ascertained. To bias correct the RCM data, quantile mapping is employed whereby each RCM grid-point timeseries is transformed such that all percentiles match up to observed percentiles. By doing this, any intrinsic model bias is removed whilst retaining persistence characteristics, which CSDI and WSDI measure. Bias correction is no utilised however for FD, ID, CWD and CDD, because it is considered that since these indices are calculated relative to absolute values, instead of arbitrary percentiles, then they are more readily comparable to the corresponding observational index. The results indicate that the RCMs assessed reproduce both CSDI and WSDI reasonably well, although for some models large differences to the observed rates are identified. ETHZ-CLM consistently simulates longer periods of warm temperatures, whereas C4I-RCA3 displays the opposite characteristic. However, these indices provide somewhat erratic performance infor-

mation over the 40-year time period when compared to all other temperature and precipitation extreme indices, raising difficulties in determining *prime facie* which RCMs are best or worst performing. The construction of these persistence indices requires at least 6 consecutive days of a particular event to occur to contribute to the overall annual counts, which generates an inherent degree of chance. It is important to note that these indices were not originally intended to be used for the evaluation of climate model simulations, but for studies focussing on observational data. Therefore it is likely that another type of persistence index construction would provide clearer differentiating information on the models considered, although WSDI and CSDI do at least show that the RCMs are simulating similar annual count magnitudes to that derived from the observed data.

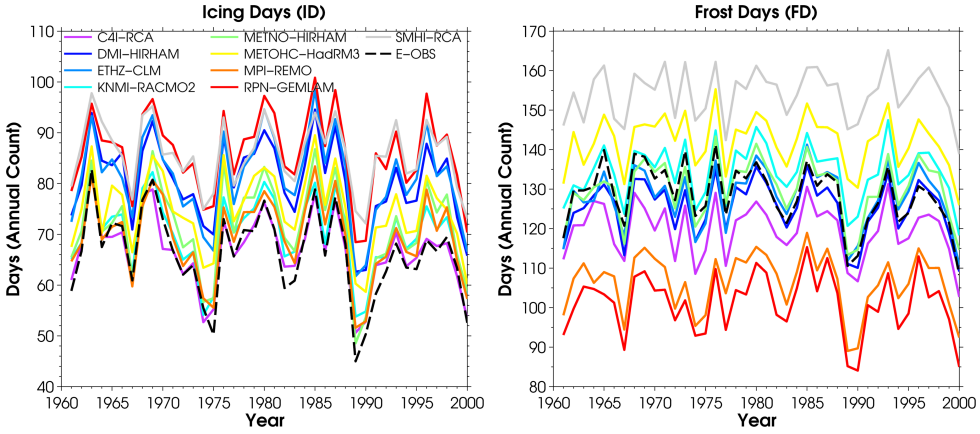

Figure 4.7: Spatial mean temperature annual time series for Icing Days and Frost Days Indices, E-OBS black dotted line.

The final temperature extreme indices considered here are the absolute threshold frequency quantities Icing Days (ID) and Frost Days (FD) which are calculated as the annual count of days in which Tmax<0 and Tmin<0 respectively. The RCMs are systematically cold in the simulation of maximum temperatures and, for all but two models (METO-HC-HADRM3 and SMHI-RCA), warm for minimum temperatures. For ID, this cold bias is seen in the higher than observed annual counts of very cold days, with the two coldest models RPN-GEMLAM and SMHI-RCA giving the largest counts. FD similarly reflects the information found from consideration of Tmin, with the warmest model RPN-GEMLAM giving the lowest counts and the coldest model, SMHI-RCA, the largest. One can conclude therefore that RCM performance information from ID and FD, at least for a general metric not considering smaller sub-domain features of RCM output, is not dissimilar to that taken from considering Tmax and Tmin alone.

Precipitation Variables

A common feature of regional and global climate model simulations is the systematic overestimation of precipitation levels, primarily caused by the lack of explicit convection resolving schemes (Kendon *et al.*, 2012). For RCMs, results from the European PRUDENCE multi-model ensemble experiment identified a tendency for models to have particularly large biases for both larger magnitude and low level precipitation events (Boberg *et al.*, 2009). The RCM's simulations of precipitation evaluated here from ENSEMBLES show an overestimation of daily precipitation magnitudes at all percentiles (Figure 4.8). Observed data shows that on average over 45% of days have negligible or no precipitation levels, in contrast to the tendency of RCMs to be dry on less than 30% of days. Additionally, the RCMs

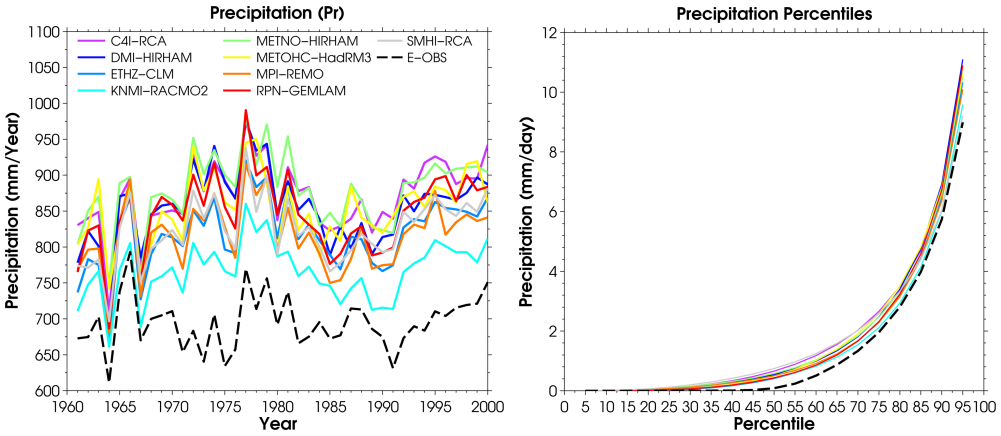

Figure 4.8: Spatial mean precipitation annual time series and percentiles for ENSEMBLES RCMs, E-OBS black dotted line.

maintain this over-simulation into the higher percentiles, although the average RCM percentage error is reduced from 500% at the 50th percentile to 14% at the 95th percentile (The error of RCMs at the 50th percentile error is approximately 5 times the observed values; 0.4mm/day for RCMs compared to 0.08mm/day observations). The best performing RCMs, KNMI-RACMO and SMHI-RCA, whilst exhibiting the low percentile 'drizzling', show good replication of observed extreme events above the 85% percentile. Total rainfall amounts for the ensemble (left) are overestimated by ∼20%, with a large discrepancy between observed levels and the best ensemble member, KNMI-RACMO. These biases are carried forward into evaluations of Consecutive Dry and Wet Days (CDD and CWD respectively) (Figure 4.9). CDD equates to the maximum annual count of days in which precipitation is less than 1mm, CWD the maximum annual count of days where precipitation is greater than 1mm. The general characteristics of the ensemble are an underestimation of

persistent dry days and an overestimation of wet days. Information can be ascertained with these two indices beyond what is found from consideration of Pr alone. SMHI-RCA (grey), for example, has the largest error for both CDD and CWD, yet performs as the second best model overall for extreme events, with one of the lower precipitation totals for high percentiles. This indicates that although the RCM simulates a large number of wet days, these days are of a much lower

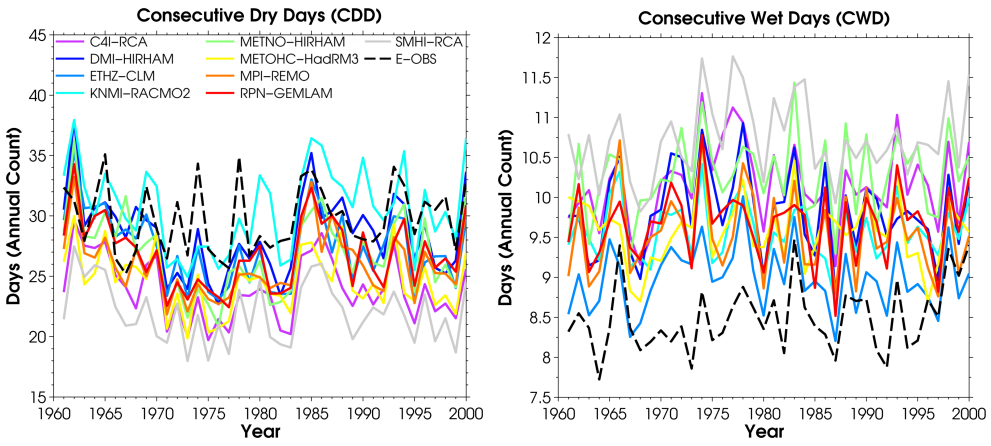

Figure 4.9: Spatial mean Consecutive Dry Days (CDD) and Consecutive Wet Days (CWD) extreme index annual time series for ENSEMBLES RCMs, E-OBS black dotted line.

magnitude than for the other ensemble members. Model performance information from either CDD or CWD tends to be repeated by the other index, which is likely due to the mirrored construction of the two, however since they are persistence measures based on a maximum period this is not necessarily guaranteed.

The high percentile precipitation events are sampled by four extreme indices (Figure 4.10), covering both frequency and magnitude: R10mm, R20mm (annual count of days Pr>10, 20mm respectively), and Rx1day, Rx5day (maximum 1-day, 5-day precipitation total respectively). The ensemble has a wide range of performance, with RCM relative performance generally consistent across both the 40-year period and each of the four indices. As with the other precipitation variables, SMHI-RCA and KNMI-RACMO show the lowest errors as these indices sample the higher magnitude percentiles. The identification of the systematic overestimation of extreme precipitation frequencies and magnitudes is seen in all four indices, however the merit of utilising the full set of indices given their high degree of correspondence is questionable. At least for this ensemble and domain, these precipitation indices are likely providing repeated information, although there are more differentiating features to extreme rainfall events if one considers spatial

and seasonal aspects of the RCM simulations. For example, larger magnitude precipitation amounts can occur in smaller concentrated regions and at different times of the year, both of which are not well sampled when taking spatial and annual averages. The following section in this Chapter considers this specific question in further detail as to what degree model performance changes when considering different time and spatial domains.

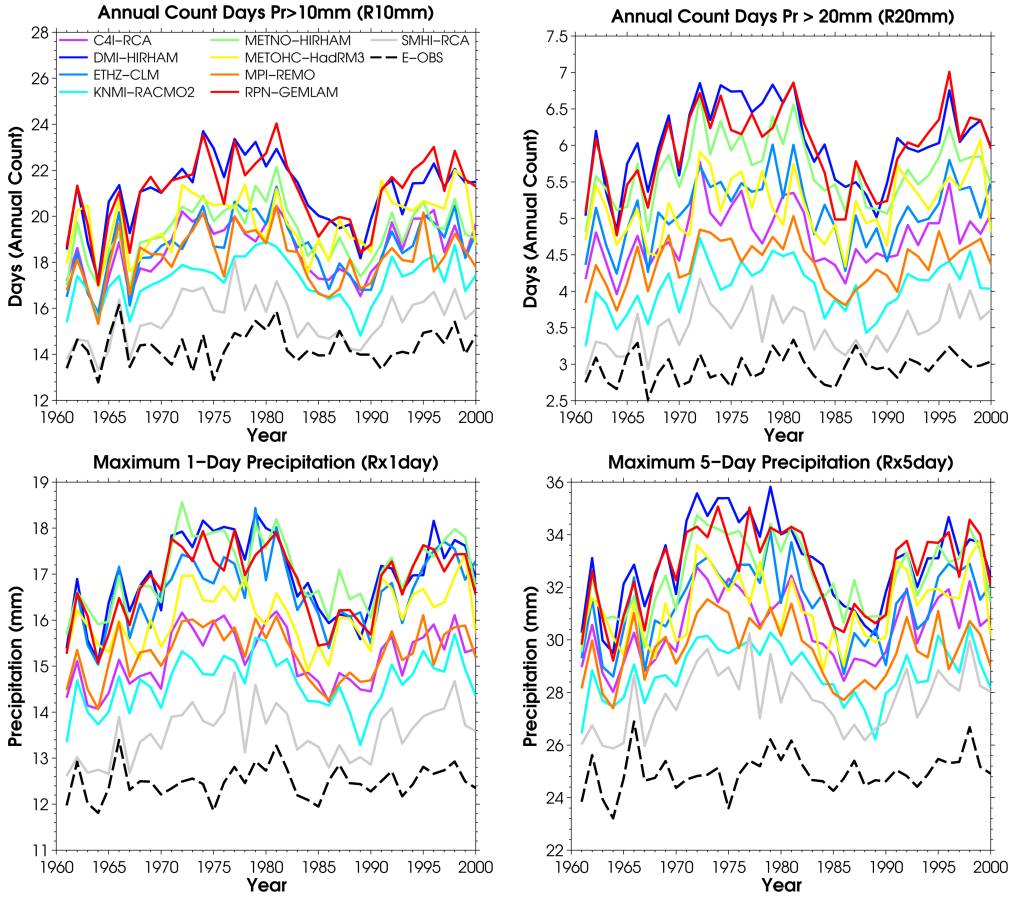

Figure 4.10: Europe spatial-average R10mm, R20mm, Rx1day and Rx5day time series for ENSEMBLES RCMs, E-OBS observations black dotted line.

The overall performance of the nine RCMs considered in this analysis, evaluated by assessing the mean spatial climatology RMSE for all 16 variables, can be viewed in a relative performance plot (Figure 4.11). A relative measure provides a means of identifying the strengths and weaknesses of each model provided that the ensemble is not systematically heavily biased (if all the models are poor, then the 'best' model from this perspective will not necessarily be of low error) or narrow in its distribution of performance (if all models are performing well, a relative measure may be misleading in penalising a model with low error). When assessing

performance over a large number of variables however, this issue will likely be less problematic to gain an overview of performance. Of the ENSEMBLES RCMs KNMI-RACMO2 scores consistently and unusually as one of the top three RCMs over all variables; all other RCMs are either poor with respect to temperature variables and strong in precipitation variables or vice versa. SMHI-RCA3 for example performs strongly for mean and extreme precipitation yet is poorly performing in simulation of temperature with a systematic cold bias. This raises a question as to what variables are most important to select models for assessing the changes to precipitation due to climate change, since this RCM is an example where the simulation is apparently skilful but possibly for the wrong reasons, and may not correctly respond to a warming climate. It is clear however that some of the variables are reproducing similar information, especially in regard to precipitation extreme indices, and therefore any final set of variables should attempt to reduce this redundancy, an issue that shall be returned to at the end of this chapter.

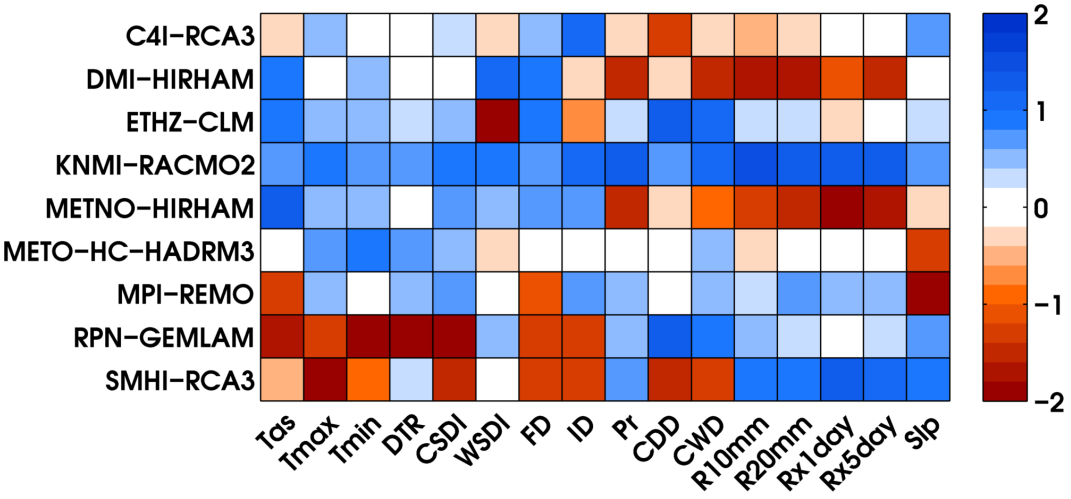

Figure 4.11: Relative performance of ENSEMBLES RCMs representing mean annual climatologies for 16 variables/extreme indices. For each variable, the RCMs metric values are normalised in the usual way; subtracting the mean of the metric values and dividing by the standard deviation. This leaves metric values centred at zero, and these values are what is represented here, with white values indicating RCMs that are middle ranking in performance, blue colours higher performing and red worse performing.

Variable Sensitivity: Conclusions

Given the wide range of possible variables for assessing RCMs it is interesting to note the similarities between some of the variables selected for this investigation. For temperature variables, it is valuable to evaluate models over the whole temper-

ature distribution since mean temperature metrics do not capture RCM simulation characteristics sufficiently well. This is made clear in the extremely low representation of Diurnal Temperature Range for RPN-GEMLAM, which scores reasonably well in its Tmean. Extreme temperatures in the form of the 90th-percentile of Tmax and 10th-percentile of Tmin although used in some extreme indices do not necessarily provide additional information over simply Tmax and Tmin, although the magnitude of biases does grow towards the tails of the temperature distribution. Temperature extreme indices are found to either provide similar information to the standard variables or somewhat random model performance in the case of persistence measures CSDI and WSDI. Precipitation variables similarly are found to exhibit common shared behaviour, reflecting the RCM's systematic overestimation of precipitation totals. Sea-level pressure as a separate variable to the other 15 not unexpectedly provides a further classification of RCM skill which is somewhat independent.

Overall, the degree to which model performance changes with respect to variable choices for RCM metric assessments depends on what variable is changing. For most temperature and precipitation variables, changing from within that group to a similar variable (e.g. mean temperature to maximum temperature, or mean precipitation to R10mm) does not change the resulting conclusions of RCM skill. However, changing from one group to the other will cause larger sensitivities in metric assessments. This is primarily due to the fact that for this ensemble and European domain, models that perform well in temperature do not generally do so for precipitation and vice versa. One would expect a reasonable degree of congruence between models that simulate temperature well and those that simulate precipitation with skill. If that were the case then the overall sensitivity to a change from temperature to precipitation variable would be lower. This conclusion is one which may be tested for another domain and ensemble to see whether these results are transferable elsewhere.

4.3 Sensitivity to Choice of Temporal/Spatial Domain

Once the assessment variables are considered and chosen there remain decisions regarding the temporal and spatial domain over which an RCM should be evaluated. 'Domain' is taken to be the three dimensional window (two spatial, one time) in which model simulations are compared to observations. Temporal domains of

annual means are commonly used (e.g Murphy *et al.*, 2004; Gleckler *et al.*, 2008; Reichler and Kim, 2008) to give a general characterisation of model performance in simulating the average climate. Alternatively, seasonal information can be used to identify how well models are replicating different processes most prevalent or relevant at that time of year. Time domains can also be chosen from different segments within long simulation runs of several decades offering further avenues for assessing performance. The possibilities for spatial domains are similarly numerous. From the starting point of evaluating performance over the whole regional area RCMs can be tested over the land only area or sub-domains either of homogeneous climatic properties or country specific focus. These choices altogether in choosing an RCM evaluation domain for use in a performance metric raise questions as to their influence on the resultant model skill scores and whether different results could be ascertained through a different method.

The ENSEMBLES RCMs having been tested in their performance in mean annual temperatures in the first section are now evaluated in their replication of seasonal and spatial variability for temperature and precipitation. Large temperature biases occur in all seasons (summer and winter observational mean and model biases shown in Figure 4.12). Common systematic errors are strongest in winter months, with all nine RCMs simulating a warm temperature bias in north-eastern Scandinavian regions, and a cold bias in southern areas, particularly the Alps. In summer the RCM characteristics are generally reversed, with a common warm bias in eastern and southern regions and cold in northern areas. The RCMs perform best for the British Isles and central Europe for both summer and winter with biases less than 1°C occurring in most cases. These large changes in RCM performance both over Europe and the seasons lead to differences in absolute and relative model performance (Figure 4.13). The lowest RMSE skill scores arise when assessing annual means, the reason for this likely being a manifestation of model biases of opposite signs from summer and winter months cancelling out. This would suggest that annual averages may produce misleading impressions of model performance over the whole annual cycle in cases such as these where biases happen to cancel out. A model of the exact opposite characteristic of a more uniform bias may have a similar magnitude of error throughout the year but appear, through such a metric, as exhibiting lower performance simply because the biases are of the same sign.

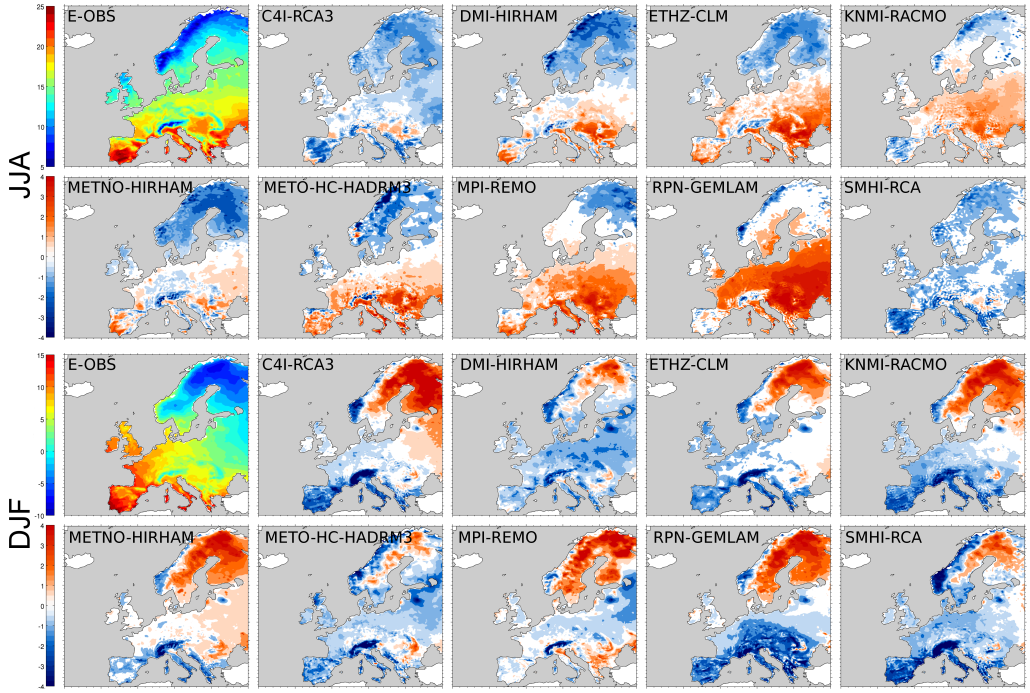

Figure 4.12: E-OBS summer/winter temperature mean climatology (°C) and ENSEMBLES RCMs mean temperature bias (°C) covering 1961-2000.

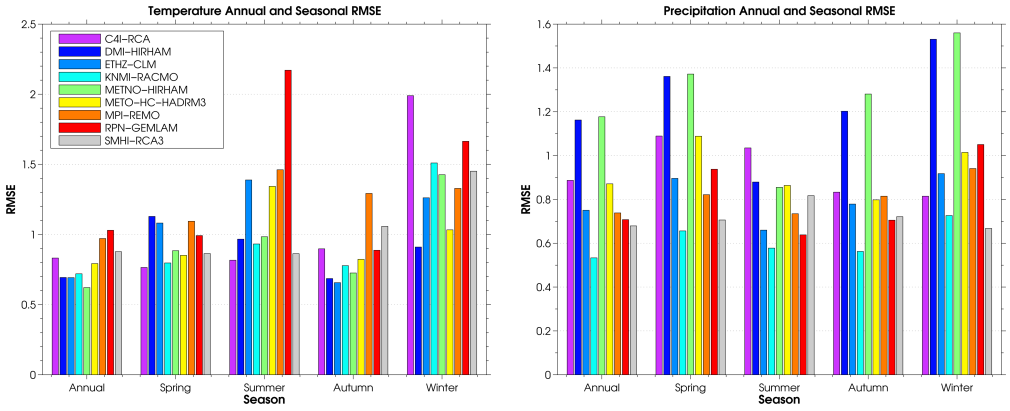

Figure 4.13: ENSEMBLES RCMs annual and seasonal temperature and precipitation mean climatological spatial RMSE skill scores for European total domain

The weakness of annual metrics does not occur in assessment of precipitation skill, as the RCMs present a uniform zero to wet bias in all seasons and regions (Figure 4.14). The characteristics of precipitation spatial variability in Europe are more affected by the time of year than temperature, and thus lends itself more to seasonal evaluations particularly on localised scales. Western coastal regions receive higher precipitation amounts in winter, especially in more mountainous regions such as the Highlands of Scotland and Norway, due to the occurrence of a stronger jet stream (Lavers *et al.*, 2013). In summer however, convective rainfall

due to the increased average temperatures is the primary process leading to higher precipitation amounts in central European regions (Berg *et al.*, 2009). In winter the RCMs tend to present lower precipitation percentage biases in the western coastal regions where high precipitation amounts form, suggesting that the RCMs are simulating inter alia the larger scale westerly extratropical cyclones with some skill. The more eastern areas see higher percentage biases where the observed precipitation totals are low, indicating that the models are simulating too much low intensity precipitation in these regions as noted in Figure 4.8. This characteristic is also prevalent in summer months where the highest biases occur in regions of lowest total rainfall. The fact that the models are uniformly biased regardless of season leads to the annual metric being more representative of model performance over the whole annual cycle than for temperature, with smaller changes in the general distribution of model performance over all seasons (Figure 4.13).

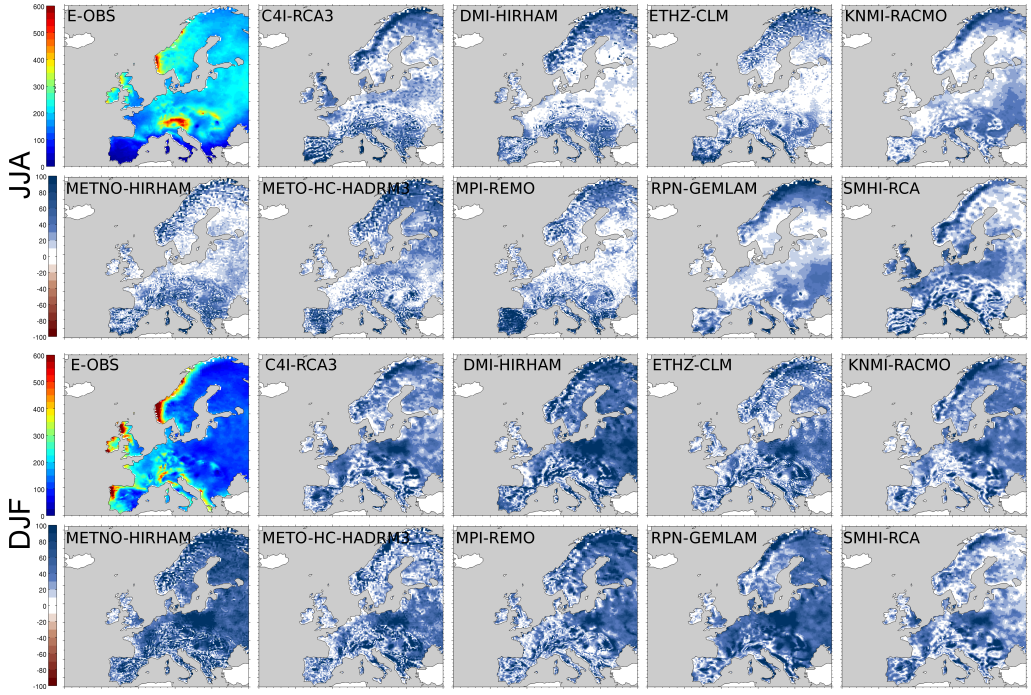

Figure 4.14: E-OBS summer/winter precipitation mean climatology (mm/season) and ENSEMBLES RCMs precipitation percentage bias (-100% - 100%) covering 1961-2000.

Assessing the RCMs on smaller sub-domains enables a quantitative characterisation of the spatial variation in simulation errors (Figure 4.15), not able to be ascertained through total area assessments. For temperature, a wide range of skill scores are found reflecting the tendency for RCMs to perform better towards the western and central regions and worse in northern, eastern and Mediterranean

areas. To make a judgement as to the 'best' model over one country scale region therefore may be misleading as to the performance over other regions. For example, in Eastern Europe C4I-RCA and SMHI-RCA3 are the two best performing RCMs, whereas in the Iberian Peninsula they are the two worst performing. Such results may provide an argument against selecting RCMs based on their performance over these smaller regions alone for use in further investigations such as climate change impacts assessments, since potentially useful performance information is lost with the narrowing of spatial domain. On the other hand, those models with high errors over a region of interest may simply not be realistic enough for purpose, and such smaller sub-domain evaluations could be of use to differentiate between RCMs in this regard. The precipitation annual sub-domain assessments show a more consistent pattern of model performance over Europe, with the relative performance of RCMs maintaining their general distribution over all sub-domains. Therefore for mean precipitation selecting models based on their overall total European domain performance is likely more robust and representative of the performance of the ensemble in a sub-domain. Individual model absolute performance will change depending on the region, but for identifying the more skilful models (at least for this ensemble and climatic region) the total domain metric may be sufficient.

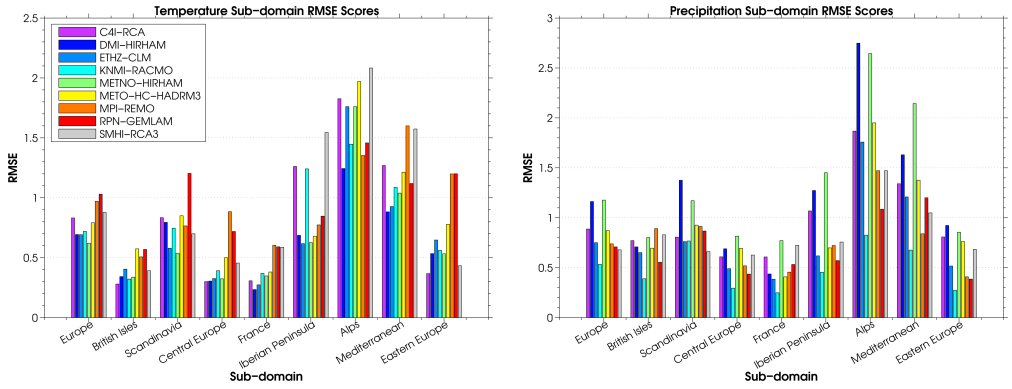

Figure 4.15: ENSEMBLES RCMs annual temperature and precipitation Europe and sub-domain spatial RMSE skill scores.

RCM ensembles with longer simulation hindcasts of several decades can be tested over the whole simulation, as was the case in the previous model evaluations, or for shorter sub-intervals. The sensitivity of model performance assessments to changes in the choice of temporal domain in this second sense, other than selecting times of the year, is tested here by evaluating RCMs over all 10-year time windows between 1961-2000. This provides an indication of the robustness of shorter interval model evaluations and also that of the full length simulation assessment.

Minimum, mean and maximum temperature RCM data are compared to the respective observational data by computing temporal RMSE values for each 10-year gridbox timeseries then taking the spatial mean of the resulting grid (Figure 4.16). Model performance varies by up to 18% for Tmin over 1961-2000 although for Tmax skill scores vary by a smaller range of 2-7%. An interesting result is that the models within the ensemble tend to increase and decrease in performance in unison for Tmin, Tmean and Tmax. The cause of this is not entirely clear, although most likely it could be due to biases in the ERA-40 driving data relative to E-OBS. However the benefit of this phenomenon is that despite model performance varying over the 40 year time-frame the relative performance between models is approximately constant, with the lowest and highest error models remaining so respectively. The implication from this is that RCM performance information for temperature is robust in as far as relative performance is concerned for any choice of time-frame. The variations in skill score for Tmin and Tmean can be substantial, although less so for Tmax, however this is not an issue when the ensemble behaves in this uniform manner.

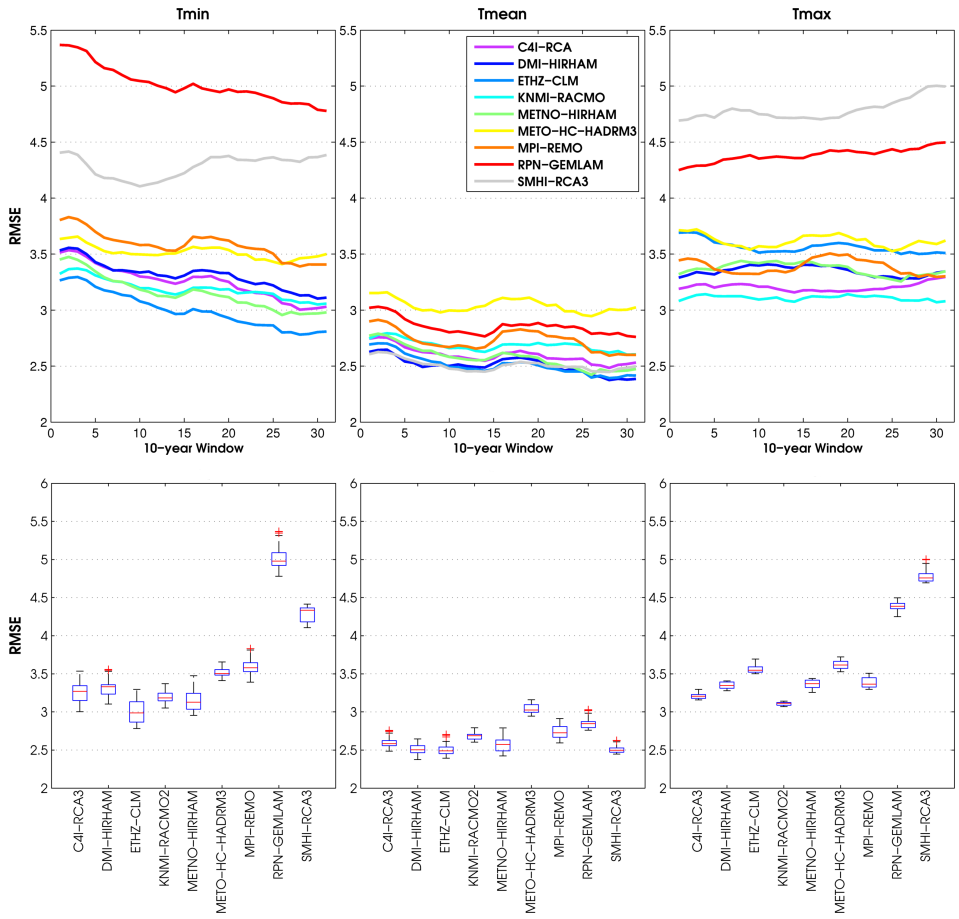

Figure 4.16: Temperature average temporal RMSE skill scores (top) and range of score output (bottom) evaluated over all 10-year time windows from 1961-2000.

Precipitation temporal time-window sensitivity is lower than for temperature with model skill scores varying up to 7%. Similarly to temperature temporal skill scores the ensemble performance distribution moves almost in unison, with very little variation in relative performance over the 31 time domains. Robustness for precipitation metrics over different time-frames therefore is high both for relative and absolute model performance. Model assessments in respect of mean climatologies can be used confidently in generalising to wider time domains. However, with respect to metrics assessing extreme events the full time-frame is recommended to ensure the indices are robust given the infrequent nature of the variable type.

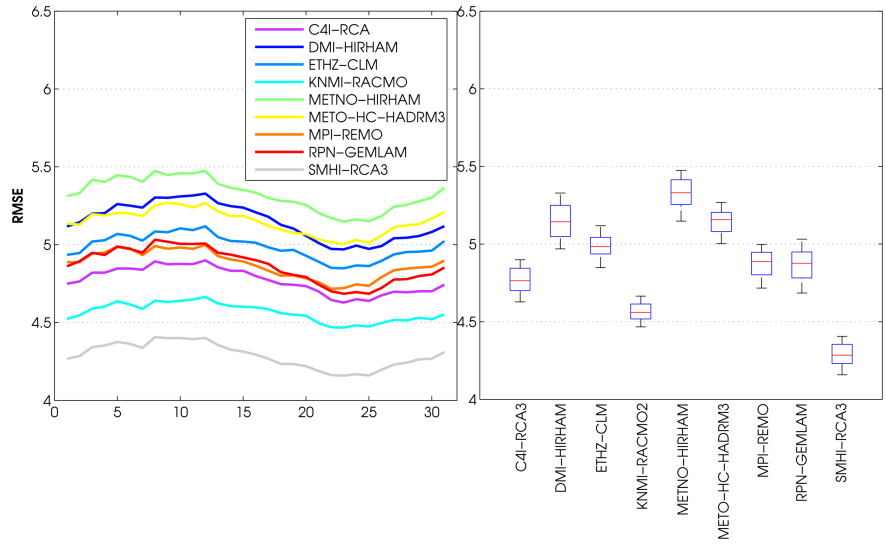

Figure 4.17: Precipitation spatial RMSE skill scores and range of score output evaluated over all 10-year time windows from 1961-2000.

Temporal/Spatial Sensitivity: Conclusions

RCMs are complex to evaluate given their three dimensional nature. Summarising performance information into scalar quantities necessitates a loss of information in both time and space, and therefore it is important to consider whether a given metric retains the required information or not. With regard to the choice of temporal domain, there are two types used in a model evaluation setup: a choice of seasonal or annual timeframe, and the choice of full length or segmented RCM simulation runs. First, on seasonal/annual sensitivity there is substantial variation in assessed model performance between seasons, whereas the annual mean tends to dampen this information. To what extent the information is lost through this time averaging depends on the characteristics of the model biases as discussed; more uniform biases will be preserved in annual absolute value metrics, whereas if errors cancel out annual bias will not be representative of the whole annual cycle. It may be possible to construct statistics which can preserve this information, but for those considered here it is desirable to include seasonal information (summer/winter is likely sufficient) to ensure a more comprehensive assessment of model behaviour. RCM assessments on shorter segments of long decadal simulations are found to be robust, providing confidence to evaluate models where, for example, limited duration observational data exists.

The second aspect of spatial sensitivity considered here is found to depend on the variable and domain of interest. Precipitation scores tend to preserve relative

performance over all sub-domains of Europe, although the absolute magnitudes of error do vary at least to an equal degree to that of temperature. The consequence of this is that for precipitation 'good' models retain this character throughout the simulation domain, and as such one can have confidence in any single domain assessment that those conclusions will hold elsewhere. For temperature however, this does not apply to the same degree although there are few anomalous results suggesting that although assessing models over all domains may be prudent, a total area assessment will not be overly misleading as to the general traits of each RCM. To what extent these conclusions hold beyond this ensemble, European region and variables is an open question however.

4.4 Sensitivity to Choice of Statistic

Spatial Pattern and Standard Error Statistics

This section investigates the sensitivity of metrics using spatial pattern and standard error statistics. To do this, metrics incorporating these statistics types are applied to mean climatological model-observation difference patterns over the whole European domain and the eight Rockel sub-domains. The following analysis aims to test how robust these metrics are to changes in the statistic when the same climatic field is evaluated. Table 4.8 shows the type of raw metric output produced by the seven metric variations, in this example the RCM skill of replicating the CDD (consecutive dry days) extreme index mean climatological spatial pattern over 1961-2000. Although this quantitative information can be useful in identifying better or worse performing models, when using more than one statistic there may be disagreement which leads to the question as to which statistics are more robust or meaningful. This section aims to answer these questions and recommend statistics producing independent information for use in RCM assessment projects.

Each statistic is constructed to quantify RCM performance as a function of the model error relative to observed dataset for each gridpoint. These functions map a non-zero positive error value to an output skill score number range which is different for each statistic type. The statistics chosen range over -∞ to 1 (e.g. Index of Agreement), 0 to ∞ (e.g. RMSE) or -1 to 1 (Correlation). Furthermore, the direction of increasing performance may not be the same between two statistics; RMSE decreases with smaller errors, whereas the level of correlation will increase. The

RCM	RMSE	MAE	STD-R	IA	R	SSS	SSM
C4I-RCA3	8.46	6.20	0.83	0.85	0.84	0.53	2.05
DMI-HIRHAM	7.13	4.86	1.30	0.93	0.91	0.67	1.47
ETHZ-CLM	4.91	3.40	1.00	0.96	0.93	0.84	2.59
KNMI-RACMO2	5.89	3.84	1.22	0.95	0.93	0.77	1.25
METNO-HIRHAM	7.28	4.60	1.09	0.91	0.86	0.65	3.42
METO-HC-HADRM3	6.70	4.72	0.83	0.91	0.90	0.70	2.62
MPI-REMO	6.65	4.56	0.91	0.92	0.87	0.71	2.45
RPN-GEMLAM	5.17	3.45	0.93	0.95	0.92	0.82	1.88
SMHI-RCA	8.83	6.43	0.75	0.73	0.86	0.49	1.41

Table 4.8: Example of ENSEMBLES RCM metric scores for Standard Error and Spatial Pattern statistics for the CDD (Consecutive Dry Days) extreme index mean climatology calculated over the 1961-2000 whole European domain.

standard deviation ratio (STD-R), defined by RCM standard deviation divided by the observational standard deviation, is different to the other statistics in that perfect performance would give an output value of 1, but with values of 0.9 and 1.1 equating to lower performance due to the fractional nature of the function. When comparing output between these statistics, relative measures are used to account for the changes in units in the same way as relative model performance is derived in Figure 4.11. For the mean temperature climatology, relative model performance information is calculated to indicate the level of agreement between the seven statistics (Figure 4.18). All of the statistics but one, the standard deviation ratio (STD-R), produce similar metric output suggesting that this aspect can be assessed with a single statistic without loss of information. The STD-R evaluates the similiarity of spatial homogeneity over the spatial field, but does not agree with the six other statistics on the best or worst performing RCMs. It is more likely that this statistic is badly defined for assessing a spatial field, and is not considered in the final set of statistics, due to its deficiencies in assessing the location and magnitude of errors. To illustrate these two problems, consider the following two examples. For each gridpoint, all values are swapped with another random gridpoint; according to STD-R, the RCM is equally as well performing. Next, consider an RCM which is qualitatively very poor in replicating magnitudes, yet simulates a similar variation in amounts; according to STD-R this RCM will be considered as well performing. Although STD-R does assess simulation aspects not evaluated by the other statistics, these problems bring into question whether this statistic should be considered equally as robust. A combination of these issues is likely the cause of the discrepency between STD-R and the other statistic output, and therefore STD-R is not included in the final set of statistics. Standard Error and spatial pattern statistics are

analysed together as they can be compared on their assessments on one single field, whereas temporal statistics cannot be applied to this input data. Event Frequency and Temporal Variability statistics in a similar vein can be applied to the same input data (timeseries) and as such it is considered reasonable to assess these together.

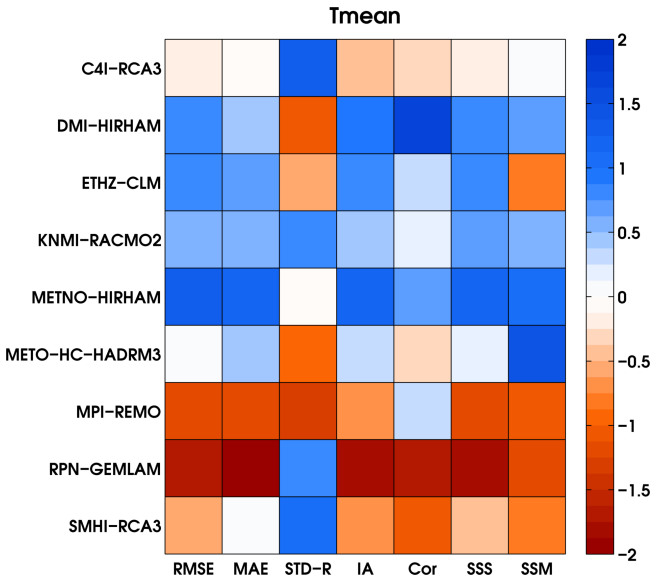

Figure 4.18: Relative model performance for mean temperature spatial evaluations using standard error and spatial pattern statistics. Blue colours indicate better performing models, red worse performing.

To investigate further the relationship between the chosen spatial pattern and standard error statistics, Principal Component Analysis is used on the normalised metric output over all 10 extreme indices (CDD, CWD, CSDI, WSDI, FD, ID, R10mm, R20mm, Rx1day, Rx5day) (Figure 4.19). This analysis is only applied to the extreme indices since they are considered to provide a wide range of input data sufficient to assess the sensitivity of metric assessments to changes in statistic. Care must be taken however in this type of analysis as two statistics could be interpreted to be similar by means of an artificial correlation; by pooling metric values from a range of variables with differing magnitude, the overall data on which the PCA is applied could lead to misleading results. To test the sensitivity of results to this possibility, the analysis can be performed by assessing not only all variables for a single domain, but also all domains for a single variable. Alternatively, the issue may be avoided by normalising input variable data before applying PCA. By normalising, such artificial correlations should not become an issue. Biplots are produced displaying the first two principal components and the corresponding loadings for the statistics. This format assists in showing relationships between statistic; the closer two vectors are the more alike they are. The first two principal

components account for between 76% and 86% of the total variation in metric output, which suggests that it is reasonable to interpret the data in a biplot as displayed, although the lower the total variance explained, the less representative the biplot of the overall metric output. The second remark is that, some statistics should be expected to produce output which is more closely related by construction to others, and this is present in the case of RMSE, MAE and the Spatial Skill Score (SSS). The small acute angle between their respective loading vectors in all domains indicates that they are highly correlated to one another, which suggests a substantial degree of redundancy if all were to be used. Additionally, these three statistics are the most representative of the majority of the variation in metric output given their close correspondence with the PC1 (x-axis in the biplot). This result however, is not overly surprising, since if the three statistics are indeed so closely related, then they would already account for a substantial portion of the total variance in the metric output, even before the variation in other statistic output is considered. In respect of PC1, all statistics are of the same sign in the loadings with very similar magnitudes, although RMSE and SSS have the highest in most cases indicating that they are most representative of the majority of the variation in the metric output. That all the statistics are correlated in the same direction in this first component indicates that the statistics are assessing the RCMs in a similar fashion; choosing one statistic over another in most cases is unlikely to produce altogether unrepresentative evaluations. However, the Spatial Skill Metric statistic shows little correlation (as seen by the 90 degree angle) to the approximate consensus of RMSE/SSS/MAE, and so it is possible that such an occurrence may happen. In respect of PC2, accounting for 13% of the metric evaluation variations, the largest loadings are held by SSM, and in some cases the Correlation statistic. This would indicate that these statistics are evaluating different aspects of the spatial pattern than the more error focussed statistics, but it is unclear to what extent the construction of SSM is a factor in these results.

Given that some of the statistics are known to be closely related to each other, the PCA analysis provides an opportunity to test these predictions. To begin with, RMSE and MAE is found to behave extremely similarly for all domains, which is not overly surprising given that they will only disagree for datasets with large extreme values within. The SSS likewise shows very close behaviour to that of RMSE/MAE for all domains, indicating that the variance in the data is very low. In some cases, such as Scandinavia/British Isles/Iberian Peninsula, the Index of Agreement behaves very closely to this set, yet for other domains such as the Alps,

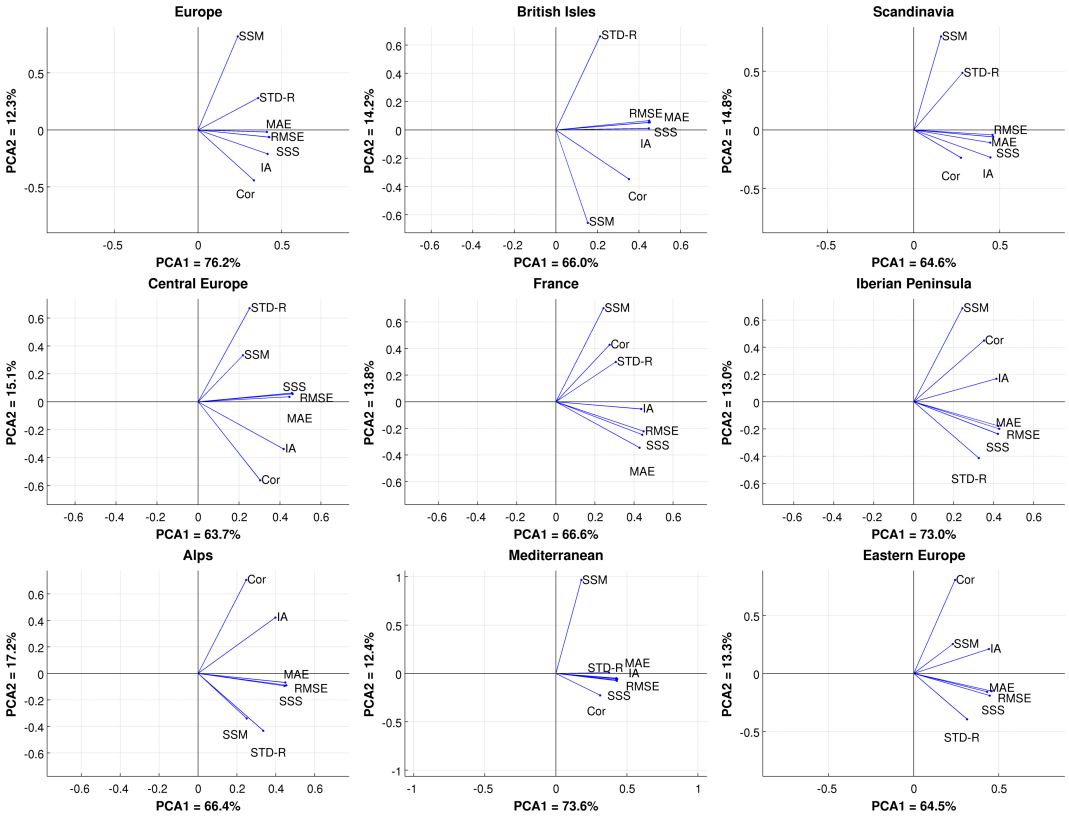

Figure 4.19: Principal component analysis of spatial pattern and 'standard error' metric output for total European domain and sub-domains. Output data has been normalised over all variables to account for differences in units and/or magnitudes of error.

Central Europe and Eastern Europe, this statistic disagrees more. This is very likely due to the 'potential error' normalising factor adjusting the metric values depending on whether RCM or observed variances are dissimilar.

To investigate the impact of the degree of metric sensitivity to changes in statistic on the relative RCM performance within the ensemble, model ranks are generated from the RCM evaluations of extreme indices mean climatological spatial patterns (Figure 4.20). Similarly to the Spatial Pattern and Standard Error statistic PCA analysis (Figure 4.19) the use of 10 extreme indices is considered sufficient to assess what the analysis is intended to determine: the degree to which RCM relative performance is sensitive to changes in statistic; it is unnecessary to apply this analysis to all variables. It is clear that there is a wide disparity in regards to the sensitivity of model ranks to changes in statistic. On the one end of the scale, CDD, CWD, CSDI and WSDI all give rise to RCM evaluations through varying the statistic with substantial disagreement as what the model rank relative to other ensemble members should be. Other extreme indices such as

R10mm, R20mm, Rx1day and Rx5day give in most cases 2-3 rank jumps with a change of statistic; larger magnitude rank differences when changing statistic are more unusual. The outcome of this is that spatial pattern and standard error statistic, although more constrained to the assessment of a single spatial field than other statistic types evaluating temporal variability aspects, can have a substantial impact on the inferred relative RCM performance within an RCM ensemble.

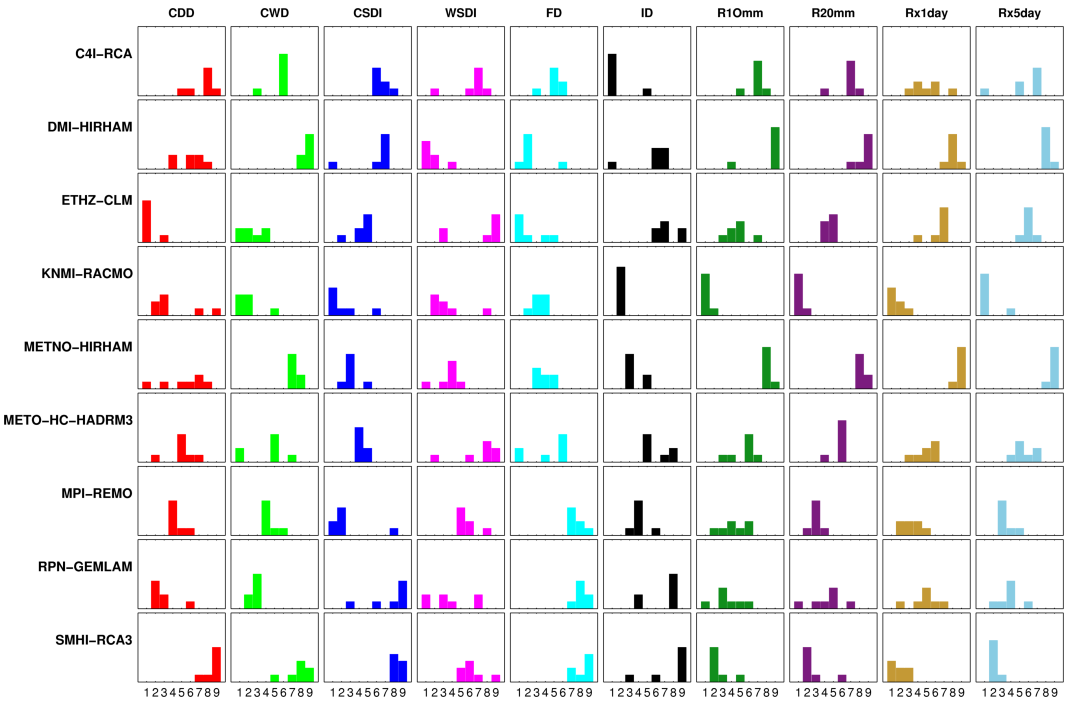

Figure 4.20: RCM Rank Sensitivity to changes in statistic for all extreme indices over the whole European domain. Bottom axis refers to model rank number from 1 to 9 (left-right) for each variable

Event Frequency and Temporal Variability Statistics

This section investigates the sensitivity of metrics utilising other, more specialised or purpose built statistics. These cannot be typically be applied outside of their intended application. They therefore are naturally not as closely related to one another as those used to assess spatial fields as in the previous section, although two statistics may well still be focussing on the same specific simulation aspect. Six distinct statistics are utilised in this section taken from Tables 4.1 and 4.7; the Annual Cycle Skill Score (ACSS), Annual Variability Metric (AVM), Interannual Variability Metric (IVM), Linear Trends Score (LT), PDF Skill Score (PDF) and CDF Skill Score (CDF). The relationship between these six statistics is assessed with PCA as

used earlier with the spatial pattern and standard error statistics (Figure 4.21). The overall total variance explained by the first two principal components in this case is much lower; between 55% and 69% is explained by PC1 and PC2 in the results. This limits the interpretive value of biplots, as much of the remaining variance

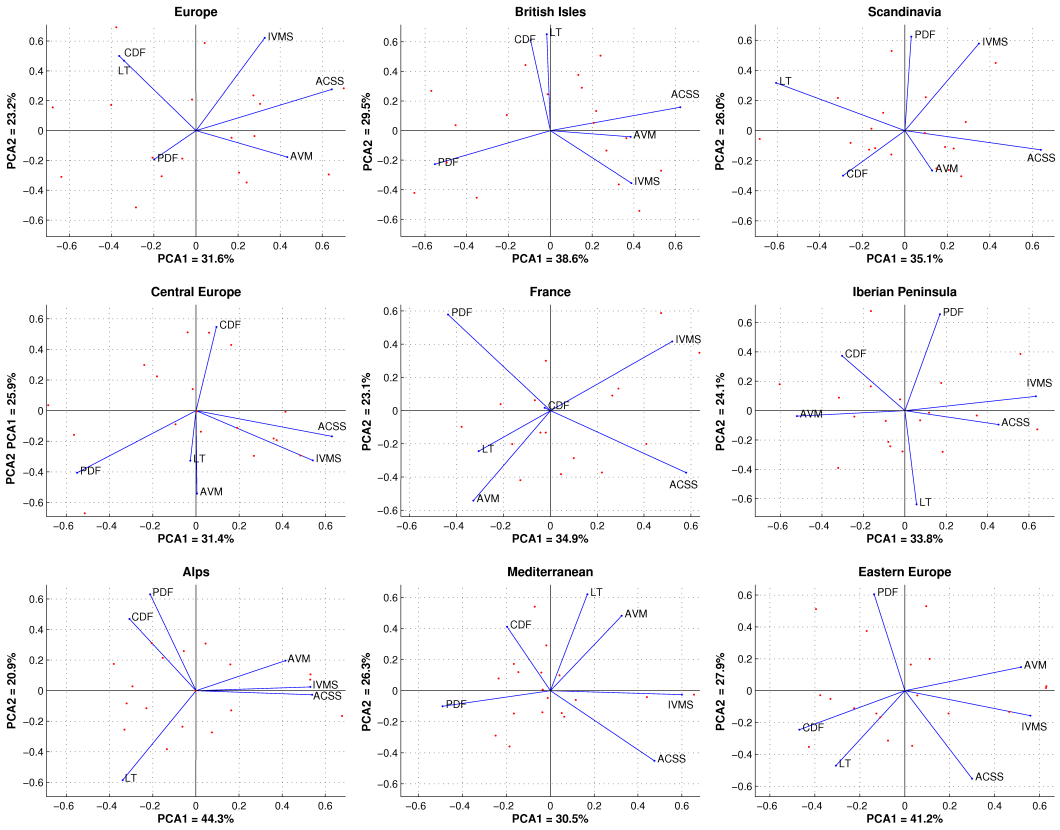

Figure 4.21: Principal component analysis of temporal variability and event frequency metric output for total European domain and sub-domains. Output data has been normalised over all variables to account for differences in units and/or magnitudes of error.

of the data is orthogonal to the 2-D plane shown. However, some general comments can be made. Firstly it is clear that the statistics produce very different metric output, with a substantial degree of strong negative correlations seen between different statistics, as seen by the near 180° angle between some of the loading vectors corresponding to each statistic. They are likely producing performance information which is of a more independent character than for statistics evaluating the same spatial climatological patterns. As a result they should all be included in the final set of statistics used to evaluate model performance. Additionally, it appears that the relationship between these statistics is not as consistent when evaluating RCMs over different domains as seen with spatial pattern and standard domain statistics. Beyond this however, it is difficult to draw more firm conclusions in part due to the

low level of explained variance in the first two components.

Statistic Sensitivity: Conclusions

The role the choice of statistic plays in RCM evaluations with performance metrics has been somewhat overlooked in the past. Although studies have evaluated climate models over a range of spatial and temporal aspects, they for the most part choose only one statistic for each individual evaluation task. The results of this section would suggest this practice may be overconfident. First, it must be noted that the assessment of the sensitivity of absolute metric output in this chapter is limited by the fact that only a small set of statistics is selected for use. Although a more exhaustive set of statistics may provide a more complete picture of how wide the range of plausible RCM assessments may be in regard to a particular aspect of interest, the limited set chosen in this case does not undermine the general point, and the outcome acts as an underestimate of the total true sensitivity of the statistic choice. The main conclusion is that the choice of statistic can have a low to substantial impact on the assessed model performance depending on the simulation aspect in question, be it spatial patterns (low) or temporal variability and event frequency (high). As seen in this second more sensitive group of statistics, there is low agreement between RCM performance assessments due to the fact that the underlying information on which the statistics are based is different, unlike those statistics investigating spatial patterns. It is highly unlikely that any two of this temporal/event frequency statistic type can be interchanged without loss of information. The main conclusion from this section therefore is that RCM evaluation studies seeking to include such a temporal dimension should consider additional statistics to reduce the likelihood of overconfident model scores.

The sensitivity of relative model performance is related to the underlying sensitivity of the absolute metric output, although there is a caveat. In most cases, if there is a high degree of agreement among the statistics as to the level of model performance over all ensemble members, then relative model performance as reflected in their ranks will be more stable and less open to uncertainty. This is important if for example some models are to be discarded for the purposes of a particular investigation due to their bottom ranking scores; if such scores are demonstrably robust then the basis for this procedure is strengthened. If on the other hand model ranks are fundamentally uncertain, yet in practice only one statistic is used, the possibility is open for models to be eliminated unfairly. The

caveat relates to if there happens to be several closely related statistics used. It may be the case that although the absolute sensitivity is low for a particular assessment, the relative performance may vary in a manner which could be reasonably interpreted as suggesting that the underlying metric sensitivity is high, when in fact it is not the case. This is the opposite to the case in which one model may be extremely good or poorly performing in relation to other models, yet this will not be communicated in relative model ranking scores. One should therefore consider the absolute scores alongside any ranked output to avoid misleading interpretations of the actual model performance. The overall conclusion is that if RCMs are tested with a limited set of statistics, any generated ranking scores should be treated with caution, as they are unlikely to be the only possibility with different statistic choices.

4.5 Metric Redundancy

The previous sections, in particular 4.2, did not analyse fully the relationships between the full variable set, but instead looked more closely at the relationships between variables of certain smaller groups. A full analysis of all variables is required, not least as it is desirable for the following chapter on metric combinations to refine which aspects of model performance are most crucial to capture. Three separate statistical analyses are undertaken to identify relationships between metrics used in the previous analysis and to reduce the number of variables taken forward to the next analysis on metric combinations. They are: Correlations, Cluster Analysis, and PCA. The methods are applied to annual mean climatologies for the 16 variables utilised in the previous investigation of metric sensitivity. For the following cluster analysis and PCA, evaluations are calculated from the RMSE mean climatology for simplicity. More detailed investigations could be undertaken for other statistics, but in this case the issue discussed previously in Section 4.4 relating to pooling of datasets of different magnitude does not arise. Correlations between metric results (Figure 4.22) show a distinct split between temperature and precipitation related variable groups. Strong anti-correlations are present between temperature and precipitation scores i.e. models performing well in temperature tend to perform poorly in precipitation. The RCMs systematically overestimate precipitation amounts over the European domain, for both mean and extreme events. Furthermore, the models on average underestimate maximum temperatures and for the most part overestimate minimum temperatures. Déry and Wood (2005) found a common anti-correlation between precipitation amounts and temperature glob-

ally in summer months, explained by the soil moisture-atmospheric feedback (Berg *et al.*, 2014). Therefore one would expect that those models overestimating minimum temperatures the most would as a result simulate lower precipitation levels. However, this is not the case with this ensemble. The best performing RCM for precipitation (with the least precipitation amounts) is systematically too cold even though one might expect this RCM to simulate more precipitation overall. Although the observed anti-correlation only applies in summer months, the seasonal temperature/precipitation correlations found in the metric scores are -0.41 in summer and -0.37 in winter, and as such the main result is not explained by this mechanism (or possibly that the RCMs do not adequately simulate this process).

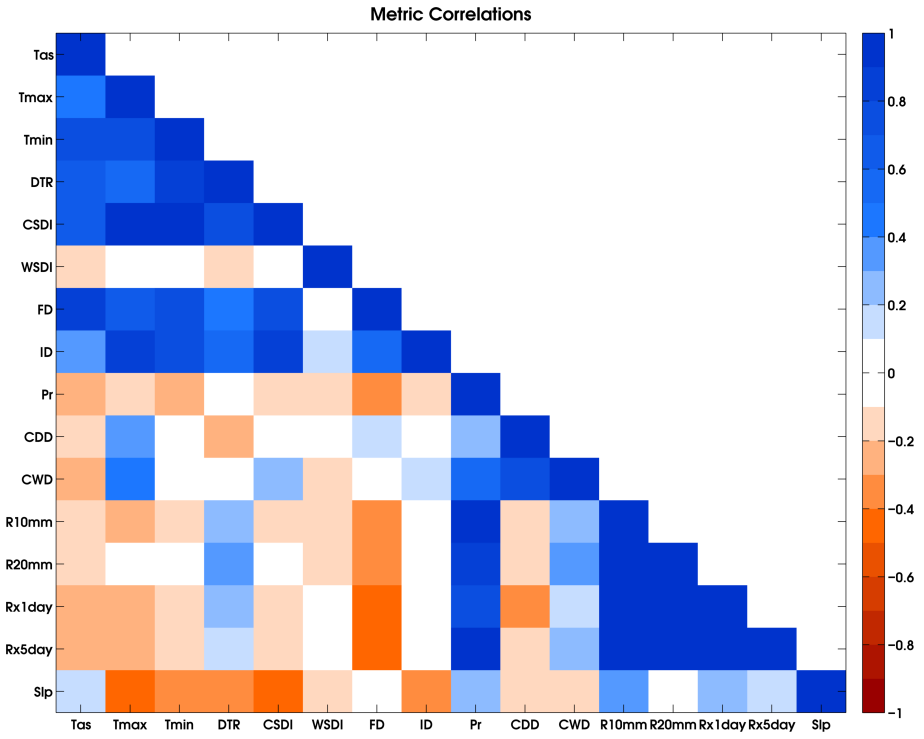

Figure 4.22: Correlations between ensemble performance metric output for temperature, precipitation, sea-level pressure and extreme indices.

Even so, this redundancy test suggests that there are three main groups of variables: temperature related, precipitation related and sea-level pressure. As a result the final set of metrics should span this range. The persistence variables CDD, CWD, CSDI and WSDI are most unique among the indices included, as they are least related to most of the other variables. These should be considered carefully when deciding the final set. Objectively setting an a priori threshold for reducing variable pairs, putting aside the issue of which to remove when a relationship is found and where to begin, is highly difficult.

The main result from the cluster analysis very much reflects the findings of the correlations, with a bifurcation of the variables into the two main groups. Close relationships between individual metric scores are seen, for example between Icing Days and DTR, however it is difficult both to ascertain the causes behind these solely based on this test due to the lack of information as to the characteristics of each model and the ensemble as a whole for each variable. As is the case with correlations, identifying which variable to remove from each pair is somewhat subjective.

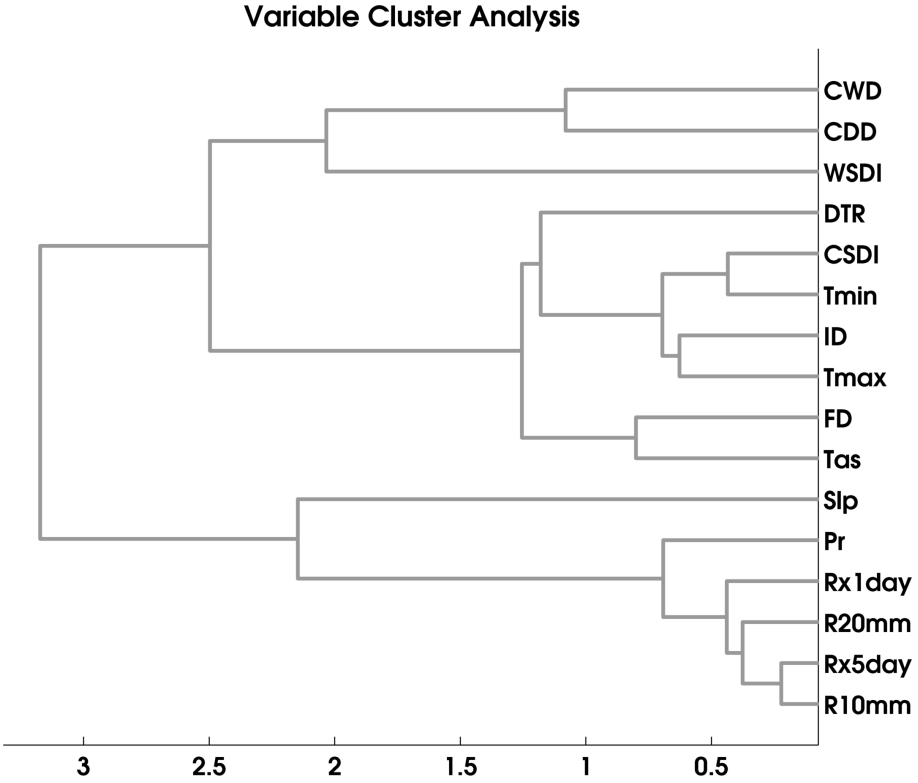

Figure 4.23: Cluster analysis calculated from ensemble performance metric output for temperature, precipitation, sea-level pressure and extreme indices. Linkages nearer to the right indicate a closer relationship between metric output.

The first three Principal Components (Figure 4.24) account for 82% of the total explained variance between metric scores, and therefore are considered sufficient to identify the primary characteristics present. The first component, accounting for 42% of the variance, displays the split between temperature and precipitation metric scores. High scoring models in Pr will tend to score highly in all the extreme precipitation indices in addition to a weaker score relationship for precipitation persistence and SLP. High scoring models in Pr also tend to score poorly in all temper-

ature variables as shown by the correlations test (Figure 4.22). The scores for PC 1 show that the two HIRHAM RCMs and RPN-GEMLAM and SMHI-RCA display this anti-correlated pattern both strongly, whereas other RCMs such as ETHZ-CLM and KNMI-RACMO2 are more evenly performing.

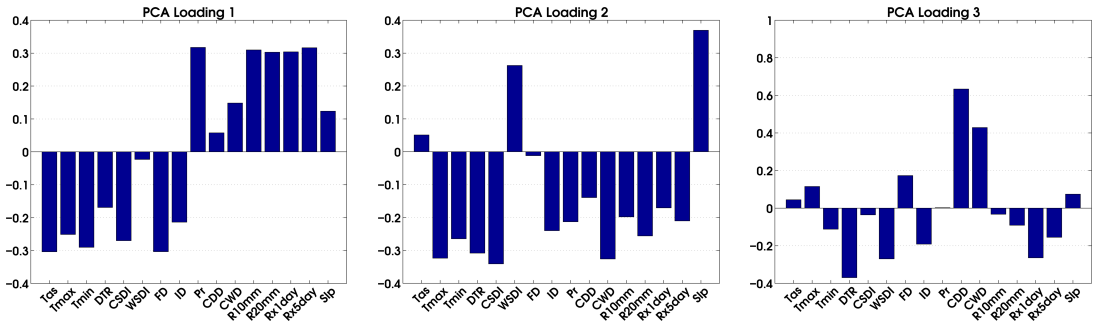

Figure 4.24: PCA loadings for the first three principal components for ENSEMBLES RCMs.

The second component (25% variance explained) is more difficult to interpret, and may relate to the correlations with the magnitude of errors found between metric scores. Principal Component 3 (12% variance explained) accounts for the relationship of persistence indices CDD and CWD to the remaining set of metric scores. This is found particularly with SMHI-RCA and C4I-RCA3, negatively correlated with ETHZ-CLM and RPN-GEMLAM (Figure 4.25, third component). The majority of RCMs do not display this pattern however. This suggests that CDD and CWD are important to characterise this behaviour for a large portion of the RCMs in this ensemble. The results of the PCA indicate that these two main groups of variables are related, in that they are anti-correlated in their measures of RCM performance. However, one cannot be considered redundant and eliminated from the final set since the second principal component indicates that both temperature and precipitation are necessary to represent a substantial degree of the variance in the metric data. Thus, both sets should both be represented fully in any final set of metrics. The CWD and CDD metrics are highlighted as important given their dominant influence on PC3.

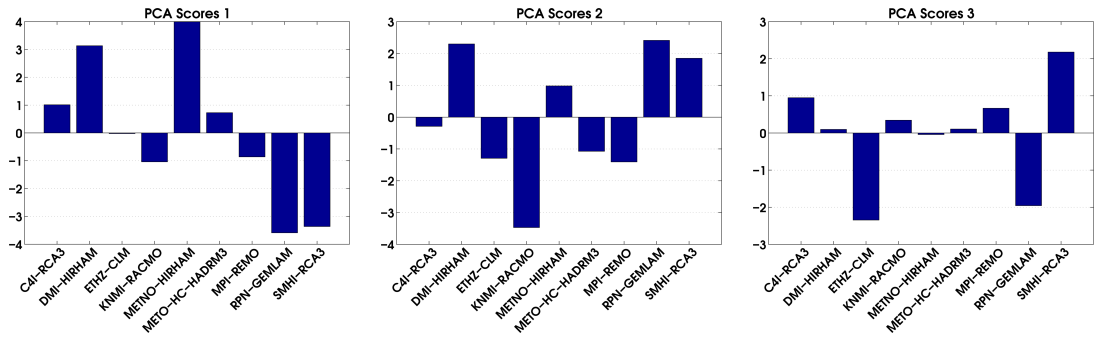

Figure 4.25: PCA scores for the first three principal components for ENSEMBLES RCMs.

These redundancy analyses have been consistent in their results, mostly confirming the presence of two main sets of closely related variables, and the relationship between the two. The tests do not however provide unambiguous advice as to the most appropriate or important variables to use from a statistical perspective however, which gives rise to a degree of subjectivity in choosing a final set. Notwithstanding, some conclusions can be made about the uniqueness of information given by each metric for assessing RCMs. Persistence variables are generally the most different to all other indices, and therefore should likely be kept whereas indices of frequency and magnitude tend to reflect performance either of the standard variables (temperature or precipitation) or that of similar indices. Fewer of this second type of index should be used therefore. Sea-level pressure is also considered qualitatively and quantitatively different from the other variables and should be kept in the final set. The precipitation extreme indices Rx1day, Rx5day, R10mm, R20mm although considered from the results of this ensemble to not be especially unique in relation to the mean precipitation, it should be noted that assessments based only on the mean of a variable's distribution can be misleading. Therefore one extreme precipitation index, Rx5day, will be included in the final set. Tmean is considered unnecessary to the final set, as are Frost Days (FD) and Icing Days (ID), given that Tmin and Tmax will sufficiently span the temperature distribution. The choice of spatial domain, given the somewhat inconclusive findings in Section 4.3, is taken to be the full European domain, partly for simplicity, but also for allowing further results to be generalisable to other regions more readily. For the statistics to be selected, it is determined that seven shall be used for further investigations in Chapter 5, namely:

Spatial Pattern RMSE, Annual Cycle Skill Score, Annual Variability Metric, Interannual Variability Metric, Linear Trends, PDF Score and CDF Score.

Final set of variables: Tmin, Tmax, DTR, WSDI, CSDI, Pr, CDD, CWD,

Rx5day, SLP

Chapter 5

Metric Combination Approaches

5.1 Introduction

A single performance metric, although able to give a concise quantitative indication of model skill, is limited in its assessment to the aspect of the specific variable in question. A broader and more comprehensive assessment of model performance across a range of variables and statistics is desirable to produce generalised performance indicators (GPIs) so that the overall skill of a climate model, quantified by a range of metrics, can be summarised in as compact a format as possible. The potential benefits of doing so would include the evaluation of improvements of multi-model ensembles over different successive generations, and by doing so providing a useful summary benchmark for users of RCM or GCM data on which they can quickly assess the viability of particular models for use. Future climate change projections can be weighted using GPI output on the basis that a higher GPI score is likely to indicate greater model reliability outside of the historical validation period. For these applications, metric combination approaches build on the objective, quantitative approach provided by performance metrics, but aim to be superior due to their ability to take into account a wide range of factors relevant to whatever is considered 'good' performance. However, there remains much scope in what such combination methods might entail, with many approaches promoting somewhat ad hoc constructions. As such whether it is possible to reach a well defined, robust and justifiable method has yet to be demonstrated (Christensen *et al.*, 2010).

There are two general components to a GPI: the choice of combination method, and the choice of what metrics are combined. A combination method is essentially

a form of averaging that brings together a range of performance metrics, be they absolute or relative, and produce a single quantitative output. The choice of metrics could relate to a broad range of variables assessed with a single statistic (e.g Murphy *et al.*, 2004; Reichler and Kim, 2008), or a narrow range of variables assessed with a range of different statistics (e.g. Coppola *et al.*, 2010; Kjellström *et al.*, 2010; Xu *et al.*, 2010). Therefore two main questions arise: how sensitive are GPIs to changes in combination method, and also to changes in the composition of metrics used. More broadly, what is the role of GPIs in assisting end users, for example an impacts modeller looking to select one or two RCMs for their study? Are GPIs a good method for choosing or eliminating models from a study? What are the benefits of using GPIs over qualitative analysis of a range of single metric results? Are GPIs of use to model developers in benchmarking their RCM?

To answer the first and more concrete questions, a number of sensitivity studies are carried out. First, ways of combining metric information are identified and applied, with their utility assessed. Second, a test of whether the number of metrics included in combination is a factor to be acknowledged is undertaken. Two final analyses investigating how different types of variable can affect GPIs and whether including seasonal and multi-statistic information is beneficial are also presented. Before this however, methodological details of these analyses are outlined.

5.2 Methods

5.2.1 Metrics, Statistics and Pre-processing

To produce and investigate the sensitivity of GPIs a range of approaches are used. In summary, GPI output is produced utilising a reduced set of metrics, found through the elimination of redundant variables and statistics in Section 4.5 (Metric Sensitivity) Analysis. This set of metrics is expanded beyond the annual assessments to include seasonal values, where appropriate, to give a larger set of metrics with which to explore the issues outlined. Seasonal values are not used in cases where an 'annual only' extreme index is used. In total this final set constitutes 182 metrics for each spatial domain. These metrics differ however in one significant respect from those used in the previous Chapter 4 analysis, as they are normalised relative to observations. This ensures that the GPI values can be compared between ensemble generations, which is not possible with relative scores. The one exception to this

however is the fourth combination method which relies upon RCM relative rankings, and as such GPI scores between different ensembles from this method cannot be compared in this way. Finally, a standardisation procedure is then applied to the metric values to ensure that the metrics do not produce illogical GPI values when combined (e.g. negative metric values giving negative GPI values).

VariableStatistic	SPR	ACSS	AVM	IVM	LT	PDF	CDF
Tmax	A,S	A	A	A,S	A,S	A,S	A,S
Tmin	A,S	A	A	A,S	A,S	A,S	A,S
DTR	A,S	A	A	A,S	A,S	A,S	A,S
WSDI	A			A	A	A	A
CSDI	A			A	A	A	A
Pr	A,S	A	A	A,S	A,S	A,S	A,S
CDD	A			A	A	A	A
CWD	A			A	A	A	A
Rx5day	A,S	A	A	A,S	A,S	A,S	A,S
SLP	A,S	A	A	A,S	A,S	A,S	A,S

Table 5.1: 182 metrics applied in the metric combinations analysis. Some variables are defined as annual counts precluding their use in seasonal evaluations. 'A' refers to annual evaluations, 'S' to four separate seasonal evaluations (Spring, Summer, Autumn, Winter). Metrics are computed for the whole European domain and eight sub-domains in line with those used in Chapter 4.

The set of statistics is that identified from the metric redundancy analysis in Chapter 4.5 constituting seven different statistics: Spatial Pattern RMSE, Annual Cycle Skill Score, Annual Variability Metric, Interannual Variability Metric, Linear Trends, PDF Score and CDF Score (Table 5.1, top row). They produce different ranges of output (Figure 5.1) and go through one stage of standardisation. This is done such that all metrics increase with greater model performance, such that combining does not produce nonsensical results (higher GPI score should always indicate higher model performance, which would be undermined with a metric that does not increase with performance). Furthermore, they are constructed to be non-zero positive-increasing (such that increasing metric values equate to greater performance - RMSE and CDF metrics are inverted to achieve this) through a linear shift. For certain '1-minus error' statistics, S, (such as ACSS, AVM, LT and CDF), this requires the following transformation to the transformed statistic, S_T:

$$S_T = \frac{1}{1 - S} \tag{5.2.1}$$

This avoids any single metric cancelling out all others all statistics, which could potentially arise in a geometric/multiplicative GPI, by ensuring that all statistics are non-zero. To enable comparison of GPI values between difference generations of RCM ensemble, whilst maintaining equality among metrics included, a normalisation stage is introduced. This stage is embedded within each metric to produce values that are normalised by the observational standard deviation for each variable in question. This way, each GPI input metric is of similar magnitude without making the final outputs ensemble specific, which would occur if metrics were normalised relative to the ensemble.

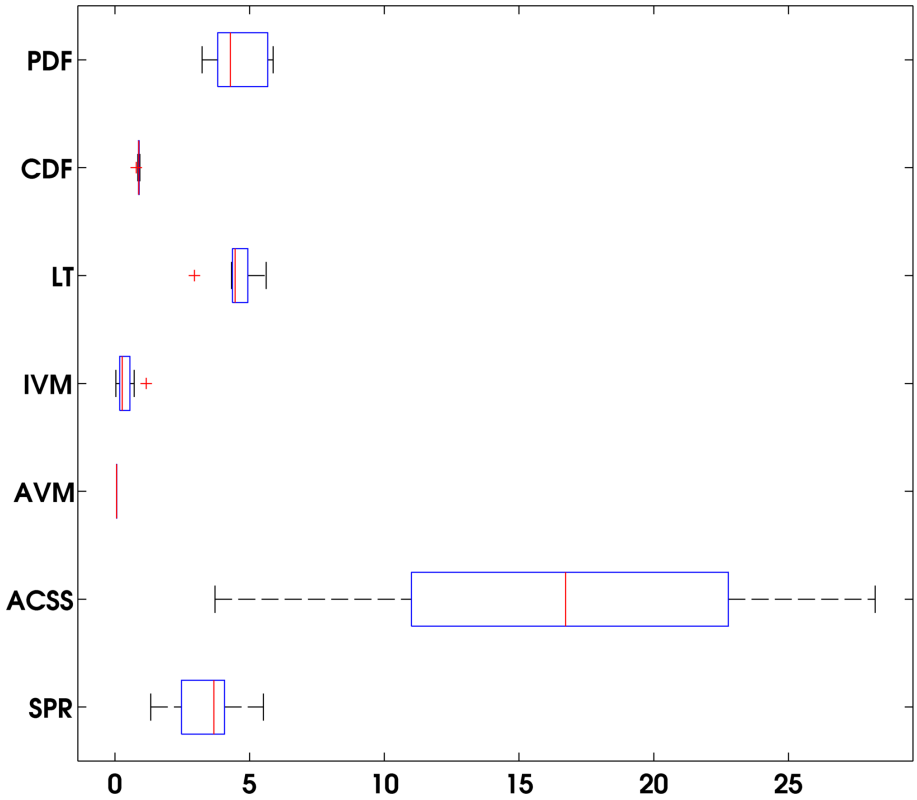

Figure 5.1: Numerical range and distribution characteristics of raw metric scores for Tmax for each of the seven statistics used in the metric combinations analysis. Some statistics present skewed distributions, whereas others are approximately normally distributed. This can have an effect on the discriminatory power of a GPI depending on the combination method used.

All of the statistics are calculated on a grid point basis except for the spatial pattern RMSE (SPR) statistic that requires a two dimensional domain of values for the metric to operate. These metric values are input into a 34x9x9x7 matrix (34 variables | 9 RCM | 9 domains | 7 statistics). The number of variables at 34 can be totalled from Table 5.1, where for each variable seasonal values are either included or excluded. Given that six variables (Tmax, Tmin, DTR, Pr, Rx5day and SLP) use seasonal values they therefore contribute 5 'variables' each (annual + 4 seasons). This equates to $6 * 5 = 30$ variables, and thus including the remaining four (annual only) variables of WSDI, CSDI, CDD and CWD, 34 variables in total are used. To produce RCM ranking scores for each metric an identical size (34,9,9,7) matrix is constructed and RCM ranks are calculated for each variable/domain/statistic combination; rank 1 equating to best performance, 9 the worst. For example for SPR temperature metric over the British Isles, the best RCM with highest metric score would be assigned rank 1, the worst RCM with lowest metric score rank 9. These two matrices (metric scores and ranking scores) are then used as input into the combination methods, the details of which are outlined below.

5.2.2 Combination Methods

Four combination methods are applied to the set of metrics generated: Geometric (Equation 5.2.2), Additive (Equation 5.2.3), Harmonic (Equation 5.2.4) and Ranking. They aim to cover a wide range of possible options for quantifying an aggregate score from a given set of input values. The equations detail how n numbers of metrics f_n are combined into a final GPI score R, and can be used on any set of non-negative, non-zero inputs ($\forall n, f_n > 0$). The ranking approach labels each RCM by its performance for a given metric relative to the other eight ensemble members. It then calculates a final overall score from the sum of the ranks over all metrics, divided by the rank sum (here this equates to 45 for a normalising factor).

$$R = \sqrt[n]{\prod_{k=1}^{n} f_k} = \sqrt[n]{f_1 \cdot f_2 \cdot f_3 \cdots f_n} \tag{5.2.2}$$

$$R = \frac{1}{n} \sum_{i=1}^{n} f_n = \frac{1}{n} \left(f_1 + f_2 + f_3 + \cdots + f_n \right) \tag{5.2.3}$$

$$R = \frac{n}{\sum_{i=1}^{n} \frac{1}{f_n}} = \frac{n}{\frac{1}{f_1} + \frac{1}{f_2} + \frac{1}{f_3} + \cdots + \frac{1}{f_n}} \tag{5.2.4}$$

The combination methods have different qualities. For example the additive approach is sensitive to large values within a set of inputs, and therefore a GPI based on this will bias its appraisal more on the good performance in some metrics rather than requiring good performance over all metric. The harmonic approach on the other hand gives a small value bias to GPIs, meaning that to get a high score, good performance in all categories is more of a requirement. The geometric method is an intermediate case to these two methods and is the most commonly applied combination scheme used in studies (e.g. Coppola *et al.*, 2010; Reichler and Kim, 2008; Xu *et al.*, 2010; Murphy *et al.*, 2004; Sánchez *et al.*, 2009). Beyond those methods selected, there are other alternative methods, for example a quadratic mean, or less well known methods such as a truncated mean where particularly high and low values are discarded before combining. The methods chosen however aim to further expand upon the limited selections used in the literature, to see the effect alternatives could have.

5.2.3 Analysis Methods

One of the main aspects of the analysis is the use of randomly sampled metric combinations. The justification for doing this, even though some samples may not be ones chosen in reality, is that provides an upper-bound with which to handle the inherent subjectivity in selecting metrics for use. To do this, since most GPI studies use in the region of 5-25 metrics, the full set of 182 metrics is subsampled to produce a range of plausible GPI output. This subsampling is done in such a way that an equal proportion of temperature, precipitation and sea-level pressure metrics are selected, subject to their relative sizes in the full set of 182. i.e since half of the metrics are temperature related, any subsample will have half of those metrics being temperature related metrics and so on. For each subsample, a GPI value is then produced for each combination method, and through this an overall GPI uncertainty range can be ascertained for a given number of input metrics. This method is highly parallelised for speed, and the monte-carlo sampling is done 1000 times for each number of input metrics. Results are generally robust to this magnitude of sampling.

$$WA = \frac{\sum_{i=1}^{n} w(i) \cdot m(i)}{\sum_{i=1}^{n} w(i)} \tag{5.2.5}$$

The method used to assess the effect of this range of GPI output is as follows. Weighted climatological averages are produced by first calculating the mean spatial climatology $m(i)$ for each RCM i. Next, these spatial fields are ascribed a GPI weight $w(i)$ (from the above systematic subsampling), before averaging into a single weighted spatial field. Finally this field is then divided by the sum of the weights to account for the magnitude of weighting (Equation 5.2.5). This weighted average is then compared to observations using the spatial pattern RMSE statistic (evaluating the RMSE of a weighted spatial pattern to an observational spatial pattern), and therefore with the range of GPI output produced from the subsampling, a range of spatial pattern RMSE can be produced. The spatial pattern RMSE, when used as an input into GPIs is unitless through the normalisation procedure, whereas here simply for assessing the error between two spatial fields this normalisation is not undertaken, leaving the output in units of degrees Celsius, or mm/month where appropriate. This provides a quantitative method to compare and contrast the different combination schemes, and the robustness thereof, which otherwise would not be able to be done in anything other than a qualitative sense since the GPI values themselves are unitless.

5.3 Sensitivity to Type of Metric Combination Method

As referred to in Section 2.2, (Xu *et al.*, 2010) considered a multiplicative (here called geometric) approach to be a rigorous test of RCM performance by requiring models to have low errors in all assessed criteria to produce a strong overall GPI score. Most GCM and RCM studies (e.g Murphy *et al.*, 2004; Reichler and Kim, 2008; Coppola *et al.*, 2010; Eum *et al.*, 2012) have taken this combination framework and applied it to their chosen set of metrics, somewhat neglecting (although sometimes acknowledging) this potential source of uncertainty in GPI methodology. Christensen *et al.* (2010), arguably the most comprehensive RCM study investigating metric combinations, did examine a further two additional combination schemes, one based on reducing the spread of weights from the geometric method, the other derived from RCM relative performance rankings.

They found that weighted ensemble errors relative to observations, calculated with these methods, were sometimes worse than the unweighted multi-model mean, and furthermore that the RCM GPI rankings were found to be insensitive to these changes in construction approach. There are two main criticisms which can be directed toward this sensitivity test of GPI construction methodology. First, the set of GPI construction methods tested may not have been as large as it appeared; the 'reduced ratio' geometric method is, as they describe, predominantly a weakened geometric GPI method, and as such may not necessarily provide the desired additional diversity. Second, RCM GPI rankings may not be especially informative of the differences between combination schemes, particularly when outlier models are considered; some methods may emphasise good or bad performance more than others for example, which could become especially relevant when constructing weighted climate change projections.

One main overarching theme emerging from the literature is the impression that ensemble weighting, which directly utilises the output of various GPIs, is an added component of uncertainty when used for constructing future climate change projections. This conclusion is based in large part on the wide array of options available to produce a GPI. The following analysis by further exploring the effect of changes in GPI construction approach, not only on the absolute GPI output itself but also on weighted ensemble errors relative to observations aims to give a fuller picture of the extent to which GPI methods are robust for use, with respect to this specific aspect.

5.3.1 Range of GPI Output for different Combination Methods

It is helpful first to consider the effect of each combination method on the absolute magnitudes of GPI scores produced for each RCM, since these values are what users of such information, such as impact modellers or those interested in constructing climate projections, would be presented with. The first basic point is that each combination method produces different magnitudes of GPI score; the Ranking scores are all less than 1, whereas Additive GPI scores vary up to 3.5+. This is not an issue which affects applications of GPIs in ensemble weighting, since usually such schemes are normalised by the sum of the weights. However, this means that GPI scores from different methods are not directly comparable in this absolute sense. Nevertheless, it is clear that the methods generally are producing GPI

distributions reflecting each of the combination scheme's underlying characteristic. For example, the Harmonic average leans towards the smallest values in a set of metric inputs, whereas the Additive is biased towards larger values. The Ranking approach on the other hand strongly emphasises models with high ranks in all metrics. The effect of these differences on GPI scores is noticeable on the RCMs towards each end of the performance spectrum (Figure 5.2). In Section 4.2 Figure 4.10, KNMI-RACMO2 was qualitatively identified as the best overall performing RCM based on relative performance ranks, and this assessment is consistent with the results from each of the four combination approaches. In particular, the Ranking method (somewhat unsurprisingly) most clearly emphasises this superior performance. At the opposite end of the spectrum, RPN-GEMLAM, primarily due to its severely unrealistic underrepresentation of the diurnal temperature range and the corresponding temperature related extreme indices (Chapter 4.2, Figure 4.1), is found to be worst performing overall.

The Harmonic approach seems best suited to identifying such low scoring RCMs, since the boxplot of RPN-GEMLAM GPI scores is substantially lower and much more restricted than all other RCMs, which is not the case in the other three combination methods. However, despite these potentially useful behaviours in identifying overall better/poorer RCMs for certain combination methods (especially for users requiring a smaller subset or single RCM for use), it is not clear from these GPI score ranges what extent the differences between models are meaningful in most cases. In other words, what is the value of finding that one RCM is 0.1 'better' than another RCM? A further issue is the apparently large overlap of GPI scores in all the methods, particularly in the Additive and less so in the Ranking approaches, and as a result GPI scores for the intermediate performing RCMs seem not to provide a definitive assessment of overall performance. This overlap however may simply be an artefact reflecting the metrics sampling in each GPI, with those metrics producing larger values leading to generally larger GPI scores for all RCMs. This would lead to a range of GPI values for each RCM that do not in fact have a real overlap in GPI scores to other models. To assess the actual effect of these two issues - the extent to which the combination methods are in fact distinct, and the extent to which the range of GPI values produced is a real contributing factor to GPI uncertainty - the GPI values are applied to the commonly used ensemble weighting evaluation with seasonal climatologies. This provides a dual indicator, first of whether a GPI weighted ensemble can be of lower error (relative to observations) than the multi-model mean, and how different each

construction methodology is in a quantitative sense.

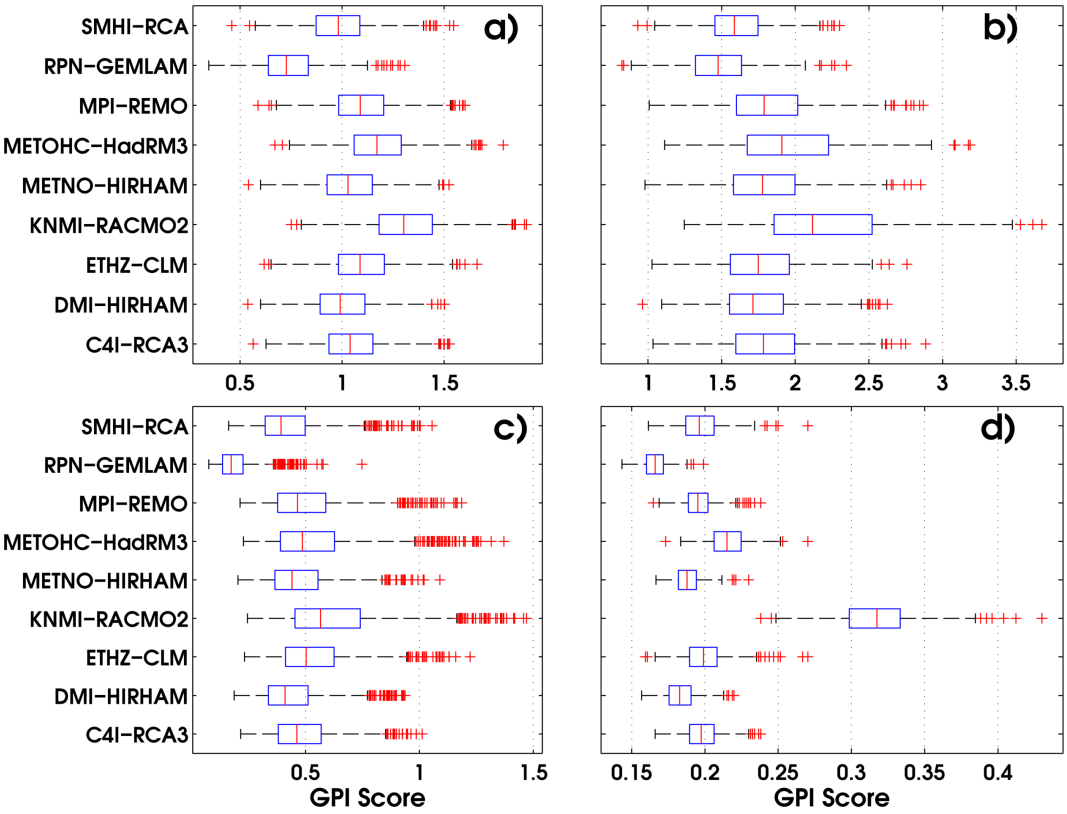

Figure 5.2: RCM European domain GPI score boxplots for four combination methods: a) Geometric, b) Additive, c) Harmonic and d) Ranking. The boxplots represent the median (red line) and interquartile range of GPI scores, with outliers as red crosses. The range of GPI scores shown are calculated from a random sampling (1000 times) of 40 metrics from the full set of 182, with a proportionate number of temperature, precipitation and sea-level pressure metrics included relative to the total number of each metric type available. Higher GPI scores equate to better overall performance.

5.3.2 Effect of Combination Procedure

Often, the utility of an ensemble weighting scheme (based from GPI scores or single metric evaluations) is appraised on its ability to have lower error than the multi-model mean (MMM) relative to observations (e.g. Christensen *et al.*, 2010). This is usually done with a single set of model GPI values, and therefore the consequences of alternative methods is ignored through this single sampling. By testing what effect the apparently uncertain range of GPI scores can have on weighted ensemble error, the overall robustness of GPIs may be inferred. This is because if there are large differences in weighted ensemble errors due to changes

in GPI methodology, then confidence in a single choice of GPI may have to treated with caution. Moreover, this test is a convenient way of assessing the degree of similarity between GPI methods in a more concrete and absolute sense than comparing unitless GPI ranges.

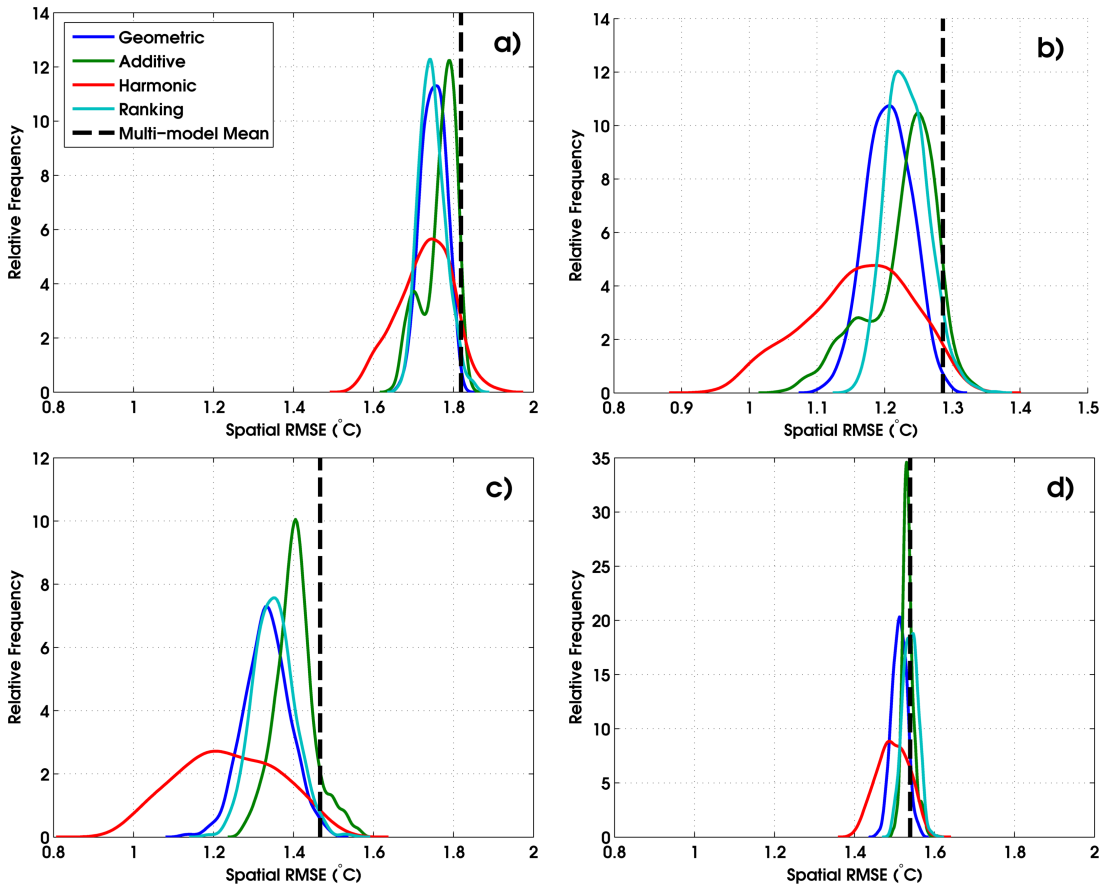

Figure 5.3: European seasonal ensemble weighted climatology average spatial RMSEs relative to E-OBS for a) Tmax Summer, b) Tmax Winter, c) Tmin Summer and d) Tmin Winter. Black dotted line refers to the RMSE of the multi-model mean. The range of weighted ensemble climatologies for each of the four GPI combination methods is shown. 1000 permutations of 20 metrics is used to produce this range of GPI output for each method.

Seasonal Tmax and Tmin spatial mean climatologies for the nine RCMs are combined into a weighted climatological average using 1000 permutations for each of the four GPI combination methods, and the spatial RMSE relative to E-OBS computed (Figure 5.3). Notably, in the majority of cases the weighted ensemble climatologies have lower error relative to E-OBS than the multi-model mean (MMM) with GPIs that consider a range of variables and not only temperature only metrics. This suggests that the principle of constructing GPIs on the basis of good performance across a range of variables and the consequential inferred improve-

ment in model reliability can be of benefit for applications considering a single variable only. However, although GPI weighting does for the most part have lower error than the MMM, the improvement is very slight in each case, being around 0.05-0.1°C. There are some differences between the four combination methods when considering the range of weighted ensemble average errors, which emerges from two sources: the underlying relative performance of the RCMs in each season and the response from each combination method. For example, the Harmonic method presents the widest error ranges and attains the greatest improvement over the MMM, particularly for summer Tmin. The remaining three methods are quite similar to one another in terms of the error range and magnitude. One particular point to make is the above average skill of the ranking GPI approach, particularly for precipitation. This is likely due to the fact that this combination method discriminates more strongly between better and worse performing RCMs, and the overall scores are possibly more inclined to be a measure of precipitation performance rather than temperature. The reason for the Harmonic method appearing best in this case goes back to the characteristics of this approach and also the RCMs themselves. Harmonic means bias output towards lower values, and therefore if an RCM attains consistently low scores with few high values, it will be scored very low. The lowest overall performing RCM, RPN-GEMLAM, happens to be worst in the representation of temperature variables, which explains the improvement in the harmonic weighted ensembles; RPN-GEMLAM is discounted more than for the other combination methods.

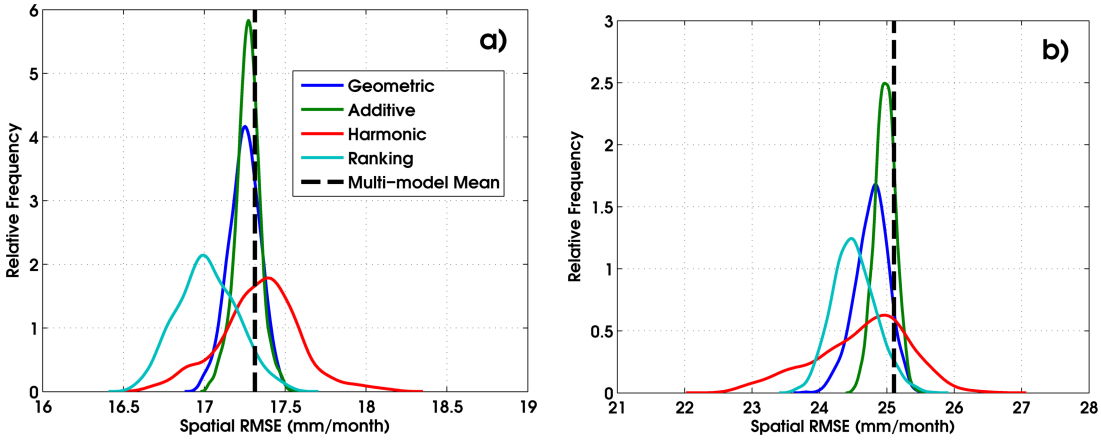

Figure 5.4: European seasonal climatology spatial RMSEs relative to E-OBS for a) Pr Summer, b) Pr Winter. Black dotted line refers to the RMSE of the multi-model mean. The range of weighted ensemble climatologies for each of the four combination methods is shown. 1000 permutations of 20 metrics is used to produce this range of GPI output for each method.

In seasonal precipitation on the other hand, the average improvement is 2-4% over the MMM, with the Harmonic approach not able to be as consistently outperforming of the MMM as in seasonal temperatures. This is due to the fact that although the two worst overall RCMs, RPN-GEMLAM and SMHI-RCA3, have low GPI scores, they do in fact perform better for precipitation. Therefore down-weighting them with this approach does not improve the weighted ensemble as clearly. The Additive and Geometric methods are both close to the MMM is terms of RMSE to observations, which suggests that these approaches, although relatively robust in terms of weighted output (having narrow ranges), are not as informative in improving upon the MMM.

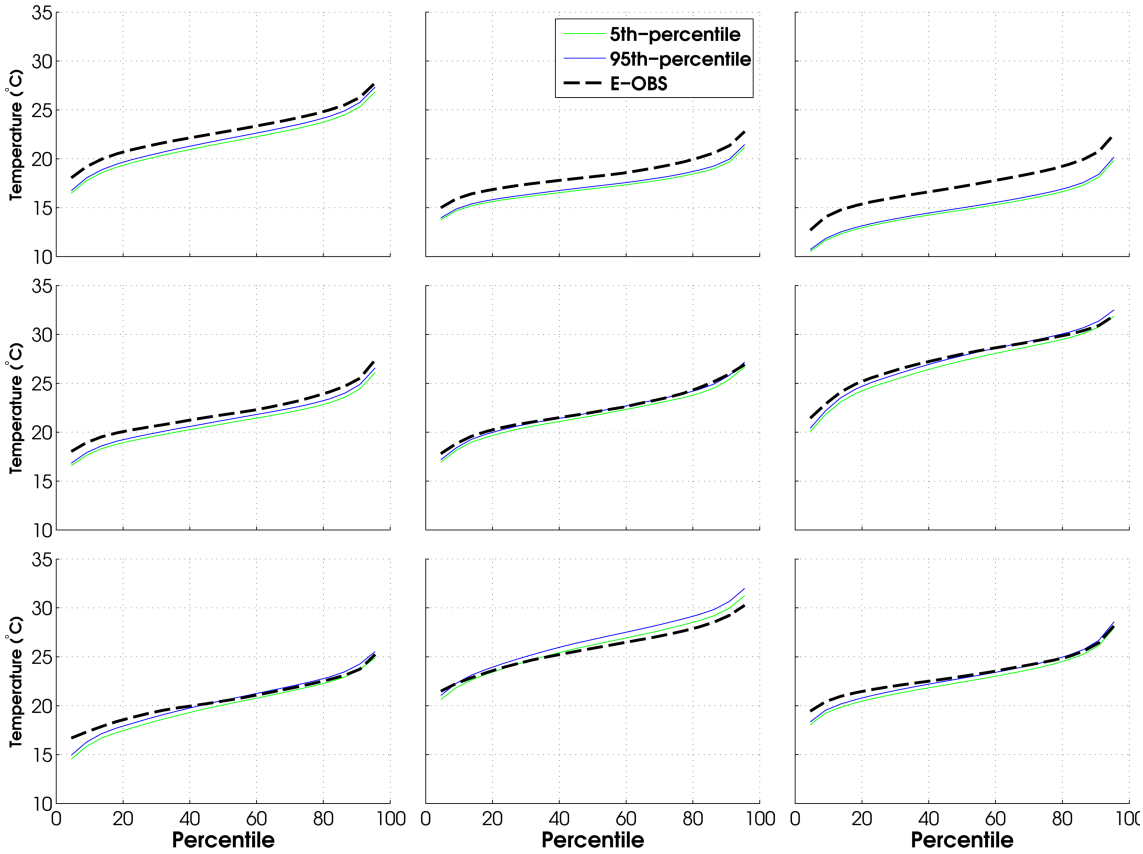

Figure 5.5: Summer monthly maximum temperature European and regional temperature (°C) CDFs for E-OBS (black dotted line) and 5th (blue) - 95th (green) percentile ranges for the nine RCM ensemble weighted with all combination methods. Each of the four combination methods are sampled with 20 input metrics 1000 times, producing 4000 individual weighted CDFs for each region. Percentiles are then calculated from this range.

A second application of GPI ensemble weighting is that applied to temperature and precipitation CDFs, similar to the study of Coppola *et al.* (2010). For the

following analysis only summer Tmax and precipitation data is used as opposed to both Tmin and winter, as the results are similar in all cases. Since the performance of RCMs is more varied at the extremes of the variable distribution, the merits of quantifying and applying overall RCM performance GPI scores may be more beneficial for improving weighted ensembles than for the mean climatology, and thus applying GPI weighting to CDFs will assess to what extent this is the case. The range of weighted Tmax CDFs produced from the four combination methods sampled 1000 times each is shown figure 5.5. The range of CDF values produced is quite narrow and for all regions and percentiles, including the higher and lower quantiles. This result is consistent regardless of overall ensemble performance; for example in Scandinavia where the RCMs show a large systematic cold bias, the GPI weighting shows the same effect as that found in France, where the ensemble performs well. For summer precipitation CDFs (Figure 5.6) RCM weighted averages replicate the observed distribution of rainfall levels, although they overestimate higher percentiles in several regions such as the Alps, Iberian Peninsula and Mediterranean. The range of GPIs values from the four combination methods has little effect on this range of CDFs, for all regions. Since this is similarly independent of ensemble performance to that seen in summer Tmax, this suggests that GPIs are insensitive to both changes in combination method and application (as far as ensemble weighting is concerned).

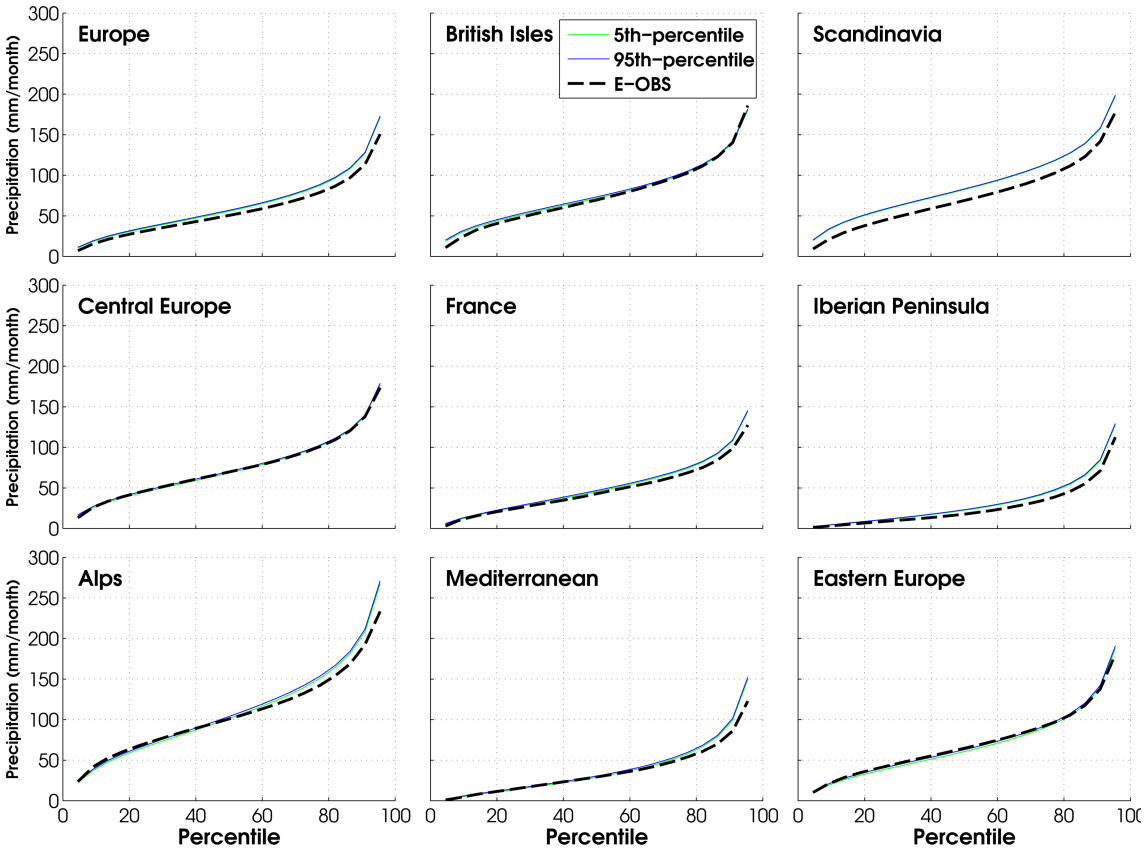

Figure 5.6: Summer monthly precipitation (mm/month) European and regional CDFs for E-OBS (black dotted line) and 5th (blue) - 95th (green) percentile ranges for the nine RCM ensemble weighted with all combination methods. Each of the four combination methods are sampled with 20 input metrics 1000 times, producing 4000 individual weighted CDFs for each region. Percentiles are then calculated from this range

5.3.3 Discussion

What do these results imply for users of GPIs as regards the choice of combination method? A common conclusion found in the literature was that the subjectivity in constructing a GPI is high (e.g. Xu *et al.*, 2010; Coppola *et al.*, 2010). This would imply that the range of values produced from different methods would therefore be wide, leading to uncertainty in any conclusions drawn from or in applications of GPIs. The results found above are consistent with the main finding of previous studies which suggested that ensemble weighting using GPI output, in the majority of cases, gave a small improvement over the multi-model mean when compared to observations (e.g. Kjellström *et al.*, 2010; Coppola *et al.*, 2010; Christensen *et al.*, 2010). The question for this section is whether the choice of combination approach have any effect on these ranges. Differences between GPI output were identified when comparing the RMSE in weighting applications (Figure 5.3), with the Harmonic and Additive methods differing most depending on the variable in

question. However, the effect of changing the combination method, as far as these four approaches are concerned, is rather small in comparison to the magnitude of the common ensemble errors. Furthermore, GPI weighting does not provide a substantial improvement over using simple unweighted ensemble averages. This suggests either that the ensemble weighting approach used to assess differences in GPI methodology might not be overly sensitive to GPI construction changes, or that the use of GPIs is unlikely to be that beneficial when compared to more straightforward ensemble average methods. However, the characteristics of this ensemble should not be overlooked, since most of the RCMs considered here are of similar overall performance. Only two RCMs, KNMI-RACMO2 and RPN-GEMLAM, were substantially different from the main group, which may have been a contributing factor to the low range of weighted values. If a wider range of RCM GPI values is found, then the potential effect of different combination methods may be more of an issue to consider when constructing a GPI. However, on the basis of the above evidence the method chosen to amalgamate metrics into one final overall score is likely robust, and therefore this suggests that the other aspects in producing GPIs are likely to be more important if methodological subjectivity is to be reduced.

5.4 Sensitivity to Number of Metrics Included

In the previous analysis, where the type of combination scheme was of interest, the number of metrics used from the identified set of 182 was held constant (20 metrics for analysis, with the proportion of temperature, precipitation and sea-level pressure metrics held constant in each permutation). Previous studies tend to hold the number of metrics constant in their GPI analysis (e.g. Murphy *et al.*, 2004; Sánchez *et al.*, 2009; Christensen *et al.*, 2010), and as such the potential for GPI output to change depending on increasing or decreasing this quantity remains uncertain. Reichler and Kim (2008) was one study that did test the effect of varying the number of variables included in their I^2 GPI with GCM assessments for CMIP3 and CMIP5 ensembles, and found that the GPI output values tended to overlap for low numbers of variables, and converge to single points when considering the full set of 15 variables. They assessed that when large enough numbers of variables are used, the overall GPI scores are robust when comparing different GCM ensembles. However, they did not further consider the effect of varying the number of variables beyond this qualitative assessment of the GPI ranges produced. The following analysis aims to answer how much the number of variables affects GPI output within a single RCM ensemble by varying the number of metrics over the full

range of values from a qualitative perspective but also using a quantitative approach similar to that used in the previous section.

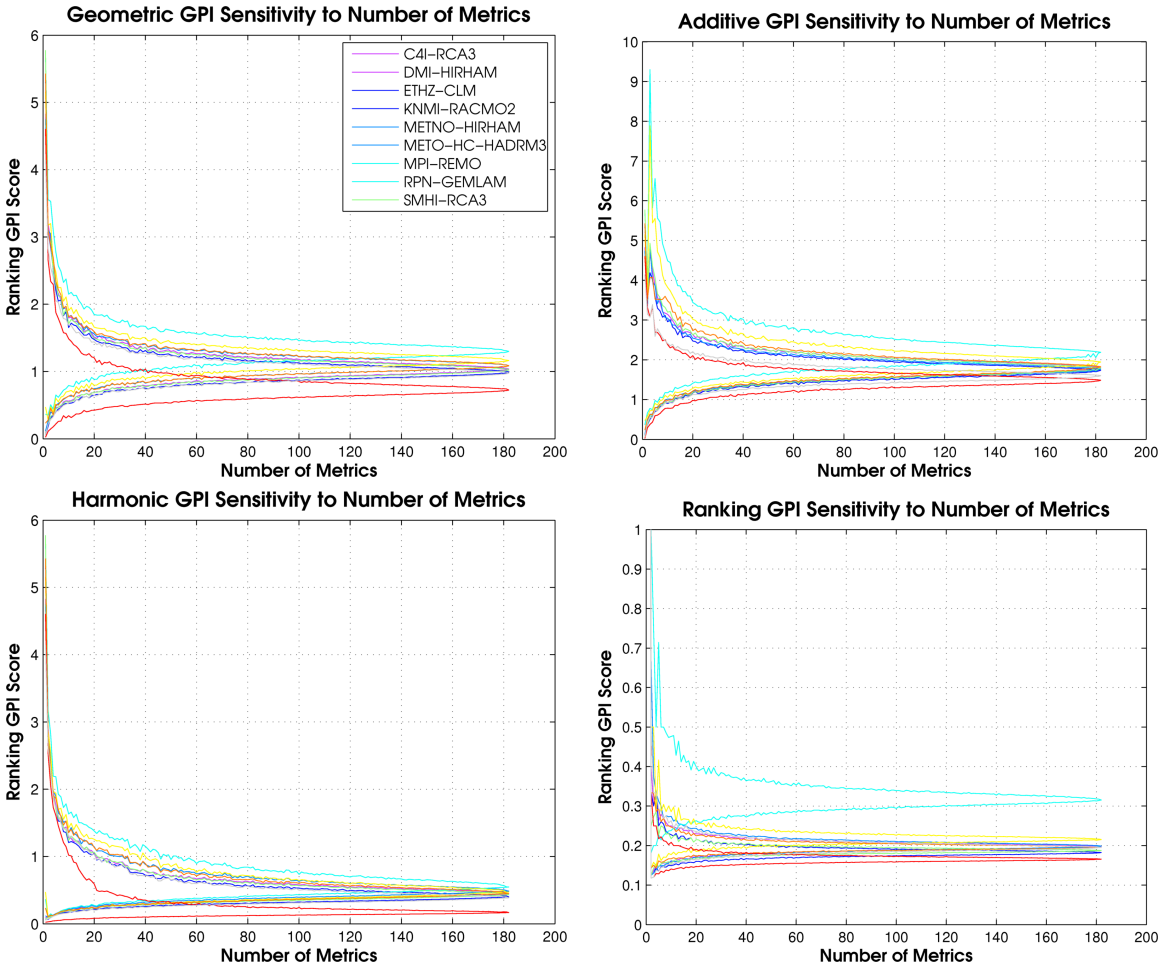

Figure 5.7: Range and absolute values of GPI output for ENSEMBLES RCMs produced for each of the four combination methods (Geometric, Additive, Harmonic, Ranking), varying the number of metrics included in combination. For each number of metrics, 1000 GPI values are produced for each combination method, and the 5th-95th percentiles of this range of values is plotted. For each number of metrics included, the proportion of temperature, precipitation and sea-level pressure metrics used was held constant, so that each sampling is a representative of potential GPI metric choice.

5.4.1 Absolute Range of GPIs with Increasing Number of Metrics

For each combination method, the number of metrics included is varied and a range of GPI output generated for all nine ENSEMBLES RCMs considered in this Chapter (Figure 5.7). The first observation is that the range of values produced with

each combination method is different, with different RCMs more emphasised as found in the range of GPI values in 5.2. This is most clearly seen in the Ranking combination method where KNMI-RACMO2 is clearly differentiated as the overall 'best' RCM when considering approximately 40+ metrics in combination. The remaining methods do not present as clearly this feature, with RCM GPI scores overlapping to a high degree. The Harmonic method on the other hand, appears to converge faster than the other methods, although this may weaken its ability to clearly identify differences between the RCMs in overall performance in this absolute sense. On the face of it, if GPIs are to be used as an overall indicator of RCM performance, then for small numbers (<20) of metrics there is an apparently high degree of uncertainty and subjectivity as to what metrics to include. These findings concur with those of Reichler and Kim (2008), as they identified a similar convergence quality to their I^2 GPI as the number of variables increases. However, the do not consider the actual effect of low variable numbers. In other words, just because the number of variables or metrics is low, that does not mean however that the corresponding differences between the RCMs is changing. If this is the case, then different metric choices may in fact provide robust performance indicators, as the overall distribution of RCMs will remain unchanged, despite the underlying method changing. The following analysis aims to answer this question as to whether the apparent large overlapping GPI output range for lower metric numbers implies that there is a high degree of uncertainty and subjectivity when considering GPIs with a low quantity of input metrics.

5.4.2 Effect of Increasing Metric Number

To do this, a similar analysis approach is taken to that used in testing the differences in GPI combination method is used, where the range of weighted climatology errors is computed for two GPI sets: one for GPIs using 20 input metrics, the other for 100 input metrics. Figure 5.8 shows the errors of GPI weighted Tmax summer/winter climatologies for 20 and 100 included metrics, for each of the four combination methods. Figure 5.7 suggests that increasing the number of metrics should discriminate between the RCMs more clearly, and therefore the GPI weighted climatologies with an increased number of metrics should become more confident. This assumption does hold, with increasing number of metrics leading to narrower more confident weighted climatologies. For Tmax (Tmin not shown but similar), the narrower range of GPI values reduces the spread in RMSE, but

does not reduce the magnitude of the error overall. This suggests that the effect of increasing the number of metrics in GPIs has diminishing returns on the final overall score, and therefore beyond a certain point does not provide additional information. However, Déqué (2007) suggests that although including a number of metrics in GPIs may not lower weighted ensemble errors in a way that differentiates itself as an improvement over 'simple' multi-model averages, the GPI scores should be considered as an improvement, since they take into account how realistic the RCMs are. Christensen *et al.* (2010) on the other hand observes that since GPIs can take into account numerous factors, GPI output is unlikely to give a complete assessment of model performance. The findings in this analysis confirm is that only some of the GPI weightings are unlikely to be a substantial improvement over the MMM, but also that including further information (at least when it comes to seasonal and extreme performance information) in GPIs may not assist in this endeavour either.

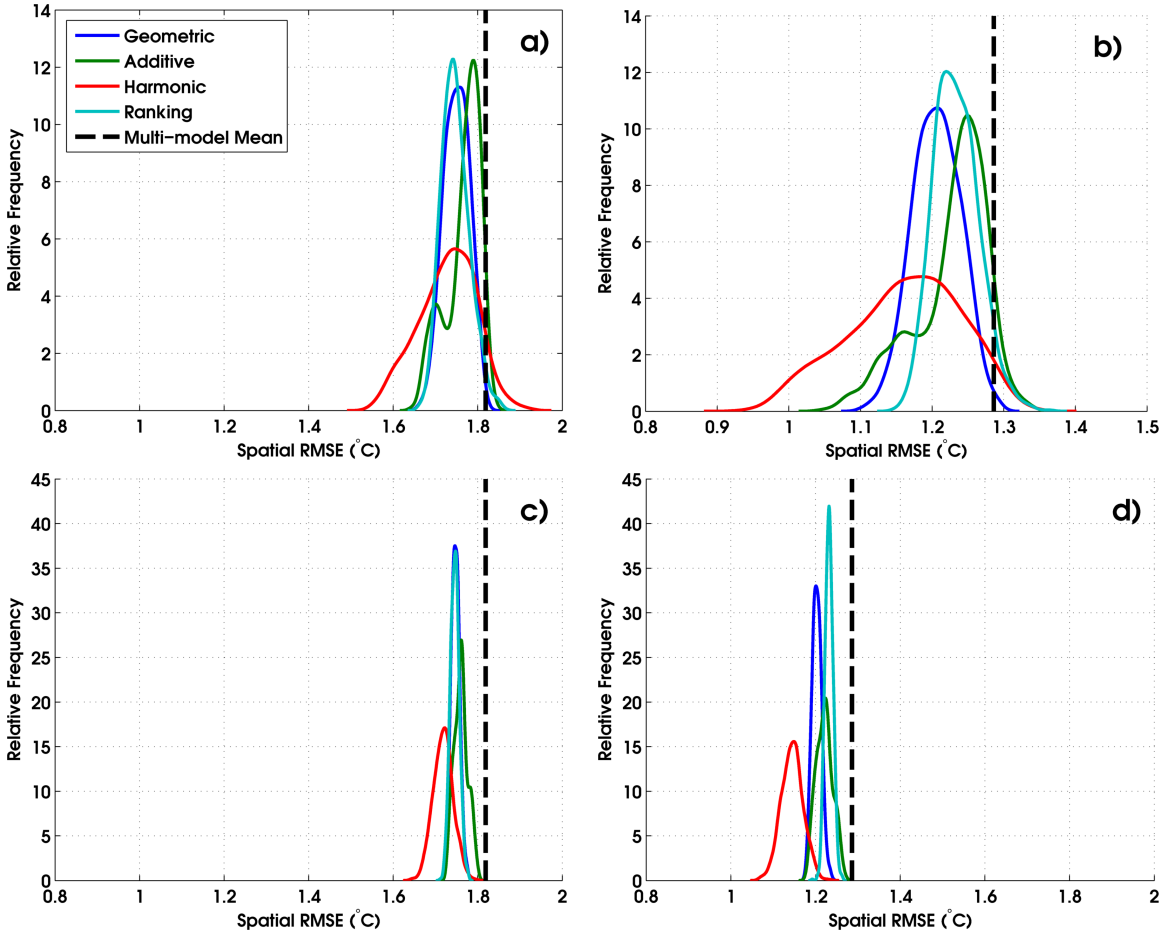

Figure 5.8: European seasonal climatology spatial RMSEs relative to E-OBS for a) Tmax Summer 20 metrics, b) Tmax Winter 20 metrics, c) Tmax Summer 100 metrics, d) Tmax Winter 100 metrics. Black dotted line refers to the RMSE of the multi-model mean. The range of weighted ensemble climatologies for each of the four combination methods is shown. 1000 permutations for each number of metrics included is generated for each combination scheme.

The results from increasing the number of metrics included in each GPI weighting for precipitation (summer/winter) was undertaken both for 20 and 100 metrics respectively (Figure 5.9). They show a related pattern whereby the range of GPI ensemble weighting errors is reduced but with the same magnitude of error present. In one case however of Pr summer for the Harmonic method, the weighted ensemble can produce a climatology that is worse than the MMM. Although one would hope that by weighting RCMs by their overall quality, and that poorer RCMs would be downgraded, that weighting would contribute to an improvement over the MMM. As mentioned previously however, some RCMs happen to perform particularly well in precipitation but poorly in other variables and therefore have lower overall GPI scores.

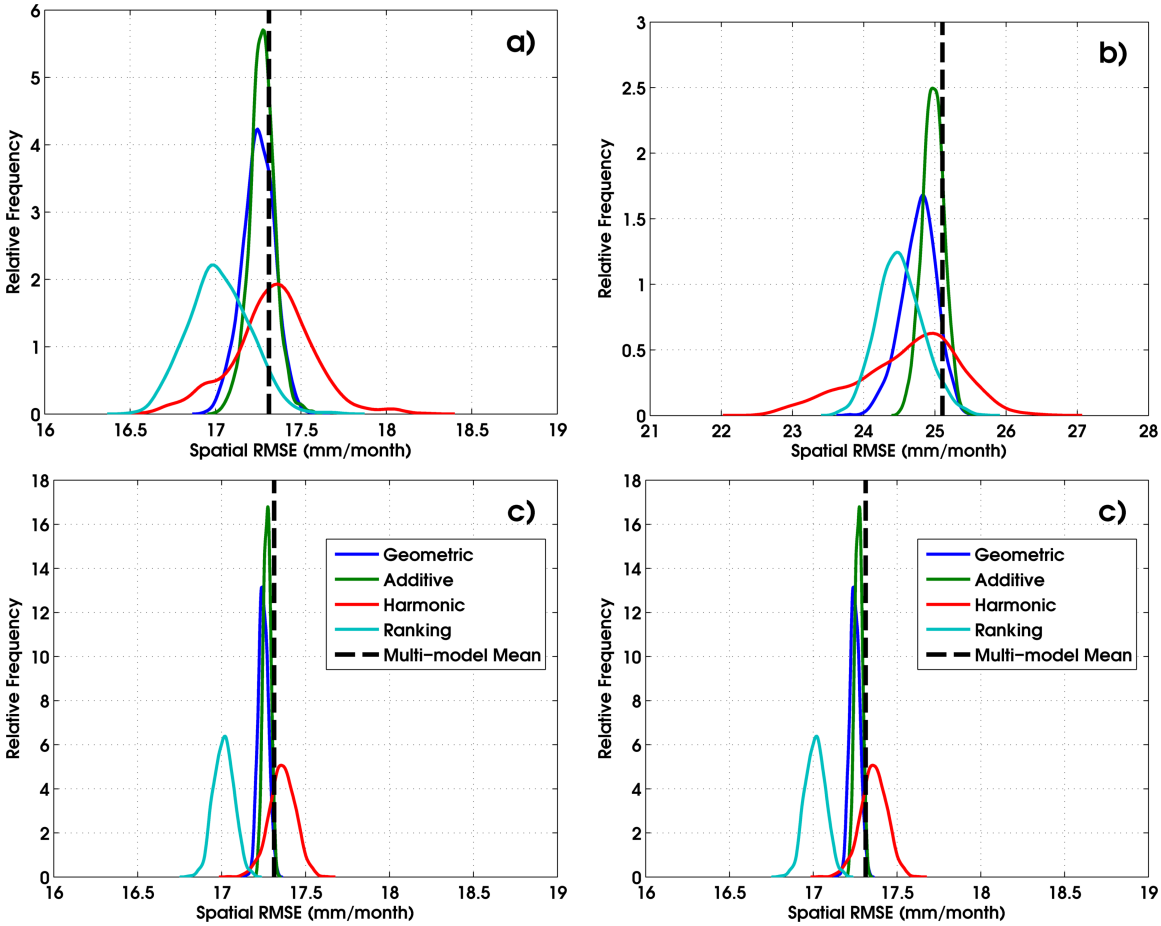

Figure 5.9: European seasonal climatology spatial RMSEs relative to E-OBS for a) Pr Summer 20 metrics, b) Pr Winter 20 metrics, c) Pr Summer 100 metrics, d) Pr Winter 100 metrics. Black dotted line refers to the RMSE of the multi-model mean. The range of weighted ensemble climatologies for each of the four combination methods is shown. 1000 permutations for each number of metrics included is generated for each combination scheme.

5.4.3 Discussion

To what extent the number of metrics included within GPIs has an effect on the final overall scores with the four combination methods has been assessed with a weighting approach. GPI methods are found to be increasing in confidence with increasing number of metrics, as GPI weighted climatological errors relative to observations are found to converge in tandem. However, since the range and effect of GPI values is in fact small, the increase in confidence of GPI values does not imply that GPIs utilising lower numbers of metrics are not robust. This suggests a counter point to the charge that GPIs are prone to subjectivity and uncertainty found in many studies (Kjellström *et al.*, 2010; Xu *et al.*, 2010; Coppola *et al.*, 2010), as the actual impacts of different methodological choices are in fact small. Although the results agree with Xu *et al.* (2010) in finding that as the number of metrics

increase the GPI value converges, this may to some degree simply be due to the averaging properties of the combination method; increasing the number of metrics would be unlikely to increase the range of values when averaging, particularly if the metrics are normalised prior to combining.

5.5 Sensitivity to Type of Variable Included

The metric set considered in this Chapter consists of a range of variables; temperature, precipitation and sea-level pressure. This set could easily be expanded upon, as was done in several GCM GPI studies (e.g. Murphy *et al.*, 2004; Reichler and Kim, 2008), considering aspects such as specific humidity, snow cover and outgoing longwave radiation. The effect of including further variables in this study is difficult to assess with the RCM and observational datasets available. However, one may infer potential changes through the comparison of GPIs utilising the full set of variables with a reduced set excluding temperature and precipitation variables respectively. If the output from these reduced GPIs is found to be different in a qualitative sense then one may conclude that to exclude further variables from the set chosen is too narrow a scope.

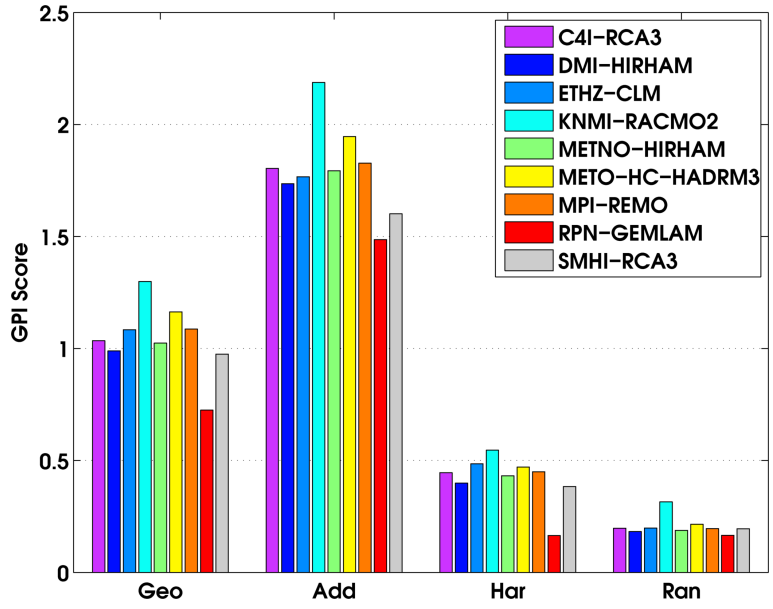

Figure 5.10: GPI output for the full set of 182 variables for each combination method.

GPI values for each of the nine ENSEMBLES RCMs and for each combination

method is produced (Figure 5.10). As noted earlier, KNMI-RACMO2 scores highest in all combination methods, but relatively more so for the Ranking approach. Furthermore, RPN-GEMLAM scores worst over all GPI methods, but relatively more so for the Harmonic method which penalises low scores more. Figure 5.11 shows the GPI output produced from the full metric set excluding precipitation variables (i.e temperature and sea-level pressure variables only). The GPI scores in this case are very much similar to those generated when precipitation was included, which suggests that those metrics assessing precipitation are not a strong determinant of overall RCM GPI scores. If this is compared to when temperature metrics are excluded (Figure 5.12) it is clear that although precipitation metrics do not have that large an effect, temperature metrics on the other hand appear to be the dominant factor in these GPI methods in evaluating overall RCM quality. RPN-GEMLAM in this GPI configuration is now considered to be substantially higher performing relative to the other RCMs in the group, particularly for the additive combination scheme. This factor alone, whereby the RCM considered to be the worst can be then assessed as 'good' when a single set of variables is removed, has consequences for the robustness of GPIs. If this is potentially true of many relevant variables which could be included in a GPI then it is difficult to regard GPI scores as in any way robust, as addition or exclusion of different variables could possibly alter the underlying pattern produced. However there are several factors related to the experimental setup which may mitigate this finding.

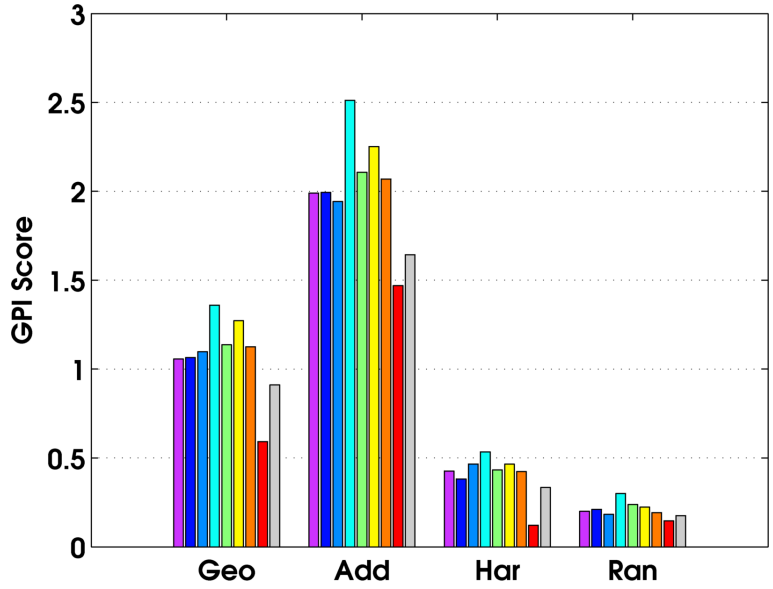

Figure 5.11: GPI output for the subset of metrics excluding those that are precipitation related for each combination method.

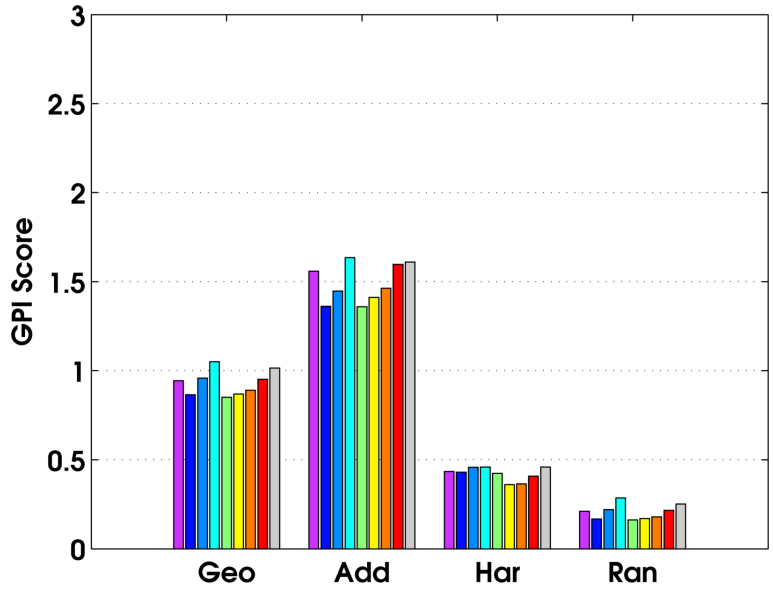

Figure 5.12: GPI output for the subset of metrics excluding those that are temperature related for each combination method.

First, the metrics considered only cover essentially three variables, albeit including a range of extreme indices and spatial and temporal aspects, and therefore the fact that excluding one of these has a large effect might not be overly surprising. Reichler and Kim (2008) in their GPI method assessed GCMs over 15 variables with a single statistic and found that GPI robustness could be achieved with 6 or more

variables of different characteristics. This could therefore imply that GPI robustness is not best achieved with a detailed sampling of different modes of variability and extremes, but by sampling a wider range of key variables. That is not to say that information on model reliability cannot be inferred from a few variables, but a more broad selection of variables for combination in a GPI may provide a better, more consistent overall measure.

5.6 Reduced Metrics: Expert Set

The previous three analyses all utilised a set of 182 metrics, ranging over a variety of variables, seasons and statistics. It is likely that some of this information is repeated to a certain extent, leading to metric redundancy (as referred to in Chapter 4). This further redundancy analysis is required for two reasons; first, with the addition of seasonal information it is likely that there is redundant information present, and second, it is impractical to recommend 182 metrics for use in RCM evaluation studies. Previous studies investigating GPIs tend to neglect this element of choosing metrics to combine, which could lead to some factors being inadvertently emphasised over others. In this section, the correlations between different metrics are considered, and resulting identified redundant metrics eliminated, simplifying the set of 182 metrics to a much reduced number. Figure 5.13 shows the spatial pattern RMSE statistic correlations for each of the 34 variables. Clearly a high level of redundancy is present between the seasonal values of Tmax, Tmin and DTR, Pr and Rx5day and the SLP variables. Objectively selecting those variables which are to be discarded is not straightforward, although some principles are applied. First, annual aspects are favoured over seasonal, since clearly any seasonal information will be encapsulated (but possibly not completely expressed) in annual evaluations. Second, if possible, variable pairs that relate to one another should be seen as an opportunity for reducing redundancy. For example for this statistic, precipitation and its corresponding extreme index Rx5day are very much closely related and therefore at least one can be removed. For this statistic then, the variables carried forward are: Tmax (A), Tmin (A), DTR (A), WSDI and CSDI (A), Pr (A), CDD (A) and SLP (A).

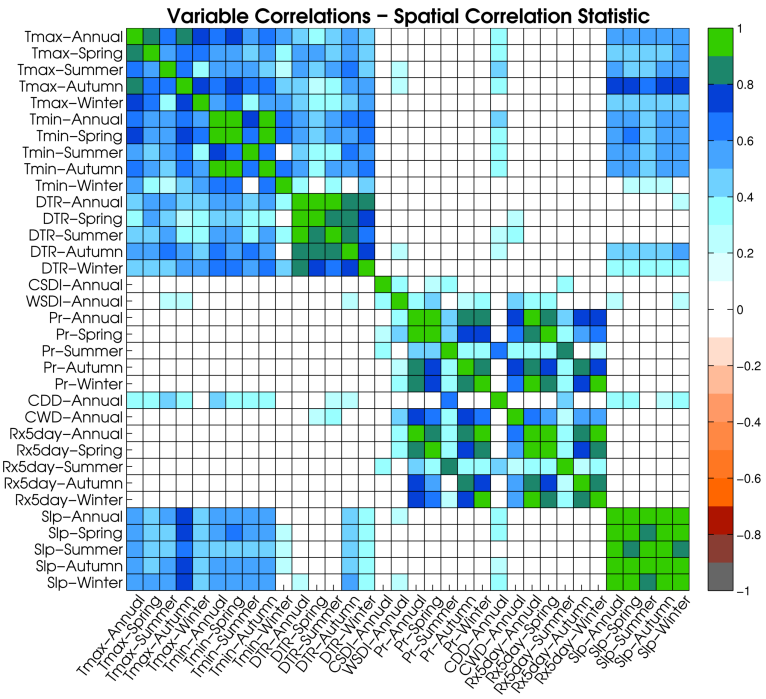

Figure 5.13: Correlation between metric values utilising the spatial RMSE statistic. Only statistically significant correlations are displayed.

Since some of the statistics are related in their approach (e.g CDF and PDF statistics aim to evaluate RCM performance across the whole variable distribution) some are also eliminated from this final set. The statistics eliminated are PDF (for the reason just stated), Interannual Variability Metric (IVM) due to a finding that the statistic was not particularly informative since it provides variable output which is uncorrelated across all variables, suggesting that it is not providing meaningful information. Furthermore the Annual Variability Metric (AVM) is excluded since the simulation performance information it is evaluating is party shared by ACSS, and additionally that it does not span a wide range of variables. Finally, the Linear Trends statistic is removed as long term errors can be assessed with other statistics. Therefore the statistics used are spatial pattern RMSE, Annual Cycle Skill Score (ACSS) and CDF skill score. Since ACSS only considers annual values the variables taken forward are Tmax, Tmin, DTR and Pr. Finally for the CDF skill score (Figure 5.14) most variables are uncorrelated outside of their local seasonal variations, except for Pr and Rx5day which are closely correlated. It is therefore considered reasonable to exclude Rx5day from this final set as its information is highly replicated by Pr. One could make an argument for including WSDI, CSDI, CDD and CWD for this final statistic, since they are not correlated particularly strongly with any other variable other than themselves, however since CDF is a

frequency statistic, it does not make as much sense to apply it to measures of extreme event persistence. For this reason, these variables are not included in RCM assessments using CDFSS.

This leaves a final total reduced set of 17 metrics consisting of:

Spatial RMSE: Tmax (A), Tmin (A), DTR (A), WSDI, CSDI, Pr (A), CDD and SLP (A)
ACSS: Tmax (A), Tmin (A), DTR (A) and SLP (A)
CDF Skill Score: Tmax (A), Tmin (A), DTR (A), Pr (A) and SLP (A)

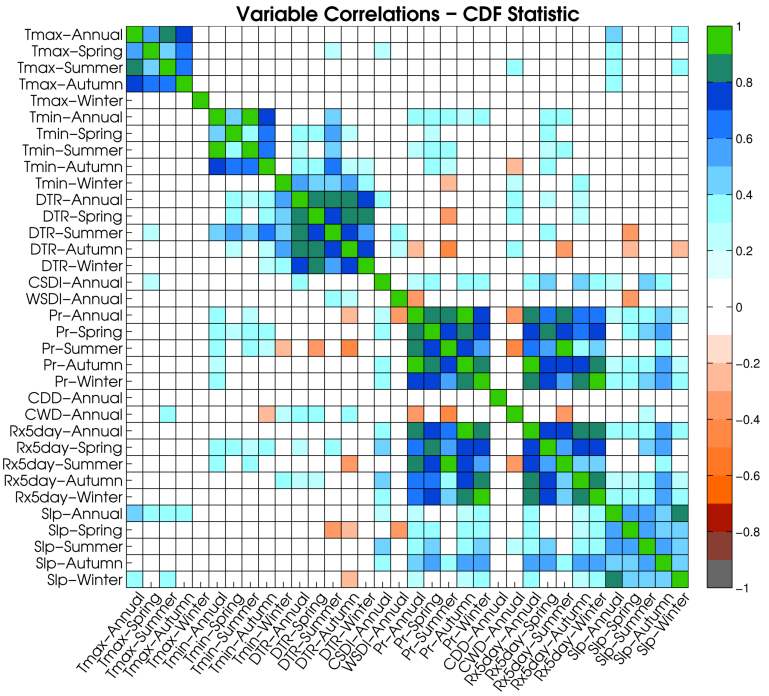

Figure 5.14: Correlation between metric values utilising the CDF skill score statistic. Only statistically significant correlations are displayed.

What is the effect of this reduced GPI? First, GPI values are computed for both the full set of 182 metrics and for the set of 17 metrics with each combination method (Figure 5.15). Previously, it was shown that GPI ensemble weighting does not substantially improve upon the multi-model mean error relative to observations (Figure 5.3). One explanation for this was that the GPI weights were not discriminating enough between the RCMs, given that the majority of the models were

similar in their overall performance levels. More concretely, the GPIs for seven of the nine RCMs were essentially replicating a simple arithmetic mean, leading to little overall improvement. Reducing the number of metrics, and importantly the level of redundant information may be a viable approach to quantifying overall performance with the least metric 'noise'. In other words, including additional related information, such as seasonal evaluations, may be having a normalising effect on the range of GPI values and therefore limiting the utility of this scheme. This can be seen most clearly with the Additive GPI approach, where the spread of GPI values over all RCMs is widened. To test whether this enhanced discriminatory power actually can be of benefit, one further analysis of GPI weighting is produced, where the errors given by the full set of metrics is compared to the reduced set.

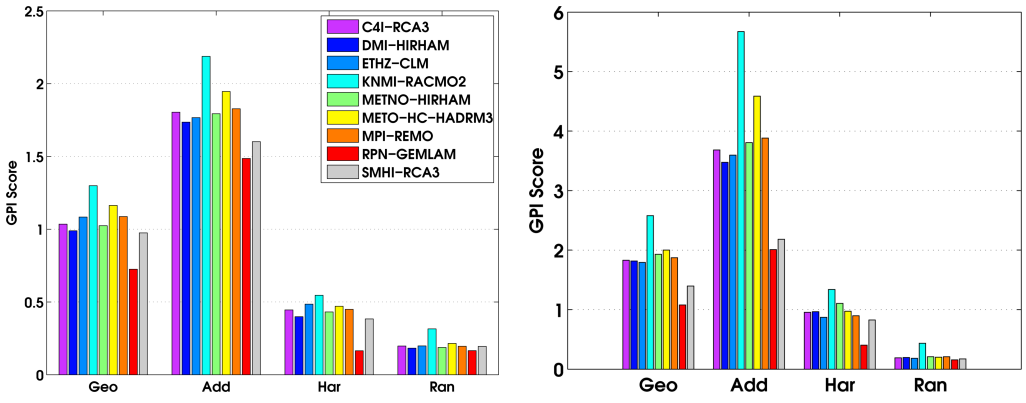

Figure 5.15: GPI output values for the full set of 182 metrics (left) and reduced set of 17 metrics (right). These are calculated for the whole European domain, and four combination methods are used: Geometric, Additive, Harmonic and Ranking.

The effect of implementing a reduced GPI with 17 metrics with ensemble weighting is evaluated through assessment of the improvement over GPI weighting that includes redundant metrics. For Tmax summer first (Tmin results similar) it is clear that the reduced set of metrics when combined with the four combination methods do for the most part improve on the average non-reduced GPI values; the additive method producing the lowest error overall relative to observations. This is likely due to the increased dispersion of the GPI values produced (Figure 5.15), making the GPI output more discriminating between the RCMs. For summer precipitation however, the reduced set can in some cases degrade the weighted climatology, undermining the hypothesis that such a redundancy reduction can only be beneficial. Overall however, the improvements or otherwise are small, and as such do not suggest that reducing the redundancy of input metrics, at least for the

GPI methods considered here, is likely to make a large difference to overall RCM performance in absolute or relative terms.

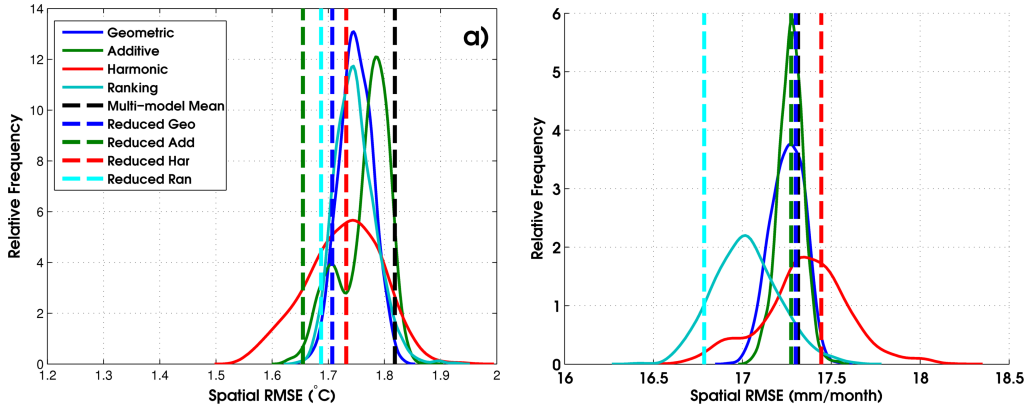

Figure 5.16: GPI weighted values for a) Tmax summer and b) Pr summer. Reduced GPI single values are given by the coloured dotted vertical lines, compared to the black vertical multi-model mean. The range of GPI values is given by the coloured PDFs for the four combination methods used throughout this chapter.

5.7 Conclusions

These analyses have had three aims, two specific and one more general: to quantify the sensitivity of GPIs to changes in combination method, to assess the sensitivity of the composition of input metrics, and more broadly what guidance can be given to different users as to the use of Generalised Performance Indicators?

First, GPI combination methods can have some effect on the output of GPIs. Qualitatively, they do emphasise different RCMs based on their underlying absolute and relative performance. This can be seen most clearly in Figure 5.1, where the best and worse RCMs are clearly identifiable with the additive and harmonic methods respectively. However, a GPI should not simply be used as a tool for finding the best or worst RCM, since the information ascertained can be interpreted as an indicator of model reliability for future projections (Déqué, 2007), although as Xu *et al.* (2010) states this can only be a 'necessary but not sufficient condition'. Quantitatively, the results are more nuanced. The actual difference produced between weighting methods (when applied to an ensemble weighting application) is small, a fact consistent with previous studies which found that weighted averages were little better than the simple multi-model mean (Kjellström *et al.*, 2010; Sánchez *et al.*, 2009; Christensen *et al.*, 2010)- not particularly strong evidence for suggesting that GPIs are related to model reliability. However, there is one upshot of this finding;

namely that GPIs are robust to different choices of combination method. As such recommendations would be to use one of the above methods with confidence, and if alternatives are to be required, to test the robustness relative to this given set in preliminary analysis.

The composition of metrics included suggests that varying the number of variables from a given set of metrics does make the already narrow uncertainty range of GPIs narrower. However, other results indicate that including or excluding a new variable which provides independent information could potentially alter overall GPI scores substantially. GPI methods do converge with additional numbers of metrics, but this could simply be an artefact of subsampling over a given set of numbers, and so this in and of itself cannot be regarded as evidence in favour of further GPI robustness with high numbers of metrics. Reducing the number of metrics to a smaller set of more independent variables can lower weighted ensemble errors relative to full GPI weighting in most cases, however this improvement is again very slight. Recommendations from these results would be to sample as many independent variables as possible. Further independent statistics are of benefit since including redundant information may cloud results with uninformative metric 'noise'.

More broadly, what is the role of GPIs in assisting end users? For users requiring the use of a single model, for example for agricultural or flood impacts modelling, a GPI can provide a useful measure of overall RCM performance, which can (as seen) identify better performing models. If the study requires an interest in specific variables, then GPIs can either be weighted with exponents on these key variables, or a GPI produced excluding other variables which are not as relevant. Model developers interested in benchmarking may find assessment with reduced redundancy GPIs helpful in rating their RCM against other institutes, but beyond this GPIs are a somewhat blunt tool; identifying where RCM biases lie and diagnosing the reasons are not what GPIs are primarily designed to do, although tailor made GPIs with specific variables aimed at certain processes may be of use. When is an RCM good enough for use? A GPI could potentially answer this with more understanding of critical thresholds (e.g. crop temperature thresholds for impacts modelling) relevant for applications, whereby some RCMs would be identified as performing adequately. However, more general a priori abstract thresholds are in practice very difficult to set or justify, and therefore this use should be more application specific. Finally, is there any benefit of using GPIs over a qualitative

analysis of a range of single metric results? Overall results would suggest that in terms of identifying specific weaknesses of RCMs, such as extreme precipitation biases or underestimation of DTR, GPIs are not that helpful. However as a benchmarking and reliability inferring tool they can be powerful in finding those models best performing over a wide range of aspects. Identifying such models will be of benefit to others inasmuch as directing model development to areas in which these better performing RCMs are superior. As regards model reliability, of course the future is unknown and therefore there is no absolute certainty that past performance will guarantee future performance, however a well-constructed GPI, sampling a range of (for the most part) independent variables, is a good start to quantifying this abstract but essential factor, particularly for applications with future projections.

Chapter 6

The Stationarity Assumption

6.1 Introduction

The assumption that the relationship of historical RCM systematic biases will remain constant to the 'true' future climate is crucial to the reliability of a number of climate change projection applications. Statistical downscaling (e.g. Hayhoe *et al.*, 2012), ensemble metric weighting (e.g. Räisänen and Ylhäisi, 2011) and bias correction (e.g. Ehret *et al.*, 2012; Teutschbein and Seibert, 2012) approaches all implicitly require that past RCM performance holds into future projections. However, this has come into question (Christensen *et al.*, 2008) given the presence of identified temperature dependent RCM biases. What this implies is that the historical performance of an RCM may not be the complete determining factor in how reliable the RCM may be in future projection, or whether those projections may be predictably over-or under-estimating the regional climate change signals. This chapter aims to answer whether RCM temperature biases are in fact stationary, using the framework of specific bias correction methods (quantile mapping) to understand how RCM temperature distributions may change in a warming climate. Only temperature is considered in this chapter, rather than others such as precipitation. This is done to focus particularly on the role of long-term temperature trends which has been discussed in Chapter 4.

6.2 Methodology

6.2.1 Stationarity of Historical Biases

The approach of Christensen *et al.* (2008); Christensen and Boberg (2012) is one version of the 'constant relation' assumption. It assesses historical biases of RCMs across the whole distribution of annual temperatures and infers that future biases are dependant on this relationship. More concretely, for their case study Mediterranean summer temperature biases were found to increase in hotter months, suggesting that in future projections RCM biases may increase in these regions. This relationship is clearly dependant on the type of physical process causing the bias, and so such an approach may not be appropriate in other regions and seasons.

An alternative method to the constant bias assumption is that of 'constant relation', where biases can change in time but given by an historically derived relationship. An example of this would be temperature-dependant model biases where simulated temperature biases increased in warmer months (Christensen *et al.*, 2008; Christensen and Boberg, 2012). This approach is essentially the same as quantile mapping (Kerkhoff *et al.*, 2014), where a transfer function is used to 'correct' the future distribution of a given variable to that found in historical observations. This assumption, although potentially more realistic than the 'constant bias' assumption requires that the (often linear) relationship used to correct future temperatures remains the same in time. This has been questioned by Bellprat *et al.* (2013) who noted that causes of model bias may not behave linearly indefinitely, as in the case of soil-moisture-feedback dependant temperature biases are restricted by how dry the soil can become. The suggested approach therefore is to link the correction function to the underlying causes of model bias, rather than assuming that linear relationships are accurate representations of bias. The questions raised by these methods is dependant on the underlying assumption used. For 'constant bias' this is straightforward: to what extent do biases in the mean (additive bias) and standard deviation (multiplicative bias) change over time? For the 'constant relation' assumption it is: is the historical relationship between observations and simulations constant in time? In the context of this chapter, the constant bias assumption can be considered by the average bias change approach, whilst the constant relation assumption is more precisely tested using the percentile bias change method.

First, to evaluate the general characteristics of historical RCM simulations

quantile-quantile (q-q) plots are produced for Scandinavia, a region of particular systematic bias. The reason this type of analysis is not implemented in Chapter 4 is because the causes of RCM bias were not as critical to questions considered, namely the general sensitivity of performance metrics, whereas in this analysis the precise behaviour of the RCMs is more important. The q-q plots are produced by first producing spatially averaged timeseries, and then calculating the 0.5-99.5th percentiles of the full 480 month timeseries (360 months for 1971-2000) for both RCMs and E-OBS. The RCM percentiles are then plotted against the observed percentiles.

The assumption that historical RCM error characteristics evaluated by comparison against observations will remain constant in time is commonly referred to as the stationarity assumption (Teutschbein and Seibert, 2012). However, what is missing from this definition is a consideration of the sources of historical biases. Following Kerkhoff *et al.* (2014), one may decompose the seasonal observational temperature $o(t)$ into three components for each year t (Equation 6.2.1): an external climate forcing signal $\lambda_o(t)$, natural multidecadal variability $\eta_o(t)$ and interannual variability $\epsilon_o(t)$. Likewise an RCM timeseries $x(t)$ can be similarly represented but with an additional error term β_x (Equation 6.2.2). Such a decomposition can aid in understanding where analysis of apparent changes in historical bias may be robust, or questionable.

$$o(t) = \lambda_o(t) + \eta_o(t) + \epsilon_o(t) \tag{6.2.1}$$

$$x(t) = \lambda_x(t) + \eta_x(t) + \epsilon_x(t) + \beta_x(t) \tag{6.2.2}$$

This historical bias β_x can be expressed as a normal distribution $N(\mu_x, \sigma_x)$, with an additive bias $\beta_{xa} = \mu_x - \mu_o$ and a multiplicative bias $\beta_{xm} = \dfrac{\sigma_x}{\sigma_o}$. The stationarity assumption has essentially two versions: one in which these bias quantities β_{xa} and β_{xm} are constant in time (termed the constant bias), or one in which the distributional relationship of these biases to the 'truth' (expressed through a transfer function) remains constant in time (constant relation). The latter version allows temperature dependant changes in bias, which are determined through an historically derived relationship to observations. The former on the other hand does not allow this to occur. Bias correction methods, though implementing variations on

these two approaches, all assume however that these historical bias characteristics do not change over time.

For historical RCM simulations forced by reanalysis data one implication is that RCMs inherit observed short and long-term natural variability (i.e. $\eta_x(t) = \eta_o(t)$). Therefore, differences in the long-term temperature trend between RCMs and observations are likely due to a differences in the external forcing response ($\lambda_x(t) \neq \lambda_o(t)$). The resulting historical simulation error therefore originates from two separate sources: RCM parametrisation and structural deficiencies or a misrepresentation of a response to changes in external forcing. An example of the latter component is in the representation of aerosols, since these are not prescribed by reanalysis, yet may have a substantial effect on regional scale climate change signals. The stationarity assumption essentially considers the effect of RCM structural deficiencies to be unchanging over time regardless of the future climate change signal. Having an RCM ensemble with a variety of local responses to a change in forcing in future projections is a desirable quality in terms of spanning a wide range of uncertainty, however in the case of historical simulations one would hope that the RCM ensemble (when forced by reanalysis) would be able to replicate the observed signal consistently. It should be noted nevertheless that climate change signals on the regional scale are difficult to detect (Bindoff *et al.*, 2013).

6.2.2 Stationarity of Future Projection Biases

To understand the potential to which RCM systematic biases may be non-stationary in future projections, it is logical first to assess the degree to which this occurs in historical simulations. This historical behaviour can be assessed through the application of bias correction methods to split-sample time series (Teutschbein and Seibert, 2012), or alternatively by assessing mean bias changes (Chen *et al.*, 2015; Maraun, 2012). The former approach can be used to assess the stationarity assumption in an indirect manner, using bias correction methods to correct an historical timeseries based on the simulated bias characteristics of the preceding 20-year period (this method is not applied in this analyses). If a bias correction method produces a future timeseries of lower error than an uncorrected one, then that may be interpreted as a case of relatively stationary bias characteristics (zero error would imply perfectly stationarity). This experiment can be seen as a necessary but

insufficient test for whether the stationarity assumption holds in future projections, since if RCM biases can change over a 40 year period then they cannot reasonably be assumed to do so in multidecadal projections. The second method of evaluating mean bias changes, by assessing the difference in climatological means between the two time periods, does not utilise any correction method, but uses a method that follows from the usual definition of systematic mean bias, that is a climatological average over an extended (e.g. 30-year) time-period of a particular variable relative to observations. This approach however assumes that any apparent change in bias does not arise from an inaccurate representation of external forcings, which may distort whether the inherent bias characteristics of the RCM are in fact stationary or not.

In section 6.3 the relationship between RCM simulations and observations in two separate analyses is evaluated. First, trends in historical bias against observations (E-OBS) covering 1961-2000 are evaluated in a similar fashion to Chapter 4. This is done for summer and winter temperatures, using reanalysis forced runs (ERA-40) to assess changes in model bias over the shorter control period. Second, quantile-quantile plots are produced for the Scandinavian 'Rockel' region for mean temperature to identify if possible temperature-dependent biases may be present.

The pseudo-reality approach has been used before for assessing potential changes in RCM bias stationarity (e.g. Maraun, 2012), however this has been done without fully considering the role that a difference in climate change response to external forcing has on this approach. For this analysis, CORDEX RCMs forced by HadGEM2 and ENSEMBLES RCMs forced by ECHAM5-ES are used as future projections. Given that there are no future observations with which to assess the RCMs, each RCM in turn is used as 'pseudo-observations' with which relative bias changes may be inferred. The RCP8.5 trajectory in conjunction with the longer time-periods of 1971-2000 and 2070-2099 were used for this analysis to ensure the greatest chance for bias non-stationarity to occur.

To begin with, GCM forced simulations are assessed in terms of their seasonal delta-t climate change signal from the temperature change from the climatological averages between 1971-2000 and 2070-2099. This is defined as the grid point future projection climatological average minus the historical climatological average for each gridpoint. This is done to identify regions where there RCMs disagree in terms of local temperature change, and to motivate a discussion of the causes for

this, and the implications for bias stationarity. The average mean climatological bias change is calculated by first evaluating the difference between an RCM and the 'pseudo-observational' RCM in the historical reference period 1971-2000 and then in the future period 2070-2099, and then assessing the change in this spatial map. This is done for both summer and winter seasons, and is compared to the second method of assessing the change in percentile biases as discussed in section 6.2.1. 'Pseudo-observations' refers to one of the RCMs in each pair that is assumed to be the 'truth', or 'future observations', with which to assess changes in bias character.

6.3 Assessment of Historical Bias Stationarity

The approach taken in this chapter to evaluate changes in RCM systematic bias over the historical period is to implement a split-sample strategy, by which a full multidecadal RCM and observational timeseries are split into two halves; the first is used to define what bias characteristics are, and the second is used to assess to what degree these biases change. This approach has been used in statistical downscaling (e.g. Gutiérrez *et al.*, 2013) and bias correction validation studies (e.g. Teutschbein and Seibert, 2012). The role of RCM climate change signal trends has not however been considered in this context. Lorenz and Jacob (2010) identified a systematic under-representation of long term temperature trends in ERA-40 forced ENSEMBLES RCMs. They found that although ERA-40 under-represented the trends in observational datasets, the RCMs produced even smaller trends. The implication of this is that using a split sample test over 1961-2000 to evaluate changes in, say, mean systematic bias will be partially masked by the RCM's warming being systematically slower.

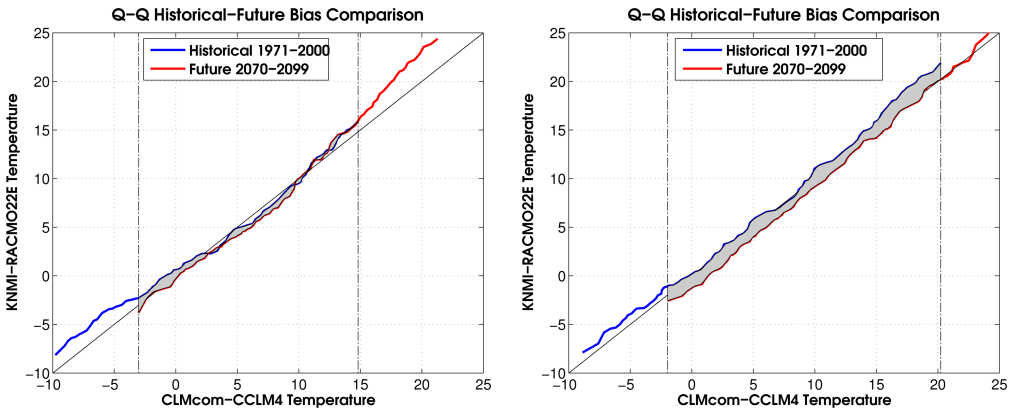

Figure 6.1: Example diagrams for how monthly quantile-quantile bias stationarity is quantified for each grid point. Both plots are generated from two gridpoints chosen for demonstration purposes only. Historical 1th-99th quantiles (°C) are calculated both for 1971-2000 and for 2071-2099 (RCP8.5) for two RCMs, and are plotted against each other. If biases are indeed stationary, for each 'pseudo-observational' temperature the corresponding RCM temperature should be the same in both historical and future projection periods. The deviation from this is given by the grey shaded regions, the total of which is used to quantify the average change in quantile bias.

The new percentile bias change method introduced in this chapter is an approach to testing the assumption that there exists a constant relation between historic and future projection temperature distributions of RCMs. Most studies do not directly test this but instead implement bias correction algorithms and evaluate the resultant corrected timeseries error against a future timeseries (e.g. Teutschbein and Seibert, 2012). By using this percentile bias change approach, it is recognised that it is not in principle possible to test the stationarity of the hottest of future temperature percentiles, since these 'pseudo-observational' temperatures do not occur in historic simulations. This can be seen in two examples (from two separate grid points) given in Figure 6.1, where the blue line represents an historic quantile-quantile plot, and the line represents the corresponding future projection q-q plot, for a grid-point for two RCMs. The percentile bias is defined as the area between either the blue or red lines and the black diagonal line, which represents where the models would have zero bias relative to one another. The change in percentile bias can be assessed for those 'pseudo-observational' temperatures (of CLMcom-CCLM4 in this case) that exist in both historic and projection time series (between the two black vertical dotted lines). The exact change is assessed by the area of the grey shaded regions. In the left hand side case, there is negligible bias change, even though there is substantial warming in both RCMs; the coldest winter months for example shift from -10°C to around -3°C in future projections. In the right hand side case on the other hand there is a shift in percentile bias since KNMI-RACMO22E systematically less than CLMcom-CCLM4 for all percentiles. This percentile bias change

method directly tests the assumption that historical distributional percentile biases are constant in time, whereas the average climatological mean bias approach does not. The autumn/spring percentile bias method is a variation on this approach, where the two timeseries compared are spring and autumn separately, although they are compared for the same time period (either historical or future projection). One issue with this approach is the fact that for higher temperature changes from historical to projection time periods, the area assessed will shrink by virtue of there being less of an overlap between the lines. To remedy this, a normalising term can be simply be introduced whereby higher temperature changes offset the smaller area with some multiplicative factor. For the results given however, this is not undertaken, and as such since temperature changes are greater in northern and high latitude regions one should be aware that the method presented may be slightly underestimating bias non-stationarity under this approach.

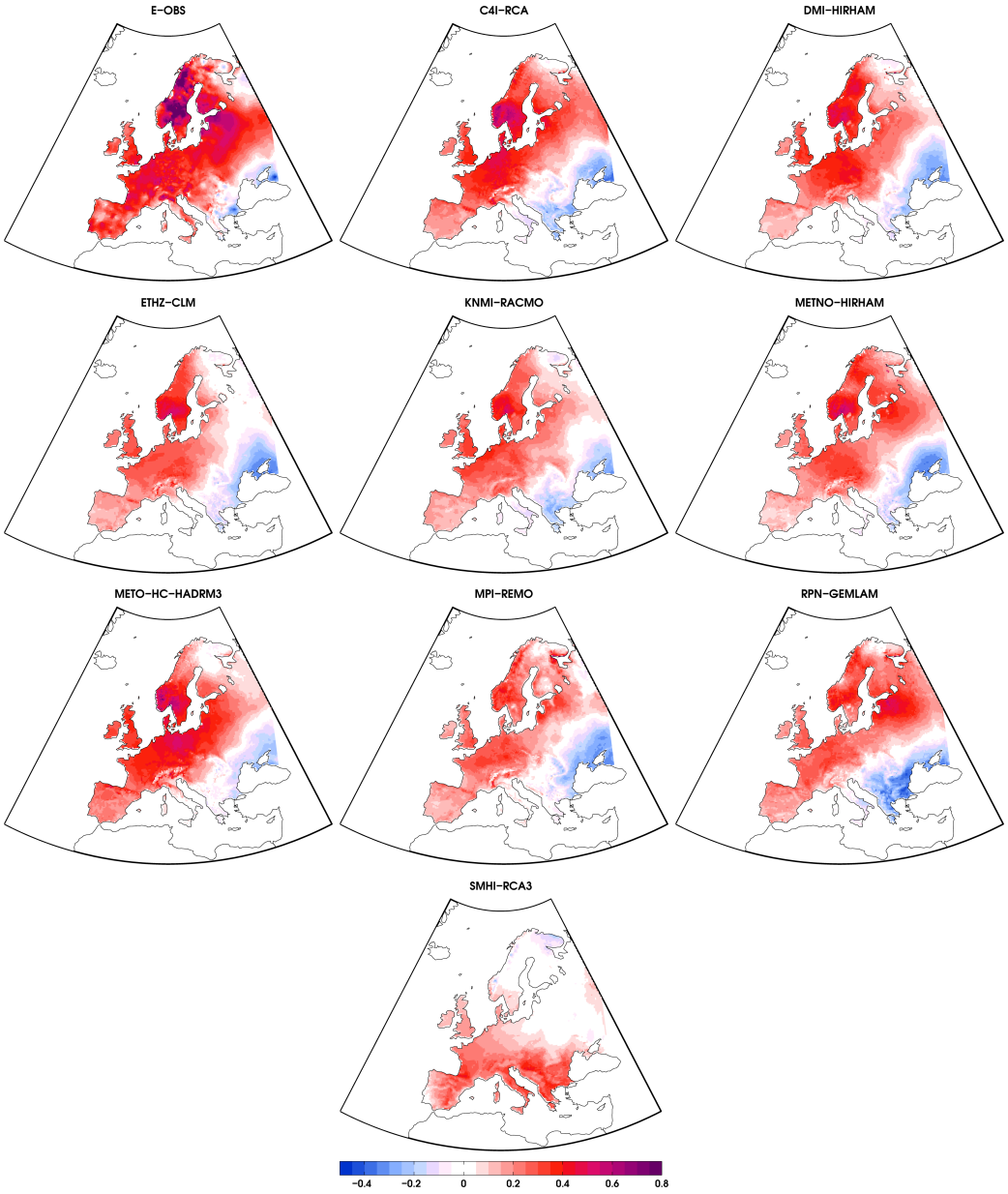

Figure 6.2: ERA-40 forced ENSEMBLES RCM and E-OBS winter temperature trend (°C/dec) 1961-2000. Trends are calculated through a linear regression of seasonal mean values for each grid point.

van Oldenborgh *et al.* (2009) assessed and suggested reasons for shortcomings in the seasonal performance in long-term simulated trends for a range of GCMs and RCM ensembles, including the ERA-40 forced ENSEMBLES simulations considered here (Fig 6.2, 6.3). Their general conclusion regarding the predominant causes for temperature changes are different for each season. In the case of winter trends, temperature changes are more dependent on large-scale atmospheric behaviour than for summer months, for which the representation of small-scale processes is more important. They further note that since RCMs inherit the ERA-40 atmospheric and sea-surface temperature forcing it is not surprising that for winter months the

RCMs perform better than for summer months, where the specific land surface and parametrisation schemes used by each RCM determine the strength of, for example, soil-moisture feedbacks and cloud cover (Lorenz and Jacob, 2010). A third type of boundary condition in the form of aerosols are also relevant in this context, since ENSEMBLES RCMs generally specify a constant aerosol climatology over the full historical simulation, even though anthropogenic aerosol emissions have decreased (Norris and Wild, 2007). Several studies indicate that although there is uncertainty as to the effects, both indirect and direct, of aerosols (van Oldenborgh *et al.*, 2013), in Europe they are likely influential on the recent warming trends (van Oldenborgh *et al.*, 2009; Ruckstuhl *et al.*, 2008). If the reduction in aerosols are in fact one of the main causes of European warming, then this would weaken the conclusion that biases in long term RCM seasonal trends are due to inherent systematic misrepresentation of physical processes, but instead are due to inaccurate external boundary conditions. The alternative is that these biases can be explained by systematic deficiencies in RCM process representation. It is not the purpose of this section to consider which of these two possibilities is the case, but it is important to consider whether or not bias non-stationarity can be inferred from differences in long-term trends. From the above in the literature it is considered reasonable to assume that changes in systematic bias are due to inherent RCM construction, for the purposes of providing an upper bound for the degree to which RCM bias characteristics may be non-stationary in historical simulations. The findings from Figures 6.2 and 6.3 concur that winter trends are better reproduced that summer trends. Whether or not one may infer anything relating to the scenario that external forcings such as the reduction in aerosols from 1960-2000 may have influenced these trends is difficult to determine without further detailed analysis.

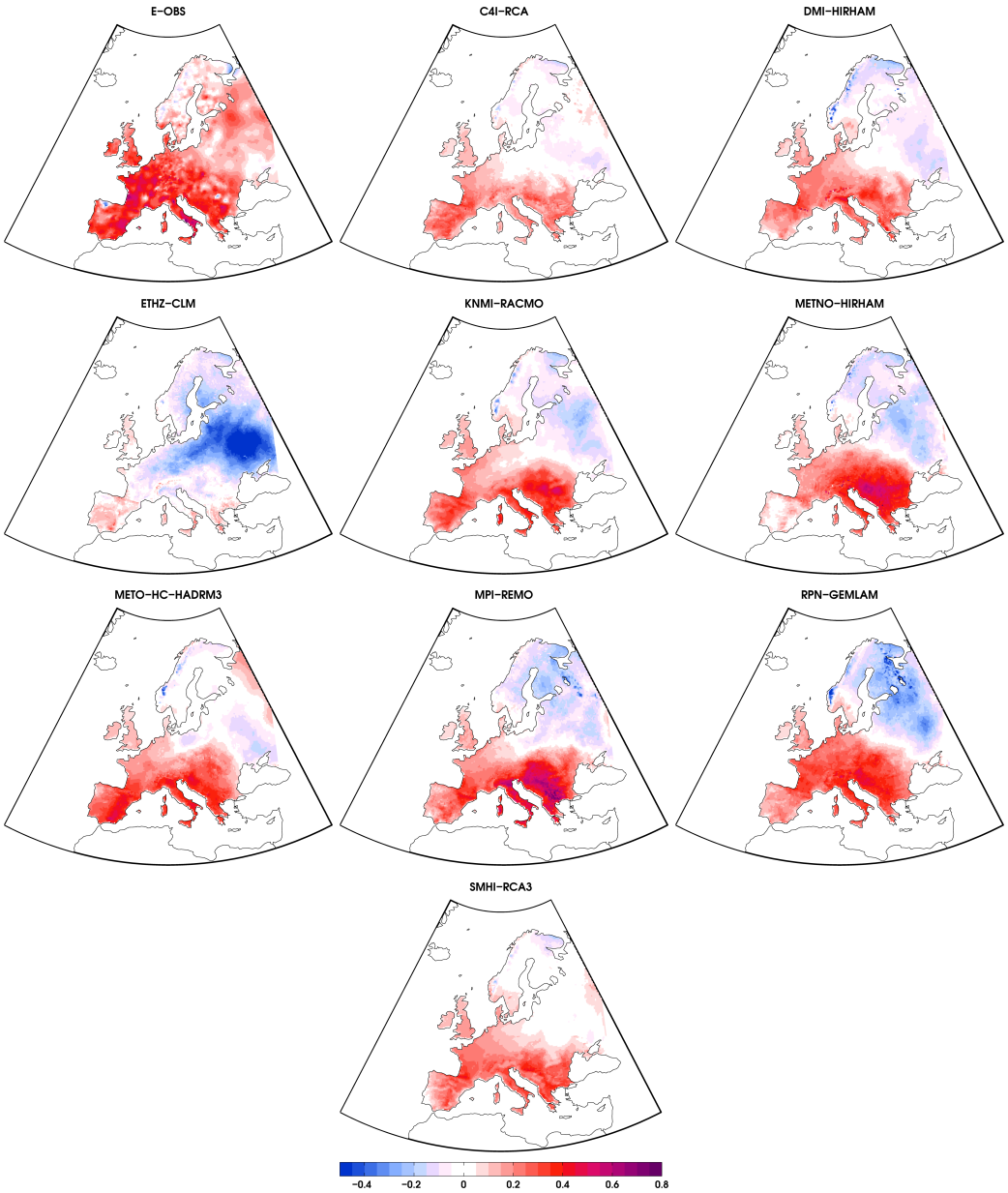

Figure 6.3: ERA-40 forced ENSEMBLES RCM and E-OBS summer temperature trend (°C/dec) 1961-2000.

The ENSEMBLES RCMs considered here and in previous chapters have common patterns of systematic seasonal mean temperature bias (Figure 6.4). In summer months RCMs overestimate southern temperatures predominantly as a result of soil-moisture feedbacks (Fischer *et al.*, 2007). Northern regions are generally closer to observations, although a slight cold bias is apparent in northern Scandinavia. Winter months see a flip in this pattern, with southern and high altitude areas simulated systematically cold whereas warm biases of up to 4°C occur in northern regions of low altitude. The preponderance of RCMs and GCMs to overestimate summer Mediterranean temperatures has been investigated in detail

and the conclusion drawn is that future temperatures may be too warm based on temperature dependent biases (Christensen *et al.*, 2008; Christensen and Boberg, 2012). This argument relies on the stationarity assumption that the historical bias distribution will remain the same in time, and also on the assumption that extrapolated biases beyond that occurring in the historical period will also follow a similar pattern. This secondary assumption is questioned by Bellprat *et al.* (2013) who suggest that because in some cases the cause of model bias may have a limiting factor (e.g. disappearance of snow cover/soil drying out) then one cannot simply apply a linear extrapolation beyond what would be physically plausible. This in turn raises another point; that the stationarity of summer biases in general is to a much greater extent unquantifiable since the temperatures attained in future projections are not reached in historical simulations, and therefore there is nothing to compare these futures to. On the other hand, the coldest monthly winter temperature biases are likely to become less relevant in a warming climate, and thus even if these biases are non-stationary, the temperatures at which these biases are relevant do not occur in future projections. In other words, the stationarity of future winter temperatures are to be assessed against historical mild autumn/spring and warm winter temperatures, not cold winters.

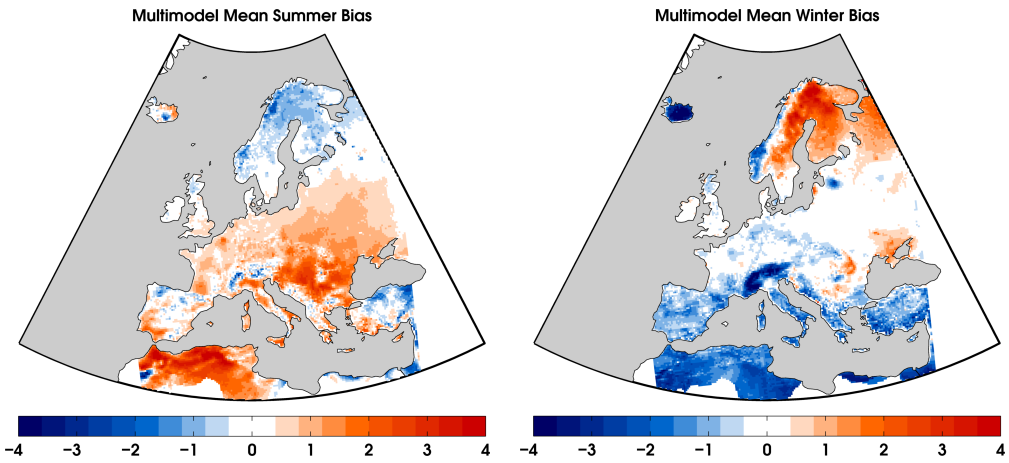

Figure 6.4: ENSEMBLES multi-model mean temperature bias against E-OBS 1961-2000 for Summer (JJA) and Winter (DJF) months (°C)

Investigating specifically the RCM biases occurring in Scandinavia in winter months, it is helpful to separate the region into those parts that are systematically warm and cold in winter, since as a whole the full region is not homogeneous (Figure 6.4). To go about this, warm and cold regions are identified from the multi-model mean winter climatology and spatial average timeseries are derived

from the model and observational data. In these quantile-quantile plots, the diagonal line indicates where an RCM would have zero bias for each temperature percentile. For the systematically cold fjord region first, RCM biases are relatively constant throughout the annual distribution although the multi-model mean bias is greater in winter months than in summer (Figure 6.4). Here grid points from Scandinavia are used as an example as in this region it is likely that the effect of climate change on snow and ice melts will influence bias characteristics far more than in more temperate regions of Europe. The fact that the RCMs are still cold in summer months might indicate that this bias is not heavily related to snow-albedo feedbacks (which might account for the slightly colder winter months) but more generally are due to other processes such as excessive cloud cover and a lack of downward shortwave radiation. The biases in this more mountainous region might reasonably be considered not overly temperature dependent. This in turn might suggest that in these regions biases may be less prone to non-stationarity in future projections. The lower regions which are systematically warm in winter months on the other hand might be due to a lack of precipitation (in the form of snow) in winter months or possibly a lack of cloud cover. Regardless, in this region biases may be more prone to non-stationarities due to the climatic characteristics of the region being more sensitive to the season. One important point however is that this region is projected to warm by 5°C in ECHAM5 ENSEMBLES and 8°C in CORDEX RCP8.5 projections, which may render the winter temperature biases less relevant, as they will not occur as frequently (see Figure 6.1).

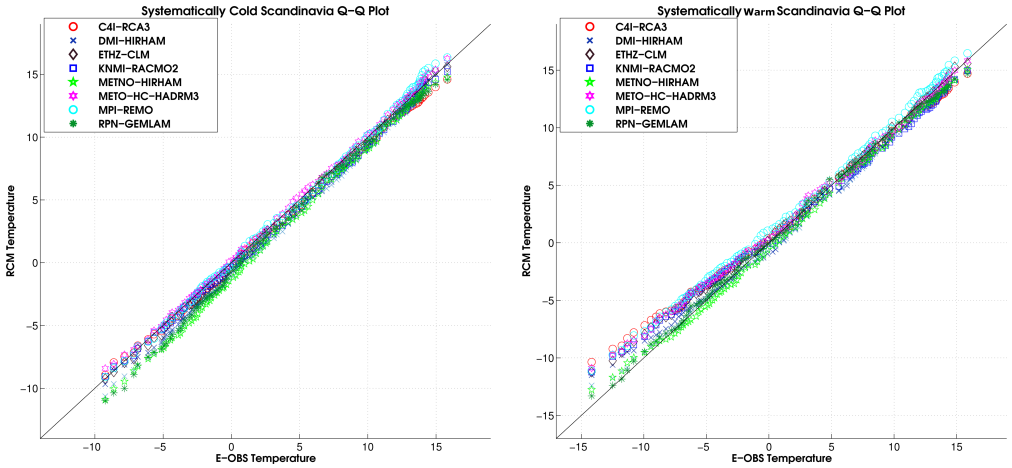

Figure 6.5: Quantile-quantile plots for nine ENSEMBLES ERA-40 forced RCMs vs E-OBS observations (°C) for systematically cold (left) and warm (right) regions. Each coloured q-q line indicates the temperature bias of the RCM at the corresponding E-OBS observational temperature.

The assessment of changes in mean seasonal bias and the stationarity of temperature percentile biases relative to E-OBS is given in Figure 6.6. For each gridpoint, the mean bias change and the average percentile bias change is calculated, and map plots are produced. The spatial patterns given in both methods are similar, although there are important differences between the left column (mean bias change) and right (percentile bias change). In summer months (top row of Figure 6.6), the mean bias change approach of Maraun (2012) produces a slightly larger change in bias than for the top right percentile bias change method. This latter method indicates low non-stationarity in most of central and southern Europe. This should be an encouraging sign that despite RCMs systematically underrepresenting the historical long term warming trend, particularly in summer, biases are quite robust when considering multidecadal future projections. In winter however, a higher degree of non-stationarity is found when assessing the percentile bias changes (Figure 6.6 right-bottom) over mean bias changes (Figure 6.6 left-bottom) particularly over western, central and northern areas. The reason for this is due to the RCM's mean winter warming being similar, but when investigated more closely the percentile biases in fact change more than a mean change would suggest. This would indicate that even if two RCMs have similar warming signals, this might not be a robust indicator of bias stationarity. It should be noted however, that these results are derived from monthly, and not daily, temperature data, and as such extremes are not considered here.

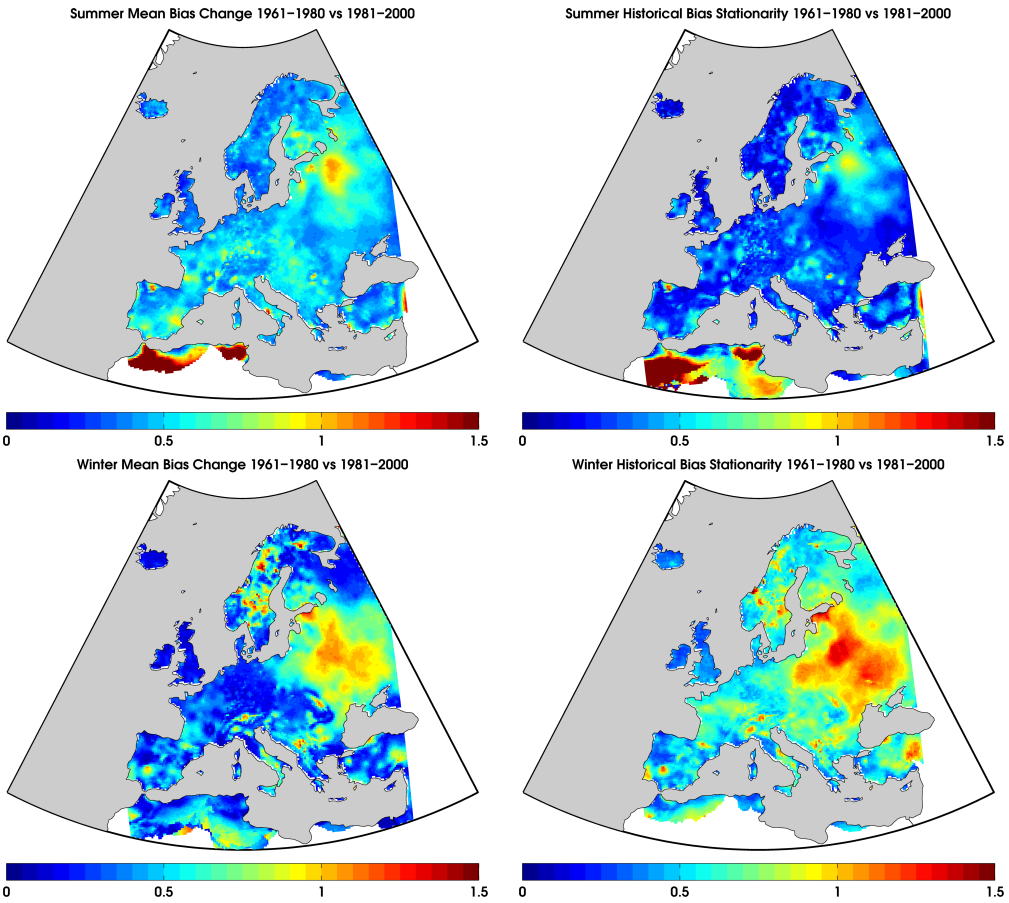

Figure 6.6: ERA-40 forced ENSEMBLES mean seasonal bias vs E-OBS change (°C) 1961-1980 vs 1981-2000 (left column) and percentile bias stationarity ($°C^2$) (right column).

To summarise, since the causes of historical RCM biases are dependent on a variety of factors - SST and atmospheric forcing, GHG/aerosol spatial and temporal concentrations and model construction/parametrisation approaches - it is difficult to assess precisely to what degree the third group of causes may be sensitive to future changes in regional climate. This is the key point when trying to assess bias stationarity, but since it is almost impossible to isolate in historical simulations one must be content with producing upper bounds of bias non-stationarity. Given that during the four decades of the ERA-40 forced ENSEMBLES simulations percentile biases may change by up to 1.5°C, then it is essential to go further to assess what the future potential for bias non-stationarity may be when European climate warms by significantly more.

6.4 Assessment of Future Projection Bias Stationarity

As Maraun (2012) states, it should not be surprising that the local climate sensitivity will be different among RCMs due to the fact that their representations of land surface and atmospheric processes are similarly varied. This is one of the main benefits of using multi-model ensembles by spanning a wider range of plausible future climate changes scenarios (Tebaldi and Knutti, 2007). Furthermore, although there are no future observations with which to assess potential bias-nonstationarity, one does not need to consider external forcing factors (e.g. aerosols) when using a pseudo-reality framework. This is because each RCM is forced with identical boundary conditions for the full historical and future projection periods. One can fairly judge that if an RCM changes its simulation characteristics relative to another RCM then that is by definition a change in bias.

The ENSEMBLES RCM's delta-t summer and winter temperature changes from 1971-2000 to 2070-2099 are given in Figures 6.7 and 6.8. For summer, the land surface warming is dictated predominantly by the degree of land-surface moisture availability (Joshi *et al.*, 2008); more moisture leads to SW radiation being transferred as latent, rather than sensible heating, whereas, in this case, a RCM soil-moisture deficiency would lead to an increased delta-t signal. The logical inference from these RCMs is that in the case of DMI-HIRHAM it would be likely to have substantially greater soil-moisture levels than the remaining three RCMs, all of which warm by very similar amounts throughout Europe.

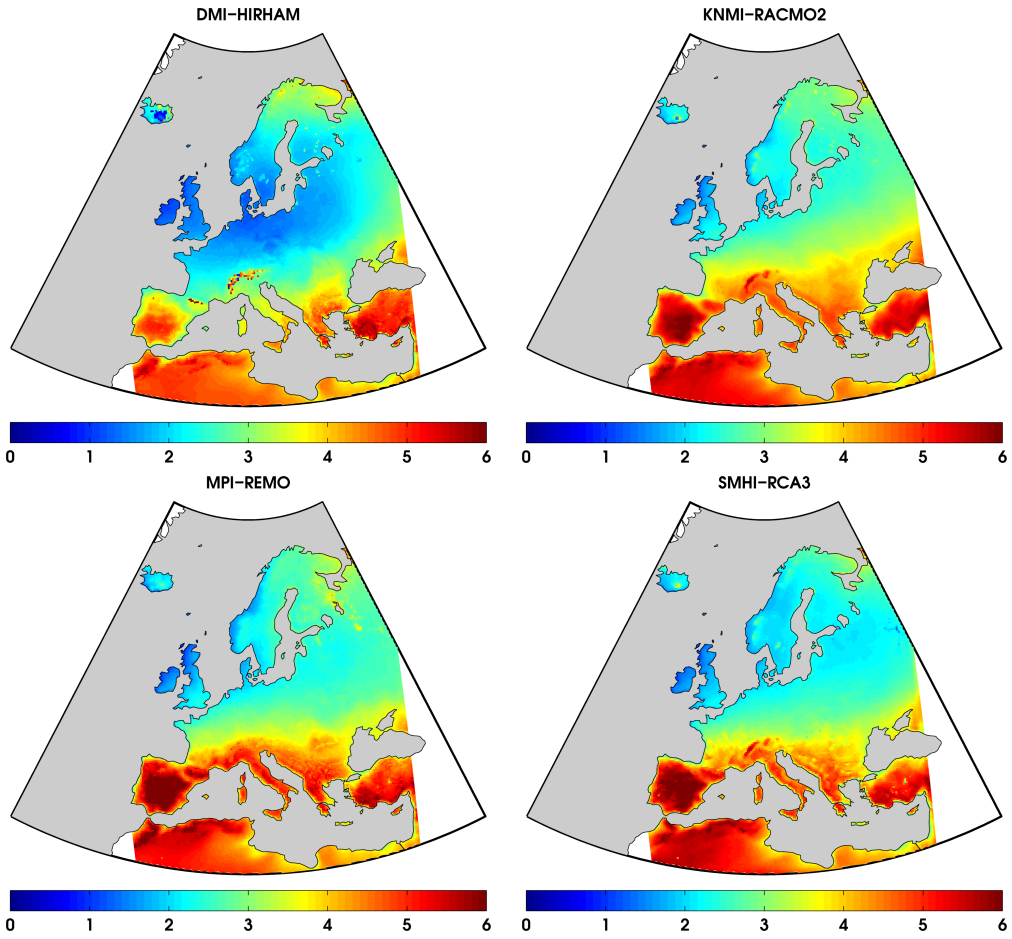

Figure 6.7: ENSEMBLES summer delta-t temperature change (°C) 2070-2099 vs 1971-2000

In winter months (Figure 6.8), large-scale atmospheric flows tend to dominate the level of warming rather than the small-scale land surface and cloud processes of summer (van Oldenborgh *et al.*, 2013). However, in high altitude regions, such as the Alps, Pyrenees, Atlas and Turkish mountain ranges, other processes such as snow-albedo feedbacks become more important (Maraun, 2012). It is notable that there is a wide spread warming across the RCMs in the Alps in particular for these four RCMs; SMHI-RCA3 warming up to 5°C whereas DMI-HIRHAM has negligible additional warming compared to the surrounding region (Figure 6.8 top-left plot). There is substantial disagreement among the RCMs therefore as to how the processes most determining of the Alpine climate will change in future, which in turn will affect the local climate change signal. Figure 6.9 shows annual surface albedo (SA) timeseries for the Alps region (as defined in the 'Rockel' regions used in Chapter 4) for those areas above 600m (to distinguish the high-altitude areas from the surrounding low-land). It is clear that the cause of DMI-HIRHAM failing to show a distinction in this region to the surrounding area is because of its failure

to simulate any substantial temporal variability and produce a long term trend in SA (Figure 6.9, blue timeseries). Things are not as clear for the other RCMs however. Although the fact that SMHI-RCA3 has the highest winter delta-t temperature change in the Alps (Figure 6.8) might be explained by its comparatively high historical SA and reduction thereafter in the future projection (Figure 6.9, turquoise timeseries). However, KNMI-RACMO2 has virtually the same SA magnitude, interannual variability and long term trend yet does not reproduce any strong warming in the Alps. MPI-REMO has a similar size of SA reduction, although starting at a lower magnitude than SMHI-RCA3, yet simulates moderate warming. Finding a way to link a process based understanding of how RCM biases may change in the future, as suggested by Bellprat *et al.* (2013), might be a more reliable approach in principle, yet in practice the system may be too non-linear to describe simply and be generally applicable for all RCMs.

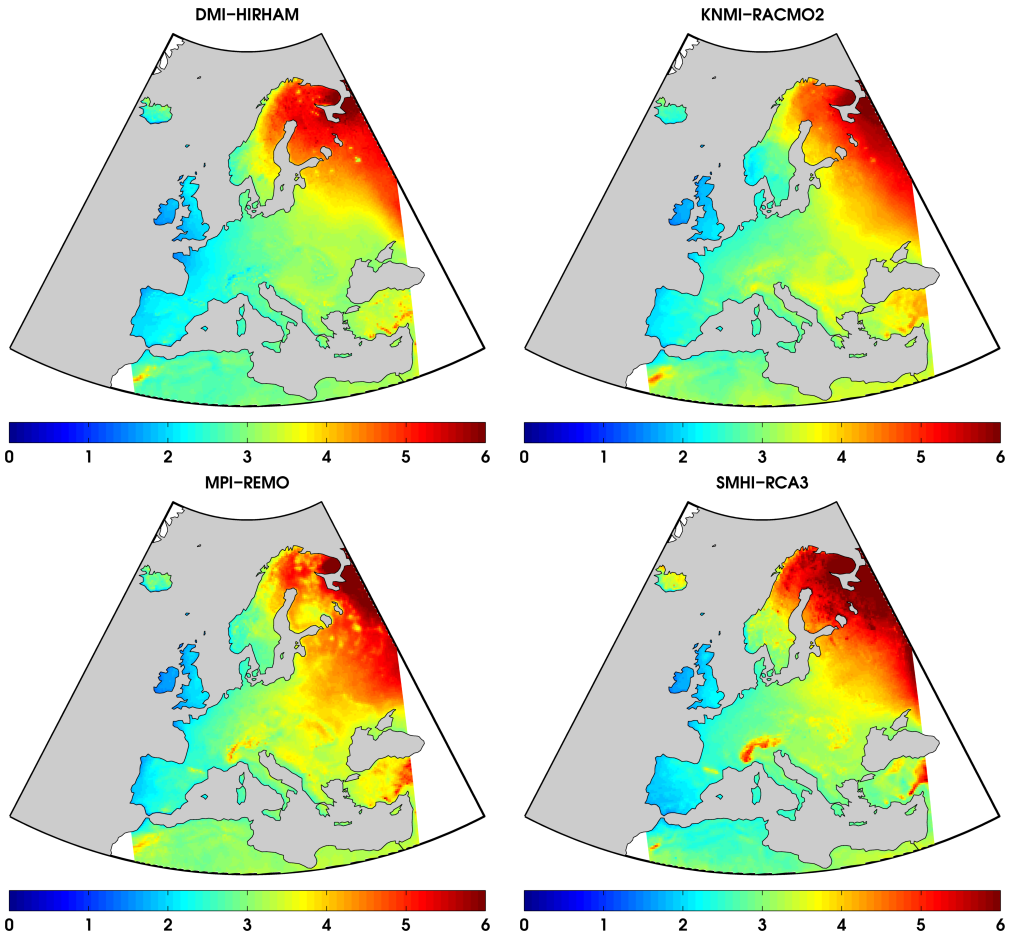

Figure 6.8: ENSEMBLES winter delta-t temperature change (°C) 2070-2099 vs 1971-2000

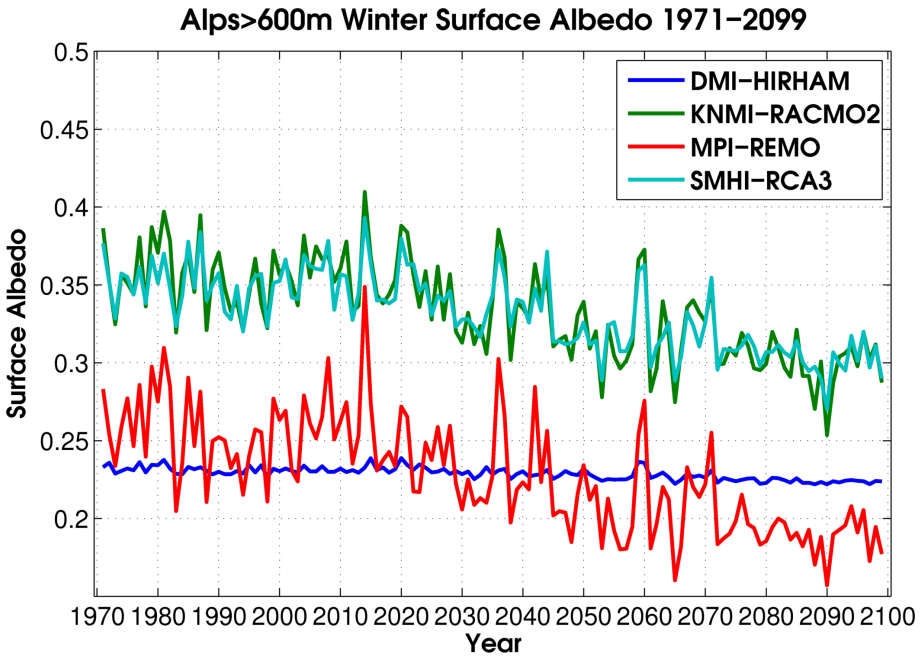

Figure 6.9: Winter Alps spatial average surface albedo annual timeseries of ECHAM5 forced RCMs for regions above 600m altitude.

Although the ECHAM5-forced RCMs show a large spread in local climate change signal in the Alps, this does not necessarily imply that there is substantial bias non-stationarity present. Figure 6.10 shows both the standard mean relative bias change as per Maraun (2012) and the bias stationarity as calculated from the change in percentile biases. In summer, as seen in Figure 6.7, there is moderate agreement among the RCMs in eastern and northern Europe, with larger disparities appearing in southern France and the Iberian Peninsula. When assessed with the bias stationarity metric however, these differences in climate change signal are not that impactful on the assessed changes in percentile biases, indicating that bias correction methods would likely be successful here. In some regions such as the Balkans and Turkey, although the RCMs agree more on the future signal, there is an increase in bias non-stationarity. For winter, the clear difference between the RCMs in Alps temperature signal is the most striking feature for most of Europe, except for the differences around the Barents Sea where sea ice levels and surface albedo levels change substantially. The bias stationarity metric on the other hand, as in the historical assessment, does not consider the Alps region to be particularly important, but generally increases the level of bias non-stationarity for most central and northern European areas.

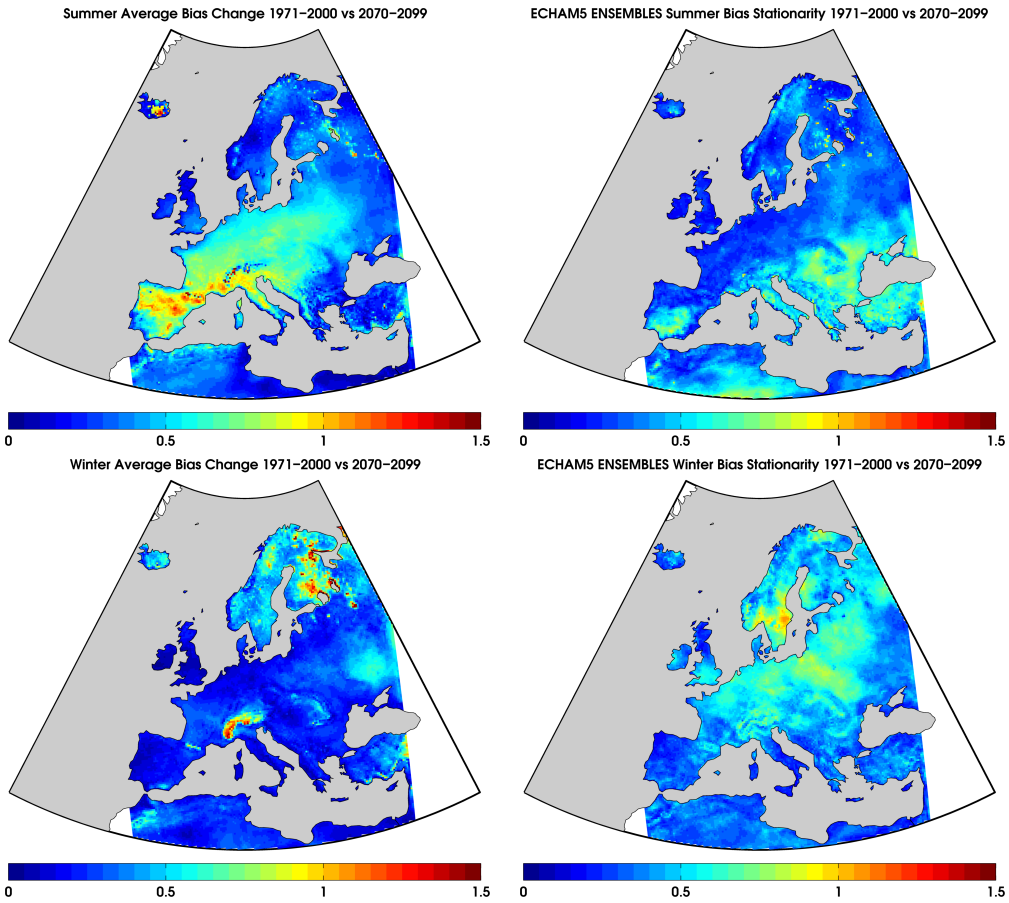

Figure 6.10: ECHAM5 forced ENSEMBLES summer and winter average relative bias change (°C) (left column) and bias stationarity (right column) ($°C^2$) for 1971-2000 vs 2070-2099. In the left hand side maps, for each grid point, each RCM pair is evaluated in their historical and future projection differences, and the average of these values is displayed as the average mean bias change. For the right hand side, for each grid point percentile areas are calculated for the same RCM pairs, and the average result plotted.

For CORDEX RCMs, the difference between the simple mean relative bias change and the stationarity metric for summer (Figure 6.11 top row) is substantially larger than for ENSEMBLES RCMs. The divergence in warming signals for southern France is not considered that large in terms of percentile bias non-stationarity. On the other hand, in regions where CCLMcom-CLM4, KNMI-RACMO22E and SMHI-RCA4 agree on the warming signal the degree of non-stationarity is large, particularly in the Balkans into Turkey, Northern Africa and Russia. This differs from the previous findings, as the stationarity of biases in summer months was previously considered more robust than winter months. Scandinavia has quite stationary percentile biases in this case. For winter months, a more moderate picture emerges, with the large discrepancy between the RCMs Alps warming signals - similarly as with ENSEMBLES - not producing large non-stationary percentile biases. Other regions such as Turkey also have more stationary biases than the

simple DJF historical - projection comparison would suggest.

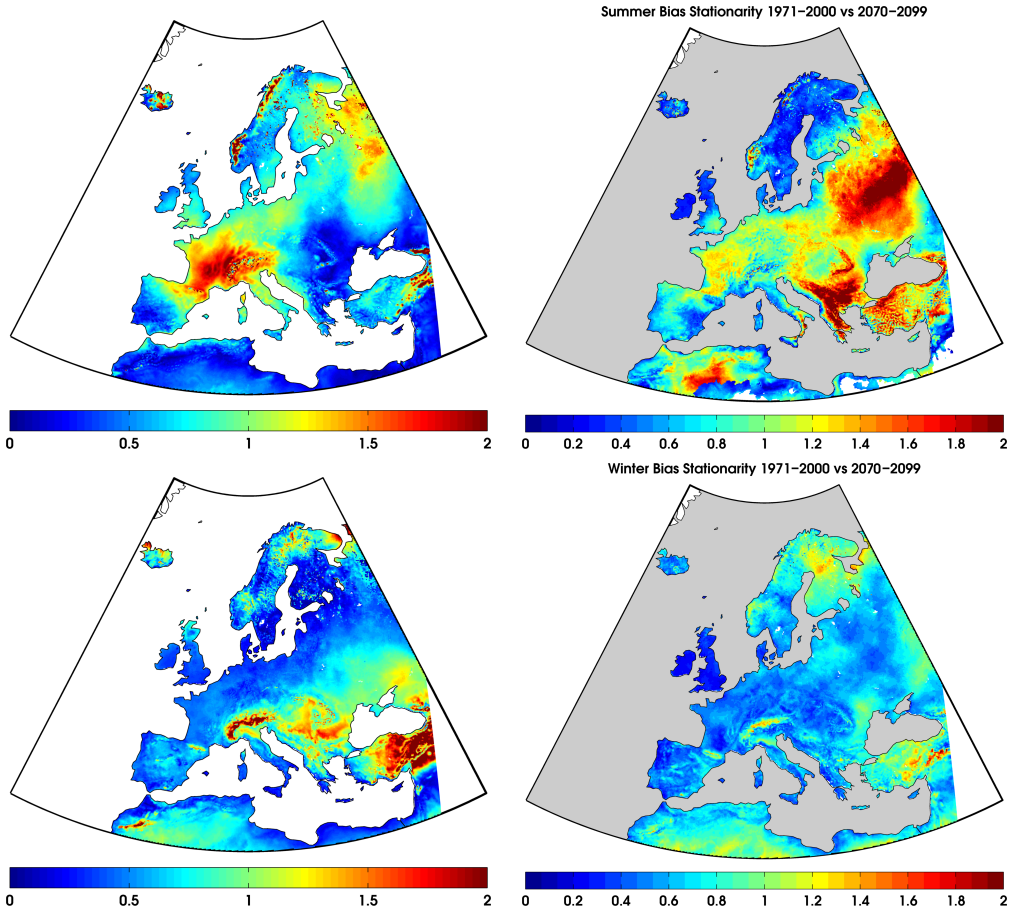

Figure 6.11: HADGEM2-ES forced CORDEX summer (top row) and winter (bottom row) average relative bias change (°C) (left column) and bias stationarity (right column) ($°C^2$) for 1971-2000 vs 2070-2099. In the left hand side maps, for each grid point, each RCM pair is evaluated in their historical and future projection differences, and the average of these values is displayed as the average mean bias change. For the right hand side, for each grid point percentile areas are calculated for the same RCM pairs, and the average result plotted.

What these results show first is that although bias non-stationarities may be occurring, they may not be as large in magnitude as some other methods may lead one to conclude. The main problem with this conclusion stems from the fact that pseudo-reality analysis will always be relative to the models within each ensemble. ENSEMBLES was found to be less likely to have non-stationarity bias characteristics than CORDEX through comparison of Figures 6.10 and 6.11. A conclusion that CORDEX RCMs require more bias correction may not be the correct interpretation however; these result should be considered as guidance for regions in which bias non-stationarity may be more of an issue for users. Some regions that have been found to be potentially susceptible for non-stationarity, namely eastern Europe in

summer CORDEX simulations, were not possible to identify before. This analysis may provide a method to look deeper into the way that RCM temperature distributions change over time relative both to historical observations, and other ensemble members.

6.4.1 Autumn vs Spring Bias Characteristics

One simple test which can be used to test whether the most basic interpretation of the stationarity assumption holds is whether percentile biases are the same throughout the annual cycle. Although one might expect RCM bias characteristics in Autumn and Spring to be dissimilar given that there are difference in the land surface (e.g. snow cover and vegetation changes) and atmospheric properties, the bias stationarity assumption in its usual form makes no such distinction. This section will aim to identify regions where the underlying regional climatic characteristics are more prone to non-stationary biases. It should be noted however, that bias correction methods can account for the seasonal cycle by correcting daily data over a (say) 30-day window. Without this, any correction function would be defined based on the bias characteristics of the whole approximated annual cycle and thus would not distinguish between seasons. For example, if an RCM has, say, a bias of $+2°$C when observations are 5°C in autumn, then according to the stationarity assumption this bias should also be the same in spring when observations are at 5°C (or any other season). This is only in the case of where one is considering percentile biases, and not mean bias, since of course it is highly likely that the autumn and spring climatological means are different.

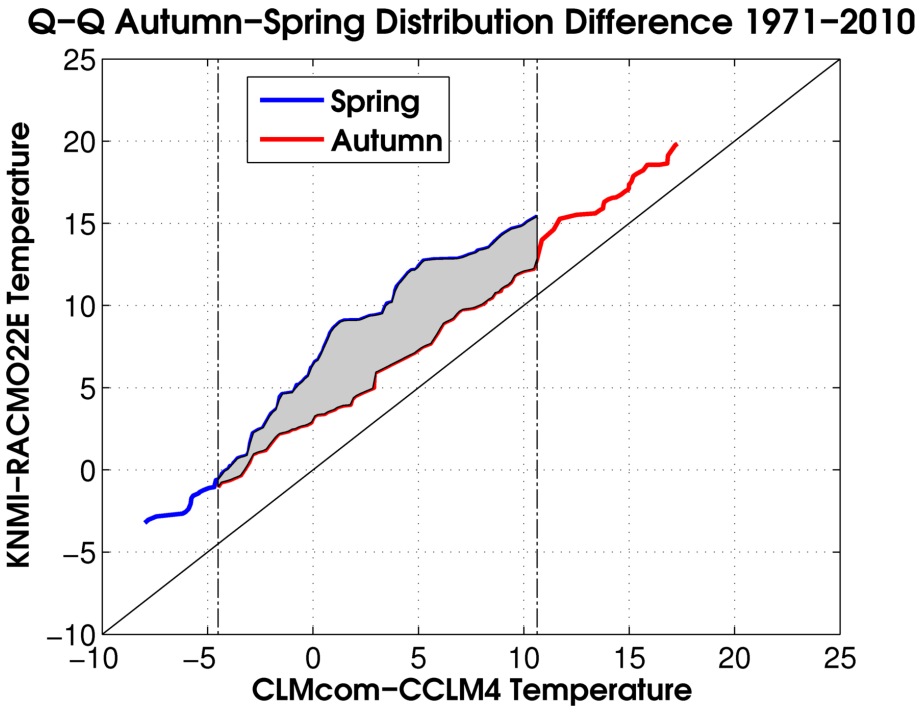

Figure 6.12: Example diagram for a single gridpoint of Autumn and Spring quantile-quantile lines (°C). Grey region is area of comparison of Spring (blue) to Autumn (red) months, when using CLMcom-CCLM4 as 'pseudo-observations'. The corresponding area of this can be used to quantify the degree of non-stationarity of RCM bias distribution. Completely stationary bias characteristics would lead to this region being of zero area.

This autumn/winter bias distribution difference test (Figure 6.12) for single gridpoint example) is done both for the historical 1971-2000 and future projection 2070-2099 periods (Fig 6.13). First, the ENSEMBLES RCMs forced by ECHAM5 (top-left 6.13) show that the percentile bias difference between Autumn and Spring months is less than reasonably low for the majority of regions, being less than 1°C average percentile bias change) except for some large differences in the high Alps and southern Finland. This pattern does not change when considering the future projection period (top-right of 6.13), indicating that the cause of the difference between the two seasons bias characteristics does not change. In other words, the underlying differences between the two seasons in terms of European climatic properties (land surface etc.) do not change substantially. For ENSEMBLES RCMs therefore, the conclusion would be that seasonal bias correction is necessary for future projections, as the qualitative differences between seasons remains from 1970-2099.

In the CORDEX RCMs forced by HADGEM2-ES however (bottom row 6.13), although there are more regions where in the historical period the bias stationarity assumption is less plausible (i.e more non-stationary biases in the Alps, Turkey,

Russia, Pyrenees and Atlas Mountains) these seasonal differences shrink when moving into the future projections as can be seen by the lower bias-stationarity values in the future projection bottom-right plot of Figure 6.13 compared to the historical bottom-left plot. This suggests that the bias stationarity assumption becomes more reasonable over time; that is, the description of seasonal biases is better approximated by the annual cycle in the future projections. This should occur if the seasonal differences in regional climatic properties shrink due to climate change (e.g less snow in the Alps in spring relative to autumn months). There are some areas where this convergence of spring and autumn bias characteristics does not occur in CORDEX RCMs (such as in western coastal Norway), but this behaviour generally holds across the region.

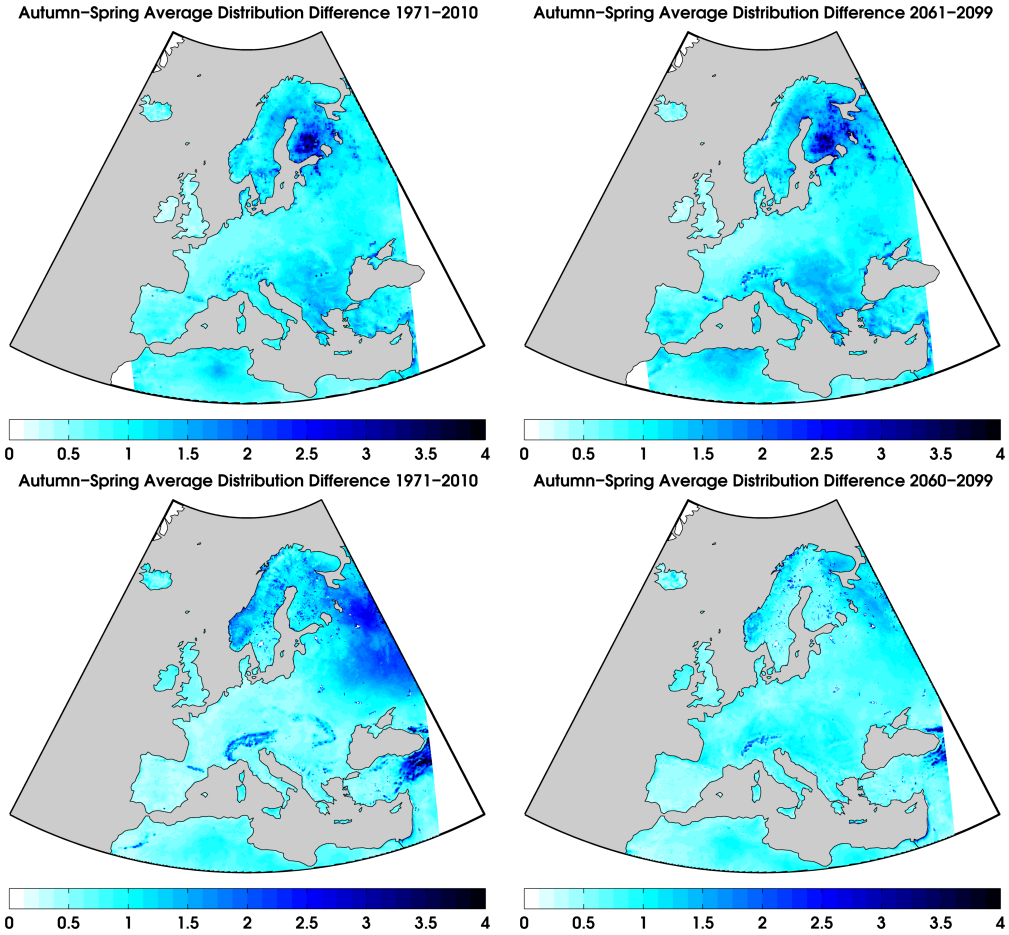

Figure 6.13: ENSEMBLES (top row) and CORDEX (bottom row) Autumn/Spring bias distribution similarity over 1971-2000 and future similarity for 2070-2099 ($°C^2$)

6.5 Conclusions

In this chapter, the stationarity assumption has been considered in the context of historical and future projections forced by GCMs. One first comment is that there is more than one way to interpret future projection biases; given that there are no observations with which to truly assess the assumption it should not be a surprise that the interpretation of results remains somewhat subjective. Historical mean biases are commonly evaluated as a difference of climatological averages, but this approach does not take into account how the stationarity assumption is dealt with specifically by correction methods. Given that quantile mapping approaches tend to outperform other more simple approaches (e.g. Terink *et al.*, 2010; Teutschbein and Seibert, 2012), it was considered reasonable to use a percentile bias method with which to evaluate bias changes over time.

First on historical bias stationarities, quantifying this is always difficult given the numerous factors to consider in whether RCMs forced by reanalysis are being provided the correct information to begin with. An inability to determine whether systematic RCM biases are a result of biases inherent to the RCM rather than inherited boundary condition error is a limitation of the analysis, but these approaches to quantifying potential regions of non-stationary biases provide a reasonable starting point for guidance. RCMs typically do not show very strong changes in bias over the historical period, although the two methods used (mean bias change and percentile bias change) produce different results due to their testing dissimilar portions of the temperature distribution. RCM biases may be more stationary than the previous mean bias change method of Maraun (2012) suggests, although this result is necessarily limited to the range of temperatures that occur in both split-samples (or both historical and future projection periods for the following section). Although proving or disproving bias stationarity is in principle impossible to determine, the methods used in this chapter may offer some ways in which to identify regions where bias non-stationarity should be considered. Assessing the stationarity of assumptions relating to model biases that have not occurred yet (i.e. temperatures outside the range of historical simulations) is a question that is even further beyond the first proposition. Approaches such as suggested by Bellprat *et al.* (2013) may be of use in regions where RCM biases can be linked to a particular physical process (e.g. soil-moisture/surface-albedo feedbacks), but validation of such methods poses a similar challenge to the more general assumption considered here.

Given the difficulties in identifying changes in historical reanalysis forced RCM simulations, pseudo-reality simulations offer an idealised experiment providing 'true' boundary conditions with 'pseudo-observations', although with the fundamental problem of subjective interpretation of results, both with respect to the method and to the RCM ensemble still built in. Finally, this analysis was done only considering temperature; other variables such as precipitation may also be considered with such methods described to determine regions of particular sensitivity to non-stationary biases.

Chapter 7

Conclusions, Recommendations and Outlook

7.1 Introduction

The aims of the thesis laid out in Chapter 1.4 are the development of objective and robust approaches for the evaluation and assessment of RCMs using performance metrics and to provide guidance to relevant groups as to how best utilise these metrics. This involves investigation and exploration of three main objectives. The first two objectives are focussed on the robustness of performance metrics, both individually and in combination, the third objective relates to the degree to which historical bias characteristics are stationary in future projections. The third aim is to develop and provide guidance on the use of performance metrics more broadly. These three objectives are principally analytical in nature, and are investigated within Chapters 4, 5 and 6 in the thesis.

Although the analysis approaches used to answer these questions of the robustness of metric use and application are predominantly quantitative in nature, they remain open to qualitative interpretation. This reflects the fact that there can never be a completely objective approach to climate model evaluation. Although different evaluation approaches may be preferred on subjective grounds, such as availability, quality or length of observational data, model resolution and the region under consideration, the analysis indicates that there may be objective reasons to avoid or prefer certain approaches over others. Furthermore, the utility of performance metric evaluations may also depend upon what they are intended to be applied to. Metrics used in standard historical validation tests for the purposes of assessing

general model quality might be not as appropriate for use in weighting future projections for example.

7.2 Summary of Findings

7.2.1 Chapter 4: Performance Metric Sensitivity

Performance metrics are quantitative measures of model performance of a specific variable relative to observations. They can be considered to provide an objective alternative to more qualitative model evaluation approaches, such as the interpretation of mean climatological difference maps. For this analysis in Chapter 4.1, performance metrics are defined by four elements; variable, domain (space/time), statistic and observational dataset. Since a purpose built high-quality gridded observational dataset was constructed specifically for the ENSEMBLES project in E-OBS, only the first three of these aspects were considered for this Chapter 4. For regions such as Africa where the observational uncertainty is greater, including this element would be recommended as a priority however.

The nine RCMs used from the ENSEMBLES project were assessed in their performance relative to the E-OBS dataset for 16 variables (Tmean, Tmax, Tmin, DTR, CSDI, WSDI, FD, ID, Pr, CDD, CWD, R10mm, R20mm, Rx1day, Rx5day and SLP), 9 domains (Europe and 8 sub-regions), and 13 statistics (RMSE, MAE, Standard Deviation, Index of Agreement, Spatial Skill Score, Correlation, Spatial Skill Metric, Annual Cycle Skill Score, Annual Variability Metric, Interannual Variability Metric, Linear Trends Metric, PDF Skill Score, CDF Metric). The general conclusions from Chapter 4.2 for the first aspect of variable sensitivity are that the overall sensitivity is quite low, although there are differences when the whole variable distribution is considered. For example, there was found to be some change in performance from one RCM to another when considering Tmax to Tmin or Tmean. For precipitation related variables, similarly low sensitivity was found, indicating that the extreme indices used are somewhat analogous to one another in many respects, thus one alone might be sufficient for an assessment of RCM precipitation skill for this ensemble at least. For the temporal sensitivity of metrics, it was found that metrics summarising RCM performance over annual time periods may lose substantial information, most clearly where errors in different seasons cancel out. Shorter timescale assessments of RCM performance over

10-year periods were found to be robust relative to the full 40-year ERA-40 period. Subdomain performance for precipitation was found to be robust indicating that more small-scale evaluations are likely consistent with the overall RCM quality. Temperature performance on the other hand is less homogeneous over Europe, although not overly so. Finally, the sensitivity to statistic was dependent on the specific aspect under consideration; for spatial patterns the choice of statistic made negligible difference to the assessed performance, whereas for temporal characteristics they are more divergent on their assessment of model quality.

One of the main uncertainties in discussing these results is to what extent they are likely to hold in a different domain, for different variables or ensembles. With respect to the sensitivity of performance metrics, the most likely conclusion to hold is the results relating to the choice of statistic, since these are for the most part mathematical findings based on the relationship of one equation to another. For example, RMSE will always produce close results to MAE, and as such one can have high confidence in selecting statistics based on the findings shown in this thesis. Other findings however, such as those regarding variable choice are most susceptible to become more susceptible to changes, for example for regions which experience monsoon conditions clearly more extreme indices will become more necessary than for Europe. As a result, such analysis into variable sensitivity should be re-run. As previously mentioned, sensitivity to the choice of observational data may become much more a significant factor depending on the availability of high quality datasets. Finally, with regard to whether these findings for Europe would apply for a different RCM ensemble in Europe, it is very likely that the answer would be yes, considering the fact that it remains the same climatic conditions, regions and available variables.

7.2.2 Chapter 5: Metric Combinations

After completing individual quantitative assessments of RCM performance with metrics, it is possible to combine them into a single overall score, here referred to as a Generalised Performance Indicator (GPI). This construction has two components; the choice of combination method, and the choice and number of metrics to include. Four combination methods were used (multiplicative, geometric, harmonic and ranking) to assess the first of these aspects, and the findings given in Chapter 5.3 indicate that this choice may have a qualitative effect on how the performance level

between RCMs is interpreted. However, quantitatively there was little difference between the methods when assessed over the many metric permutations. Therefore although using a combination method may provide a comprehensive assessment in a single value, this value is not sensitive to the method used to combine the metrics. The second factor assessed is the type and number of metrics included, assessed in Chapter 5.4 and 5.5. It was found that as the number of metrics included increases, the range of GPI values converge, although this may be a statistical artefact of random subsampling of a set of numbers. However, sampling more independent variables, and not highly correlated ones (e.g. Tmean/Tmax) may be the source of largest uncertainty due to the potential of GPI values to change substantially when including further such variables in combination.

Of all the analyses, the findings from the Chapter 5 are the most likely to be robust when considering applications to alternative domains, variables or ensembles. This fundamentally is due to the nature of the analysis, in that the results are broadly mathematical, and as such are not as dependent on the underlying climatic behaviour as other investigations. For example, the findings relating to the choice, number and type of metric to be used within a GPI are found through statistical assessments and redundancy analyses, which are likely to be insensitive to changes in domain. The combination methods investigated for example, will maintain their properties relative to one another regardless of the application. Existing GPI methods, such as those by Murphy *et al.* (2004); Reichler and Kim (2008); Giorgi and Mearns (2002), should consider applying a wider range of combination procedures, since this was found to produce substantial variation in GPI output. Furthermore, GPIs should aim to span not just several variables, but also statistics.

7.2.3 Chapter 6: Stationarity Assumption

RCM temperature biases assessed by comparison to historical observations can be described by the relationship between the two distributions. The stationarity assumption (SA) in essence assumes that the nature of this relationship is time-invariant, that is that the RCM projection will continue to exhibit identical bias behaviour. Since this assumption is not directly verifiable due to the lack of future observations, only a inferential approach may be taken to investigate further. Thus a 'pseudo-reality' framework is adopted whereby several RCMs forced by the same GCM are used in turn as the future 'truth'. By doing so, whilst implicitly assuming

that the relationship between RCM and RCM is the same as RCM to reality (Whetton *et al.*, 2007), one may examine the degree to which the SA is reasonable. An historical 'evaluation' period covering 1971-2000 and a future 'assessment' period covering 2070-2099 are used with RCMs from both the ENSEMBLES and CORDEX projects.

The results from this final analysis build on, yet raise questions of the work of Maraun (2012) in their pseudo-reality analysis of ENSEMBLES RCMs using a simple mean-bias change approach. This method involved the assessment of historical mean climatological bias from one RCM to the 'pseudo-observation' RCM and calculated the difference to the future projection mean bias, thus providing an indication of the mean-bias change. A new method is proposed in this analysis in Chapter 6.3 and 6.4, more in line with how bias correction is commonly undertaken using quantile-quantile mapping approaches (Teutschbein and Seibert, 2012). This method, referred to as the percentile bias change approach detailed in Chapter 6.2, involves the calculation of the temperature biases from one RCM to the 'pseudo-observational' RCM for each 0.5th percentile for both historical, and future projection periods, and the average change given by these two values plotted in map plots. By doing so, one may compare how these different methods of assessing bias non-stationarity affect the interpretation of how the RCMs behave relative to one another. As in Maraun (2012), RCMs were found to show potential non-stationary bias characteristics, but this is region, season and method dependent. By implementing a new approach, it was shown that in some cases bias-nonstationarity is less likely than suggested by using the method of Maraun (2012). For example, in winter months for ENSEMBLES RCMs a high degree of mean bias change was apparent (Figure 6.10), whereas using a percentile bias change approach this region was found to have negligible difference in the level of bias-nonstationarity found throughout Europe. Such a finding may suggest that one should be cautious when interpreting mean bias changes, since they may not capture how the bias distributions are or are not changing relative to one another.

One key point raised in this analysis is the fact that it only makes sense to refer to a change in a particular bias in 'pseudo-observational' temperature t between two periods if in fact that temperature t occurs in both periods. For example, the temperature in the coldest historical months may not occur in future and thus the question as whether that model's bias is stationary into the future does not arise. This scenario is mirrored in the case of the warmest future temperatures, which may

likewise not occur in the past. Thus it only makes sense from this percentile bias change perspective to focus on those 'pseudo-observational' temperatures t that occur in both the historical and future periods. Inferences regarding the stationarity of winter biases are thus likely more robust than for summer since it is required to further assume some form of bias extrapolation (e.g. Bellprat *et al.*, 2013) beyond the observed historical range of temperatures. For winter this is not required as future winter biases can be 'corrected' from historical winter/autumn/spring temperatures.

The underlying cause for biases to change from historical simulations to future projections was found to be due to a shift in the physical characteristics in that region. The region focused on in Chapter 6.4 of snow cover and the Alps demonstrates this point. An accurate modelling of the snow-albedo feedback, in a similar fashion to the soil-moisture/cloud cover/precipitation feedback, is essential to a reliable simulation of historical climate. Even if they are represented well, future changes in the surface properties can have a substantial effect on the future distribution of seasonal temperature bias, as seen in figures 6.8 and 6.9. Further work in this area could take the form suggested by Bellprat *et al.* (2013), who suggested the linking of bias correction methods to the underlying physical processes involved, instead of artificial statistical processing methods.

For this final stationarity analysis, the finding that RCM bias characteristics can change substantially in a temperate region such as Europe should lead one to be wary of this occurrence in all domains. However, since the largest discrepancies between RCMs only occur when a physical process is near a non-linear tipping point, such as snow melting leading to an albedo feedback or a depletion of soil moisture, such occurences will only be present in very particular regions. The benefit of the analysis undertaken here is to highlight where such factors might occur, and thus an awareness of bias non-stationarity should be raised. With regard to other variables, however, it is difficult to generalise too much given that the Chapter 6 analysis was only undertaken with temperature.

7.3 Implications for Regional Climate Model Communities and Users

One general comment should be made first regarding the transferability of the results of this thesis investigated using European RCMs to other regions and model ensembles. As regards the Chapter 4 analysis on the sensitivity of performance metrics, it is helpful to recall the definition provided in Chapter 1.3 of the components: variable, statistic, domain and observational dataset. Europe has a varied climate, spanning regions such as the temperate western coastal, colder mountainous alpine, dry hot Mediterranean and central and eastern continental regions of larger temperature variation from winter to summer. Missing however are the characteristics of extreme climates, such as equatorial tropics or arid dry landscapes. For the choice of variable, this factor may have consequences for the interpretation of results found in Chapter 4.2, since in such climates the use of extreme indices is likely to be of increased utility, and thus although it was found that these indices may have substantial redundancy in the European context, in other climates this may not be true. The findings for the sensitivity of statistics are likely to hold in other regions, since they are independent of the climatic characteristics of the domain in question. Moreover, the choice of domain may become less relevant for more climatically homogeneous regions. Observational quality however is highly likely to become more of a concern in other regions, given that the temporal and spatial density of data in Europe is particularly good.

What are the implications of these findings for different groups such as climate modellers, analysts and impact modellers, who use information from RCMs in both the historical and future context? Fundamentally, this thesis investigates the use of performance metrics for the interpretation of RCM errors and the potential further application of this information. For each of these groups, although performance metrics essentially are performing the same task (quantitative evaluation of model errors against observations), the utility of a metric as such derives from the purpose of the application.

One main finding is that one cannot assume that two metrics are providing independent information, as shown in Chapter 4.5 on metric redundancy. Much more emphasis therefore should be made on testing whether a proposed set of metrics is in fact achieving this 'independence'. This will have several benefits, such as not focussing too much model assessment on broadly the same behaviour,

given that the underlying processes are similar. Another benefit would be to ascertain if further proposed metrics may provide additional information not previously identified.

One novel approach taken in the thesis as outlined in Chapter 4.4 is the breaking down of the statistic choices into four types: standard error measures, spatial patterns, temporal and frequency. Any set of metrics used to assess RCMs would benefit from utilising metrics from all four categories, as they are found to provide independent information, increasing the objectivity of their selection (Chapter 4.4).

For the different RCM communities themselves, model validation groups may look for a wide set of metrics taking into account many physical variables for a process-based understanding of an RCM's performance Although the Chapter 4 analysis focus is on temperature, precipitation and sea-level pressure, the analysis in Chapter 5.5 indicates that further variables are highly likely to provide independent information. Analysts interested in assessing ensembles of RCMs for a more general overview of performance may not require as in depth an approach, instead requiring a reduced set of metrics providing the 'key points', strengths and weaknesses of the models. The findings of the Chapter 5 analysis on GPIs is that any combination method will likely be robust for this task, although those wishing to distinguish better or worse performing RCMs may wish to use the additive or harmonic methods respectively to give this qualitative distinction in output.

Several implications relate to users wishing to construct climate change projections using metrics either as a guide for RCM elimination or as weighting. Performance metrics used as indicators will need to be assessed twice, both for their robustness and independence in historical simulations (Chapter 4), but also in their stationarity of the model errors in question (Chapter 6.4), which underpins the reliability of any weighting scheme. Weighted projections in regions of high bias-nonstationarity such as the Alps (see Chapter 6.4) should be used with care. However, it should be noted that weighting schemes using GPIs with a high number of metrics may not change the projected output overly (Chapter 5.3), but for a narrow range of variables (with a small number of metrics) GPI values can change substantially (Chapter 5.4 and 5.5). Metric combinations in the form of GPIs are most likely appropriate for users who require quick sources of information on model quality, in particular intra-ensemble performance over consecutive generations. However, application of GPIs may yield little additional value for ensemble

weighting applications, as the difference compared to a simple multi-model mean is negligible, at least for this ensemble and variables considered. More specialised GPIs on the other hand, considering extremes or certain physical parameters such as soil-moisture/surface albedo, may be suitable for particular purposes such as for the impacts/vulnerability community where numerous factors may be considered in one quantitative assessment. Finally, the impacts and vulnerability community might need more tailored metrics, reflecting the context of their work, for example extreme precipitation metrics for flood modelling, soil moisture and WSDI (warm spell duration index) for heatwaves. The Chapter 4 analysis suggests that for such work some extreme indices may be redundant in certain regions, and thus metric independence tests should be undertaken, such as those given in Chapter 4.5.

Clearly, the robustness of performance metrics is essential for all of these applications. If metrics are liable to change in their quantitative and qualitative assessment of RCM performance when small changes in experimental setup are introduced then one cannot rely on metrics as a reliable objective arbiter of model quality. Although RCM evaluation studies may use metrics, if a more process-based approach alongside a quantitative performance metric approach is taken, uncertainties in whether metrics are robust may be of less significance, since the underlying physical cause of RCM error will be diagnosed. However, those who do not conduct such extensive evaluation, but use metrics alone are more reliant on a robust, objective sets of metrics. This can be seen in the assessments of delta-t future temperature changes in the Alps in Chapter 6.4, as potential mechanisms such as surface-albedo feedback will not be considered by a straightforward quantification of the regional temperature change, or of mean historical temperature bias. One implication of this is that more in depth process-based analysis is essential to fully understand bias causes and changes, and metrics designed to assess key physical components are potentially the way forward.

The stationarity assumption naturally is not particularly relevant for the modellers or ensemble analysts in the historical context, but for those interested in applying quantitative metric information in future projections this should be taken into account. There remains however some level of uncertainty as to the best method with which to test this assumption, and whether in fact the relationship of RCM to RCM is analogous to RCM to reality (Whetton *et al.*, 2007). This concept of bias stationarity is essential however if one wishes to infer meaning to metric assessments beyond a simple scalar score, relevant only to a single time-period. The

fact that future model reliability must be inferred from historical simulation quality is both an essential component and yet unverifiable gives further urgency to the task of finding better and more coherent tests of bias stationarity. This may provide one avenue with which to better characterise ensemble projection uncertainty, as regions of high non-stationarity may give overconfident results.

7.4 Recommendations and Outlook

The first aim of the thesis laid out in Chapter 1.4 emphasises the need for objectivity and robustness in the task of evaluating RCMs with performance metrics and applying that information thereafter. The second aim is to develop criteria and analysis methods more likely to provide robust outcomes. The analysis undertaken in the thesis has required the adoption and development of new novel analysis methods with which to test the underlying assumptions of model evaluation, GPIs and bias stationarity. These can therefore by utilised by others wishing to carry out sensitivity analyses for their own purposes. There are three aspects which are essential to the selection of a set of performance metrics:

- **objectivity**
- **robustness**
- **redundancy**

This process is essentially one that begins with the question: what variable or variables are to be evaluated? This could specify types of process, standard variables (temperature/precipitation) or a more general overview of model skill. Thus to meet the **objectivity criterion**, metrics used would have to span a range of statistics assessing the four components identified in Chapter 4.1: standard error measures, spatial patterns, temporal and frequency. Although it was not done in this thesis, the use of more than one observational dataset is also recommended.

Next, to meet the **robustness criterion**, statistics that have been shown to be so in Chapter 4.4 are recommended for use; that is a variety of temporal statistics (interannual variability/annual cycle/diurnal temperature range) single standard error, spatial pattern type statistics and frequency metrics (PDF and CDF methods). The spatial robustness of metric output should be analysed with sub-domain

assessments (nationally or for regions of homogeneous climatic character where appropriate) to ensure that metrics are not overly sensitive to the potentially arbitrary choice of domain.

Finally, to meet the **redundancy criterion**, analysis should be undertaken to assess the output of metrics against each other to leave a final set of metrics of independent information.

The results of analysis investigating the use of GPIs to form overall indicators of RCM performance similarly involved the development and use of several new methods with which to test objectivity and robustness for both the combination method, and number/type of variable included. The recommendations for use of GPIs for the first aspect are that any combination method is likely robust in its output, and therefore any of those in Chapter 5.3 are suggested as appropriate. For any new method it is suggested that a brief sensitivity analysis is undertaken in line with the methods used in Chapter 5 to ascertain whether this method is generally robust with respect to other methods. For the number and type of variable included in a GPI it is recommended that as wide a range of variables is included as possible. This is to ensure that no relevant performance information is lost. Chapter 5.4 on the number of metrics to use yields uncertain results, and as such no recommendations are provided here beyond the suggestion that at least 10-20 are utilised given that fewer than this will not provide a sufficient variable number required under the previous recommendation.

The final analysis Chapter 6 exploring the stationarity assumption provides an alternative approach in the percentile bias change method to the mean bias change method of Maraun (2012) to assess future projection bias stationarity, although it remains unclear which can be considered superior. However, for the specific question of quantile mapping bias-correction methods, since the quantile-bias change method is in part based on this distributional approach, this new method is recommended as an appropriate measure of whether the assumption is plausible or not. For those wishing to weight future climate change projections it is recommended that after a set of objective, robust, and independent combined metrics are identified as per Chapter 4 and 5, that the underlying variable biases are assessed for non-stationarity as this may influence results greatly, depending on the region in question.

Temperature metrics using extremes (Tn10p/Tx90p) should be aware that in the quantile-bias change paradigm there may not be an answer to ask whether or how much RCM biases in these variables are stationary, given that these historical temperatures either do not occur in the future, or the future Tx90p does not occur in the historical simulations. It is therefore suggested that analysis is undertaken to ascertain to what degree this is true for all gridpoint under consideration using methods described in Chapter 6.2.

The more considered use of performance metrics towards more objective approach to RCM evaluation is a constructive trend for both practical and scientific purposes. Practical because performance metrics provide a computationally inexpensive way to gather a large amount of information on model quality into a concise format. Scientific because by removing potential areas of subjectivity the field may move towards a more systematic framework for assessing RCMs, as in numerical weather prediction, and potentially improving results as shown in that discipline. The specific topics investigated in this thesis were identified in order to further the understanding of how performance metrics can be developed and used to provide a more robust foundation with which to analyse and use RCM simulations.

Many of the recommendations from the thesis are quite specific, some general recommendations relating to good practice can be made. Overall, when evaluating RCMs, it is better to assess as wide a range of aspects as possible; process based assessments, although not directly the focus of this thesis, certainly offer a greater understanding than simpler quantitative evaluations, yet may not be undertaken for every variable and season. As such a number of statistics considering both temporal and spatial aspects are recommended, in addition to a wide range of variables including those spanning the distribution range. For more targeted analysis, such as for a specific impacts task, a narrow range of variables may be of more utility, yet the wider context should still be considered; is bias stationarity important for this region/season/variable? Is my ensemble of models overconfident in its projections? Those requiring 'quick-look' information may use metric combination methods, and most averaging methods will be appropriate for use. It should be noted that the choice of variable is the predominant factor in the final output, and as such variables should be chosen with care.

To further examine these issues, certain ensemble designs would be highly beneficial. This is most clearly the case for assessment of the stationarity assumption,

where the main requirement is several RCMs forced by the same GCM. Large ensembles would provide a much clearer understanding both of how RCM/GCM biases interact and also the degree to which these biases change over time. Since the only other alternative is to wait for validation data, this seems to be the clear best course of action. Alternatively, exploitation of the availability of several future RCP radiative forcing pathways may be another avenue for testing the assumption, as one would be able to test how the stationarity of RCM errors is influenced by using one or another RCP. Furthering the suggestions of Bellprat *et al.* (2013) in developing bias correction methods linked directly to the underlying causes of RCM bias, such as soil-moisture-precipitation feedbacks or surface-albedo/snow depth, is suggested as one potential avenue for exploration. For further investigation of metric sensitivity, it is suggested that a wider range of variables including physical properties, such as incoming solar radiation, are considered. This may provide, subject to observational availability, a clearer understanding of the causes of RCM bias, which would be of benefit to all modelling groups.

References

Abe, M., H. Shiogama, J. Hargreaves, J. Annan, T. Nozawa, and S. Emori (2009), Correlation between inter-model similarities in spatial pattern for present and projected future mean climate, *Sola*, *5*(0), 133–136.

Abramowitz, G., and H. Gupta (2008), Toward a model space and model independence metric, *Geophysical Research Letters*, *35*(5), L05,705.

Andrews, T., J. M. Gregory, M. J. Webb, and K. E. Taylor (2012), Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models, *Geophysical Research Letters*, *39*(9), L09,712.

Annan, J., and J. Hargreaves (2011), Understanding the CMIP3 multimodel ensemble, *Journal of Climate*, *24*(16), 4529–4538.

Baklanov, A. A., B. Grisogono, R. Bornstein, L. Mahrt, S. S. Zilitinkevich, P. Taylor, S. E. Larsen, M. W. Rotach, and H. Fernando (2011), The nature, theory, and modeling of atmospheric planetary boundary layers, *Bulletin of the American Meteorological Society*, *92*(2), 123–128.

Bao, X., and F. Zhang (2013), Evaluation of NCEP–CFSR, NCEP–NCAR, ERA-Interim, and ERA-40 reanalysis datasets against independent sounding observations over the Tibetan Plateau, *Journal of Climate*, *26*(1), 206–214.

Barnett, J., and W. N. Adger (2007), Climate change, human security and violent conflict, *Political Geography*, *26*(6), 639–655.

Bellard, C., C. Bertelsmeier, P. Leadley, W. Thuiller, and F. Courchamp (2012), Impacts of climate change on the future of biodiversity, *Ecology letters*, *15*(4), 365–377.

Bellenger, H., E. Guilyardi, J. Leloup, M. Lengaigne, and J. Vialard (2014), ENSO representation in climate models: from CMIP3 to CMIP5, *Climate Dynamics*, *42*(7-8), 1999–2018.

Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär (2013), Physical constraints for temperature biases in climate models, *Geophysical Research Letters*, *40*(15), 4042–4047.

Berg, A., B. R. Lintner, K. L. Findell, S. Malyshev, P. C. Loikith, and P. Gentine (2014), Impact of soil moisture–atmosphere interactions on surface temperature distribution, *Journal of Climate*, *27*(21), 7976–7993.

Berg, P., J. Haerter, P. Thejll, C. Piani, S. Hagemann, and J. Christensen (2009), Seasonal characteristics of the relationship between daily precipitation intensity and surface temperature, *Journal of Geophysical Research: Atmospheres*, *114*(D18102).

Bindoff, N., P. Stott, K. AchutaRao, M. Allen, N. Gillett, D. Gutzler, K. Hansingo, G. Hegerl, Y. Hu, S. Jain, I. Mokhov, J. Overland, J. Perlwitz, R. Sebbari, and X. Zhang (2013), 2013: Detection and Attribution of Climate Change: from Global to Regional. 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.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Boberg, F., P. Berg, P. Thejll, W. J. Gutowski, and J. H. Christensen (2009), Improved confidence in climate change projections of precipitation evaluated using daily statistics from the PRUDENCE ensemble, *Climate Dynamics*, *32*(7-8), 1097–1106.

Boone, A., B. Decharme, F. Guichard, P. de Rosnay, G. Balsamo, A. Beljaars, F. Chopin, T. Orgeval, J. Polcher, C. Delire, *et al.* (2009), The AMMA land surface model intercomparison project (ALMIP), *Bulletin of the American Meteorological Society*, *90*(12), 1865–1880.

Boulard, D., B. Pohl, J. Crétat, N. Vigaud, and T. Pham-Xuan (2013), Downscaling large-scale climate variability using a regional climate model: the case of ENSO over Southern Africa, *Climate Dynamics*, *40*(5-6), 1141–1168.

Buizza, R. (2008), The value of probabilistic prediction, *Atmospheric Science Letters*, *9*(2), 36–42.

Burton, A., H. Fowler, S. Blenkinsop, and C. Kilsby (2010), Downscaling transient climate change using a Neyman–Scott Rectangular Pulses stochastic rainfall model, *Journal of Hydrology*, *381*(1), 18–32.

Buser, C., H. Künsch, D. Lüthi, M. Wild, and C. Schär (2009), Bayesian multimodel projection of climate: bias assumptions and interannual variability, *Climate Dynamics*, *33*(6), 849–868.

Casanueva, A., S. Herrera, J. Fernández, M. Frías, and J. Gutiérrez (2013), Evaluation and projection of daily temperature percentiles from statistical and dynamical downscaling methods, *Natural Hazards and Earth System Science*, *13*(8), 2089–2099.

Chen, J., F. P. Brissette, and P. Lucas-Picher (2015), Assessing the limits of bias-correcting climate model outputs for climate change impact studies, *Journal of Geophysical Research: Atmospheres*, *120*(3), 1123–1136.

Christensen, J., F. Boberg, O. Christensen, and P. Lucas-Picher (2008), On the need for bias correction of regional climate change projections of temperature and precipitation, *Geophysical Research Letters*, *35*(20), L20,709.

Christensen, J., T. Carter, and F. Giorgi (2002), PRUDENCE employs new methods to assess European climate change, *EOS, Transactions American Geophysical Union*, *83*(13), 147.

Christensen, J., E. Kjellström, F. Giorgi, G. Lenderink, and M. Rummukainen (2010), Weight assignment in regional climate models, *Climate Research*, *44*(2), 179.

Christensen, J. H., and F. Boberg (2012), Temperature dependent climate projection deficiencies in CMIP5 models, *Geophysical Research Letters*, *39*(24), L24,705.

Collins, M. (2002), Climate predictability on interannual to decadal time scales: the initial value problem, *Climate Dynamics*, *19*(8), 671–692.

Collins, M., R. Knutti, J. Arblaster, J.-L. Dufresne, T. Fichefet, P. Friedlingstein, X. Gao, W. Gutowski, T. Johns, G. Krinner, M. Shongwe, C. Tebaldi, A. Weaver, and M. Wehner (2013), Long-term Climate Change: Projections, Commitments and Irreversibility, 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.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Collins, W., N. Bellouin, M. Doutriaux-Boucher, N. Gedney, T. Hinton, C. Jones, S. Liddicoat, G. Martin, F. O'Connor, J. Rae, *et al.* (2008), Evaluation of the HadGEM2 model, *Hadley Cent. Tech. Note*, *74*.

Coppola, E., F. Giorgi, S. Rauscher, and C. Piani (2010), Model weighting based on mesoscale structures in precipitation and temperature in an ensemble of regional climate models, *Climate Research*, *44*(2), 121.

Daron, J. D., and D. A. Stainforth (2013), On predicting climate under climate change, *Environmental Research Letters*, *8*(3), 034,021.

Déqué, M. (2007), Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values, *Global and Planetary Change*, *57*(1), 16–26.

Déry, S. J., and E. F. Wood (2005), Observed twentieth century land surface air temperature and precipitation covariability, *Geophysical Research Letters*, *32*(21), L21,414.

Dessai, S., and M. Hulme (2004), Does climate adaptation policy need probabilities?, *Climate Policy*, *4*(2), 107–128.

Doblas-Reyes, F., R. Hagedorn, and T. Palmer (2005), The rationale behind the success of multi-model ensembles in seasonal forecasting–II. Calibration and combination, *Tellus A*, *57*(3), 234–252.

Donat, M. G., G. C. Leckebusch, S. Wild, and U. Ulbrich (2010), Benefits and limitations of regional multi-model ensembles for storm loss estimations, *Climate Research*, *44*(2), 211.

Doney, S. C., V. J. Fabry, R. A. Feely, and J. A. Kleypas (2009), Ocean acidification: the other CO2 problem, *Marine Science*, *1*, 169– 192.

Dosio, A., P. Paruolo, and R. Rojas (2012), Bias correction of the ENSEMBLES high resolution climate change projections for use by impact models: Analysis of the climate change signal, *Journal of Geophysical Research: Atmospheres*, *117*(D17110).

Ehret, U., E. Zehe, V. Wulfmeyer, K. Warrach-Sagi, and J. Liebert (2012), HESS Opinions "Should we apply bias correction to global and regional climate model data?", *Hydrology and Earth System Sciences Discussions*, *9*(4), 5355–5387.

Eum, H.-I., P. Gachon, R. Laprise, and T. Ouarda (2012), Evaluation of regional climate model simulations versus gridded observed and regional reanalysis products using a combined weighting scheme, *Climate dynamics*, *38*(7-8), 1433–1457.

Evans, J., F. Ji, C. Lee, P. Smith, D. Argüeso, and L. Fita (2014), Design of a regional climate modelling projection ensemble experiment–NARCliM, *Geoscientific Model Development*, *7*(2), 621–629.

Field, C., V. Barros, K. Mach, M. Mastrandrea, M. van Aalst, W. Adger, D. Arent, J. Barnett, R. Betts, T. Bilir, J. Birkmann, J. Carmin, D. Chadee, A. Challinor, M. Chatterjee, W. Cramer, D. Davidson, Y. Estrada, J.-P. Gattuso, Y. Hijioka, O. Hoegh-Guldberg, H.-Q. Huang, G. Insarov, R. Jones, R. Kovats, P. R. Lankao, J. Larsen, I. Losada, J. Marengo, R. McLean, L. Mearns, R. Mechler, J. Morton, I. Niang, T. Oki, J. Olwoch, M. Opondo, E. Poloczanska, H.-O. Pörtner, M. Redsteer, A. Reisinger, A. Revi, D. Schmidt, M. Shaw, W. Solecki, D. Stone, J. Stone, K. Strzepek, A. Suarez, P. Tschakert, R. Valentini, S. Vicuña, A. Villamizar, K. Vincent, R. Warren, L. White, T. Wilbanks, P. Wong, and G. Yohe (2014), Technical Summary. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects, Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L. White (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. XXX-YYY.

Fischer, E. M., S. Seneviratne, P. Vidale, D. Lüthi, and C. Schär (2007), Soil moisture-atmosphere interactions during the 2003 European summer heat wave, *Journal of Climate*, *20*(20), 5081–5099.

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S. 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 (2013), Evaluation of Climate Models, 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.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Fowler, H., and M. Ekström (2009), Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes, *International Journal of Climatology*, *29*(3), 385–416.

Frame, D., N. Faull, M. Joshi, and M. Allen (2007), Probabilistic climate forecasts and inductive problems, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 1971–1992.

Furrer, R., S. R. Sain, D. Nychka, and G. A. Meehl (2007), Multivariate Bayesian analysis of atmosphere–ocean general circulation models, *Environmental and Ecological Statistics*, *14*(3), 249–266.

Gates, W. L. (1992), AMIP: The atmospheric model intercomparison project, *Bulletin of the American Meteorological Society*, *73*(12), 1962–1970.

Giorgi, F., E. Coppola, F. Solmon, L. Mariotti, M. Sylla, X. Bi, N. Elguindi, G. Diro, V. Nair, G. Giuliani, *et al.* (2012), RegCM4: model description and preliminary tests over multiple CORDEX domains, *Climate Research*, *2*(52), 7–29.

Giorgi, F., C. Jones, G. Asrar, *et al.* (2009), Addressing climate information needs at the regional level: the CORDEX framework, *World Meteorological Organization (WMO) Bulletin*, *58*(3), 175.

Giorgi, F., and L. Mearns (2002), 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.

Gleckler, P., K. Taylor, and C. Doutriaux (2008), Performance metrics for climate models, *Journal of Geophysical Research*, *113*, D06,104.

Goddard, L., A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, V. Kharin, W. Merryfield, C. Deser, S. J. Mason, *et al.* (2013), A verification framework for interannual-to-decadal predictions experiments, *Climate Dynamics*, *40*(1-2), 245–272.

Gómez-Navarro, J., J. Montávez, S. Jerez, P. Jiménez-Guerrero, and E. Zorita (2012), What is the role of the observational dataset in the evaluation and scoring of climate models?, *Geophysical Research Letters*, *39*(24).

Guilyardi, E., V. Balaji, B. Lawrence, S. Callaghan, C. Deluca, S. Denvil, M. Lautenschlager, M. Morgan, S. Murphy, and K. E. Taylor (2013), Documenting climate models and their simulations, *Bulletin of the American Meteorological Society*, *94*(5), 623–627.

Gutiérrez, J. M., D. San-Martín, S. Brands, R. Manzanas, and S. Herrera (2013), Reassessing statistical downscaling techniques for their robust application under climate change conditions, *Journal of Climate*, *26*(1), 171–188.

Hagedorn, R., F. Doblas-Reyes, and T. Palmer (2005), The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept, *Tellus A*, *57*(3), 219–233.

Hanjra, M. A., and M. E. Qureshi (2010), Global water crisis and future food security in an era of climate change, *Food Policy*, *35*(5), 365–377.

Hansen, J. (2006), Can we still avoid dangerous human-made climate change?, *Social Research: An International Quarterly*, *73*(3), 949–974.

Hay, L., and M. Clark (2003), Use of statistically and dynamically downscaled atmospheric model output for hydrologic simulations in three mountainous basins in the western United States, *Journal of Hydrology*, *282*(1), 56–75.

Hayhoe, K., K. Dixon, A. Stoner, J. Lanzante, and A. Radhakrishnan (2012), Is the past a guide to the future? Evaluating the assumption of climate stationarity in statistical downscaling, in: *AGU Fall Meeting Abstracts*, vol. 1, p. 04.

Haylock, M., N. Hofstra, A. Klein Tank, E. Klok, P. Jones, and M. New (2008), A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *113*(D20).

Heinrich, G., and A. Gobiet (2012), The future of dry and wet spells in Europe: A comprehensive study based on the ENSEMBLES regional climate models, *International Journal of Climatology*, *32*(13), 1951–1970.

Herrera, S., J. M. Gutiérrez, R. Ancell, M. Pons, M. Frías, and J. Fernández (2012), Development and analysis of a 50-year high-resolution daily gridded precipitation dataset over Spain (Spain02), *International Journal of Climatology*, *32*(1), 74–85.

Hofstra, N., M. New, and C. McSweeney (2010), The influence of interpolation and station network density on the distributions and trends of climate variables in gridded daily data, *Climate Dynamics*, *35*(5), 841–858.

Holtanová, E., J. Mikšovskỳ, J. Kalvová, P. Pišoft, and M. Motl (2012), Performance of ENSEMBLES regional climate models over Central Europe using various metrics, *Theoretical and Applied Climatology*, *108*(3-4), 463–470.

Hourdin, F., J.-Y. Grandpeix, C. Rio, S. Bony, A. Jam, F. Cheruy, N. Rochetin, L. Fairhead, A. Idelkadi, I. Musat, *et al.* (2013), LMDZ5B: the atmospheric component of the IPSL climate model with revisited parameterizations for clouds and convection, *Climate Dynamics*, *40*(9-10), 2193–2222.

Huffman, G. J., D. T. Bolvin, E. J. Nelkin, D. B. Wolff, R. F. Adler, G. Gu, Y. Hong, K. P. Bowman, and E. F. Stocker (2007), The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales, *Journal of Hydrometeorology*, *8*(1), 38–55.

Jacob, D., J. Petersen, B. Eggert, A. Alias, O. B. Christensen, L. M. Bouwer, A. Braun, A. Colette, M. Déqué, G. Georgievski, *et al.* (2014), EURO-CORDEX: new high-resolution climate change projections for European impact research, *Regional Environmental Change*, *14*(2), 563–578.

Jaeger, E., I. Anders, D. Luthi, B. Rockel, C. Schar, and S. Seneviratne (2008), Analysis of ERA40-driven CLM simulations for Europe, *Meteorologische Zeitschrift*, *17*(4), 349–367.

Jaeger, E., and S. Seneviratne (2011), Impact of soil moisture–atmosphere coupling on European climate extremes and trends in a regional climate model, *Climate Dynamics*, *36*(9-10), 1919–1939.

Jakob, C. (2010), Accelerating progress in global atmospheric model development through improved parameterizations: Challenges, opportunities, and strategies, *Bulletin of the American Meteorological Society*, *91*(7), 869–875.

Johnson, F., and A. Sharma (2012), A nesting model for bias correction of variability at multiple time scales in general circulation model precipitation simulations, *Water Resources Research*, *48*(1), W01,504.

Joshi, M. M., J. M. Gregory, M. J. Webb, D. M. Sexton, and T. C. Johns (2008), Mechanisms for the land/sea warming contrast exhibited by simulations of climate change, *Climate Dynamics*, *30*(5), 455–465.

Kallberg, P., A. Simmons, S. Uppala, and M. Fuentes (2004), The ERA-40 archive, ERA-40 Project Report Series, No. 17, ECMWF, *Reading: UK*.

Kang, E. L., N. Cressie, and S. R. Sain (2012), Combining outputs from the North American regional climate change assessment program by using a Bayesian hierarchical model, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *61*(2), 291–313.

Kendon, E. J., N. M. Roberts, C. A. Senior, and M. J. Roberts (2012), Realism of rainfall in a very high-resolution regional climate model, *Journal of Climate*, *25*(17), 5791–5806.

Kerkhoff, C., H. R. Künsch, and C. Schär (2014), Assessment of bias assumptions for climate models, *Journal of Climate*, *27*(17), 6799–6818.

Kim, D., K. Sperber, W. Stern, D. Waliser, I.-S. Kang, E. Maloney, W. Wang, K. Weickmann, J. Benedict, M. Khairoutdinov, *et al.* (2009), Application of MJO simulation diagnostics to climate models, *Journal of Climate*, *22*(23), 6413–6436.

Kim, J., D. E. Waliser, C. A. Mattmann, C. E. Goodale, A. F. Hart, P. A. Zimdars, D. J. Crichton, C. Jones, G. Nikulin, B. Hewitson, *et al.* (2014), Evaluation of the CORDEX-Africa multi-RCM hindcast: systematic model errors, *Climate Dynamics*, *42*(5-6), 1189–1202.

Kirtman, B., S. Power, J. Adedoyin, G. Boer, R. Bojariu, I. Camilloni, F. Doblas-Reyes, A. Fiore, M. Kimoto, G. Meehl, M. Prather, A. Sarr, C. Schär, R. Sutton, G. van Oldenborgh, G. Vecchi, and H. Wang (2013), Near-term Climate Change: Projections and Predictability, 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.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Kjellström, E. (2005), *A 140-year simulation of European climate with the new version of the Rossby Centre regional atmospheric climate model (RCA3)*, SMHI.

Kjellström, E., F. Boberg, M. Castro, J. Christensen, G. Nikulin, and E. Sánchez (2010), Daily and monthly temperature and precipitation statistics as performance indicators for regional climate models, *Climate Research*, *44*(2), 135.

Klein, S. A., Y. Zhang, M. D. Zelinka, R. Pincus, J. Boyle, and P. J. Gleckler (2013), Are climate model simulations of clouds improving? An evaluation using the ISCCP simulator, *Journal of Geophysical Research: Atmospheres*, *118*(3), 1329–1342.

Knutti, R. (2008), 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.

Knutti, R. (2010), The end of model democracy?, *Climatic C*, *102*(3), 395–404.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. Meehl (2010), Challenges in combining projections from multiple climate models, *Journal of Climate*, *23*(10), 2739–2758.

Knutti, R., and G. Hegerl (2008), The equilibrium sensitivity of the Earth's temperature to radiation changes, *Nature Geoscience*, *1*(11), 735–743.

Knutti, R., and J. Sedláček (2013), Robustness and uncertainties in the new CMIP5 climate model projections, *Nature Climate Change*, *3*(4), 369–373.

Køltzow, M. A., T. Iversen, and J. E. Haugen (2011), The Importance of Lateral Boundaries, Surface Forcing and Choice of Domain Size for Dynamical Downscaling of Global Climate Simulations, *Atmosphere*, *2*(2), 67–95.

Kotlarski, S., K. Keuler, O. B. Christensen, A. Colette, M. Déqué, A. Gobiet, K. Goergen, D. Jacob, D. Lüthi, E. van Meijgaard, *et al.* (2014), Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble, *Geoscientific Model Development Discussions*, *7*(1), 217–293.

Kyselỳ, J., and E. Plavcová (2010), A critical remark on the applicability of E-OBS European gridded temperature data set for validating control climate simulations, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *115*(D23).

Lambert, S., and G. Boer (2001), CMIP1 evaluation and intercomparison of coupled climate models, *Climate Dynamics*, *17*(2), 83–106.

Laprise, R. (2008), Regional climate modelling, *Journal of Computational Physics*, *227*(7), 3641–3666.

Lavers, D., C. Prudhomme, and D. M. Hannah (2013), European precipitation connections with large-scale mean sea-level pressure (MSLP) fields, *Hydrological Sciences Journal*, *58*(2), 310–327.

Lee, D. D., and H. S. Seung (1999), Learning the parts of objects by non-negative matrix factorization, *Nature*, *401*(6755), 788–791.

Legates, D., and G. McCabe Jr (1999), Evaluating the use of" goodness-of-fit" measures in hydrologic and hydroclimatic model validation, *Water Resources Research*, *35*(1), 233–241.

Lenderink, G. (2010), Exploring metrics of extreme daily precipitation in a large ensemble of regional climate model simulations, *Climate Research*, *44*(2), 151.

Liang, X.-Z., K. E. Kunkel, and A. N. Samel (2001), Development of a regional climate model for US Midwest applications. Part I: Sensitivity to buffer zone treatment, *Journal of Climate*, *14*(23), 4363–4378.

Lim, E.-P., H. H. Hendon, D. L. Anderson, A. Charles, and O. Alves (2011), Dynamical, statistical-dynamical, and multimodel ensemble forecasts of Australian spring season rainfall, *Monthly Weather Review*, *139*(3), 958–975.

Lorenz, P., and D. Jacob (2010), Validation of temperature trends in the ENSEMBLES regional climate model runs driven by ERA40, *Climate Research*, *44*(2), 167.

Ma, H.-Y., S. Xie, J. Boyle, S. Klein, and Y. Zhang (2013), Metrics and diagnostics for precipitation-related processes in climate model short-range hindcasts, *Journal of Climate*, *26*(5), 1516–1534.

Magnusson, L., and E. Källén (2013), Factors influencing skill improvements in the ECMWF forecasting system, *Monthly Weather Review*, *141*(9), 3142–3153.

Mahlstein, I., and R. Knutti (2012), September Arctic sea ice predicted to disappear near 2 C global warming above present, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *117*(D6).

Maraun, D. (2012), Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums, *Geophysical Research Letters*, *39*(6), L06,706.

Marzban, C. (2003), A comment on the ROC curve and the area under it as performance measures, *Weather Forecasting*.

Mauritsen, T., B. Stevens, E. Roeckner, T. Crueger, M. Esch, M. Giorgetta, H. Haak, J. Jungclaus, D. Klocke, D. Matei, *et al.* (2012), Tuning the climate of a global model, *Journal of Advances in Modeling Earth Systems*, *4*(3), M00A01.

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

Meehl, G. A., G. J. Boer, C. Covey, M. Latif, and R. J. Stouffer (2000), The coupled model intercomparison project (CMIP), *Bulletin of the American Meteorological Society*, *81*(2), 313–318.

Moise, A., and F. Delage (2011), New climate model metrics based on object-orientated pattern matching of rainfall, *Journal of Geophysical Research*, *116*(D12), D12,108.

Mooney, P., F. Mulligan, and R. Fealy (2011), Comparison of ERA-40, ERA-Interim and NCEP/NCAR reanalysis data with observed surface air temperatures over Ireland, *International Journal of Climatology*, *31*(4), 545–557.

Moss, R., J. Edmonds, K. Hibbard, M. Manning, S. Rose, D. van Vuuren, T. Carter, S. Emori, M. Kainuma, T. Kram, *et al.* (2010), The next generation of scenarios for climate change research and assessment, *Nature*, *463*(7282), 747–756.

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

Nakicenovic, N., J. Alcamo, G. Davis, B. de Vries, J. Fenhann, S. Gaffin, K. Gregory, A. Grubler, T. Y. Jung, T. Kram, *et al.* (2000), *Special report on emissions scenarios: a special report of Working Group III of the Intergovernmental Panel on Climate Change*, Tech. rep., Pacific Northwest National Laboratory, Richland, WA (US), Environmental Molecular Sciences Laboratory (US).

Nicholls, R. J., and A. Cazenave (2010), Sea-level rise and its impact on coastal zones, *Science*, *328*(5985), 1517–1520.

Nikulin, G., C. Jones, F. Giorgi, G. Asrar, M. Büchner, R. Cerezo-Mota, O. B. Christensen, M. Déqué, J. Fernandez, A. Hänsler, *et al.* (2012), Precipitation climatology in an ensemble of CORDEX-Africa regional climate simulations, *Journal of Climate*, *25*(18), 6057–6078.

Nishii, K., T. Miyasaka, H. Nakamura, Y. Kosaka, S. Yokoi, Y. Takayabu, H. Endo, H. Ichikawa, T. Inoue, K. Oshima, *et al.* (2012), Relationship of the reproducibility of multiple variables among global climate models, *Journal of the Meteorological Society of Japan*, *90*(0), 87–100.

Norris, J. R., and M. Wild (2007), Trends in aerosol radiative effects over Europe inferred from observed cloud cover, solar "dimming," and solar "brightening", *Journal of Geophysical Research: Atmospheres (1984–2012)*, *112*(D8).

Oreskes, N., K. Shrader-Frechette, K. Belitz, *et al.* (1994), Verification, validation, and confirmation of numerical models in the earth sciences, *Science*, *263*(5147), 641–646.

Parker, W. (2011), When Climate Models Agree: The Significance of Robust Model Predictions, *Philosophy of Science*, *78*(4), 579–600.

Patz, J. A., D. Campbell-Lendrum, T. Holloway, and J. A. Foley (2005), Impact of regional climate change on human health, *Nature*, *438*(7066), 310–317.

Perkins, S., A. Pitman, N. Holbrook, and J. McAneney (2007), Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions, *Journal of Climate*, *20*(17), 4356–4376.

Phillips, T., G. Potter, D. Williamson, R. Cederwall, J. Boyle, M. Fiorino, J. Hnilo, J. Olson, S. Xie, and J. Yio (2004), Evaluating parameterizations in general circulation models: Climate simulation meets weather prediction, *Bulletin of the American Meteorological Society*, *85*(12), 1903–1915.

Piani, C., J. Haerter, and E. Coppola (2010), Statistical bias correction for daily precipitation in regional climate models over Europe, *Theoretical and Applied Climatology*, *99*(1-2), 187–192.

Pierce, D., T. Barnett, B. Santer, and P. Gleckler (2009), Selecting global climate models for regional climate change studies, *Proceedings of the National Academy of Sciences*, *106*(21), 8441.

Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Glecker (2008), Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, *Journal of Geophysical Research*, *113*, D14,209.

Pocock, J., S. Dance, A. Lawless, N. Nichols, and J. Eyre (2012), Representativity error for temperature and humidity using the Met Office high resolution model, *Mathematics Preprint Series MPS-2012*, *18*.

Power, S. B., F. Delage, R. Colman, and A. Moise (2012), Consensus on twenty-first-century rainfall projections in climate models more widespread than previously thought, *Journal of Climate*, *25*(11), 3792–3809.

Prömmel, K., B. Geyer, J. Jones, and M. Widmann (2010), Evaluation of the skill and added value of a reanalysis-driven regional simulation for Alpine temperature, *International Journal of Climatology*, *30*(5), 760–773.

Radic, V., and G. K. Clarke (2011), Evaluation of IPCC Models' Performance in Simulating Late-Twentieth-Century Climatologies and Weather Patterns over North America, *Journal of Climate*, *24*(20), 5257–5274.

Räisänen, J. (1997), Objective comparison of patterns of CO2 induced climate change in coupled GCM experiments, *Climate Dynamics*, *13*(3), 197–211.

Räisänen, J., and J. Ylhäisi (2011), Can model weighting improve probabilistic projections of climate change?, *Climate Dynamics*, pp. 1–18.

Reichler, T., and J. Kim (2008), How Well Do Coupled Models Simulate Today's Climate?, *Bulletin of the American Meteorological Society*, *89*, 303.

Reifen, C., and R. Toumi (2009), Climate projections: Past performance no guarantee of future skill?, *Geophysical Research Letters*, *36*(13), L13,704.

Rockel, B., and K. Woth (2007), Extremes of near-surface wind speed over Europe and their future changes as estimated from an ensemble of RCM simulations, *Climatic Change*, *81*(1), 267–280.

Rosenzweig, C., D. Karoly, M. Vicarelli, P. Neofotis, Q. Wu, G. Casassa, A. Menzel, T. Root, N. Estrella, B. Seguin, *et al.* (2008), Attributing physical and biological impacts to anthropogenic climate change, *Nature*, *453*(7193), 353–357.

Rosolem, R., H. V. Gupta, W. J. Shuttleworth, L. G. G. Gonçalves, and X. Zeng (2013), Towards a comprehensive approach to parameter estimation in land surface parameterization schemes, *Hydrological Processes*, *27*(14), 2075–2097.

Ruckstuhl, C., R. Philipona, K. Behrens, M. Collaud Coen, B. Dürr, A. Heimo, C. Mätzler, S. Nyeki, A. Ohmura, L. Vuilleumier, *et al.* (2008), Aerosol and cloud effects on solar brightening and the recent rapid warming, *Geophysical Research Letters*, *35*(12), D08,214.

Sánchez, E., R. Romera, M. Gaertner, C. Gallardo, and M. Castro (2009), A weighting proposal for an ensemble of regional climate models over Europe driven by 1961–2000 ERA40 based on monthly precipitation probability density functions, *Atmospheric Science Letters*, *10*(4), 241–248.

Sanderson, B. M. (2011), A multimodel study of parametric uncertainty in predictions of climate response to rising greenhouse gas concentrations, *Journal of Climate*, *24*(5), 1362–1377.

Sargent, R. G. (1998), Verification and validation of simulation models, in: *Proceedings of the 30th conference on Winter simulation*, pp. 121–130, IEEE Computer Society Press.

Schwalm, C. R., D. N. Huntinzger, A. M. Michalak, J. B. Fisher, J. S. Kimball, B. Mueller, K. Zhang, and Y. Zhang (2013), Sensitivity of inferred climate model skill to evaluation decisions: a case study using CMIP5 evapotranspiration, *Environmental Research Letters*, *8*(2), 024,028.

Separovic, L., R. de Elía, and R. Laprise (2012), Impact of spectral nudging and domain size in studies of RCM response to parameter modification, *Climate dynamics*, *38*(7-8), 1325–1343.

Sharma, D., A. Das Gupta, and M. Babel (2007), Spatial disaggregation of bias-corrected GCM precipitation for improved hydrologic simulation: Ping River Basin, Thailand, *Hydrology and Earth System Sciences*, *11*(4), 1373–1390.

Sillmann, J., V. Kharin, X. Zhang, F. Zwiers, and D. Bronaugh (2013), Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate, *Journal of Geophysical Research: Atmospheres*, *118*(4), 1716–1733.

Sillmann, J., V. Kharin, F. Zwiers, X. Zhang, D. Bronaugh, and M. Donat (2014), Evaluating model-simulated variability in temperature extremes using modified percentile indices, *International Journal of Climatology*, *34*(11), 3304–3311.

Simmons, A., P. Jones, V. da Costa Bechtold, A. Beljaars, P. Kållberg, S. Saarinen, S. Uppala, P. Viterbo, and N. Wedi (2004), Comparison of trends and low-frequency variability in CRU, ERA-40, and NCEP/NCAR analyses of surface air temperature, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *109*(D24).

Smith, D. M., A. A. Scaife, and B. P. Kirtman (2012), What is the current state of scientific knowledge with regard to seasonal and decadal forecasting, *Environmental Research Letters*, *7*, 015,602.

Smith, R. L., C. Tebaldi, D. Nychka, and L. O. Mearns (2009), Bayesian modeling of uncertainty in ensembles of climate models, *Journal of the American Statistical Association*, *104*(485), 97–116.

Stainforth, D., T. Aina, C. Christensen, M. Collins, N. Faull, D. Frame, J. Kettleborough, S. Knight, A. Martin, J. Murphy, *et al.* (2005), Uncertainty in predictions of the climate response to rising levels of greenhouse gases, *Nature*, *433*(7024), 403–406.

Stan, C., M. Khairoutdinov, C. DeMott, V. Krishnamurthy, D. Straus, D. Randall, J. Kinter, and J. Shukla (2010), An ocean-atmosphere climate simulation with an embedded cloud resolving model, *Geophysical Research Letters*, *37*(1), L01,702.

Stevenson, S., B. Fox-Kemper, M. Jochum, R. Neale, C. Deser, and G. Meehl (2012), Will there be a significant change to El Niño in the twenty-first century?, *Journal of Climate*, *25*(6), 2129–2145.

Stocker, T. (2011), *Introduction to climate modelling*, Springer Science & Business Media.

Stocker, T., D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley (2013), IPCC, 2013: Summary for Policymakers. 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.

Stringer, L. C., J. C. Dyer, M. S. Reed, A. J. Dougill, C. Twyman, and D. Mkwambisi (2009), Adaptations to climate change, drought and desertification: local insights to enhance policy in southern Africa, *Environmental Science & Policy*, *12*(7), 748–765.

Stroeve, J., M. M. Holland, W. Meier, T. Scambos, and M. Serreze (2007), Arctic sea ice decline: Faster than forecast, *Geophysical Research Letters*, *34*(9), L09,501.

Stroeve, J. C., V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, and W. Meier (2012), Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations, *Geophysical Research Letters*, *39*(16).

Sylla, M., F. Giorgi, E. Coppola, and L. Mariotti (2013), Uncertainties in daily rainfall over Africa: assessment of gridded observation products and evaluation of a regional climate model simulation, *International Journal of Climatology*, *33*(7), 1805–1817.

Taylor, K. (2001), Summarizing multiple aspects of model performance in a single diagram, *Journal of Geophysical Research*, *106*(D7), 7183–7192.

Taylor, K., R. Stouffer, and G. Meehl (2012), An overview of CMIP5 and the experiment design, *Bulletin of the American Meteorological Society*, *93*(4), 485.

Tebaldi, C., K. Hayhoe, J. Arblaster, and G. Meehl (2006), Going to the extremes, *Climatic Change*, *79*(3), 185–211.

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

Tebaldi, C., R. L. Smith, D. Nychka, and L. O. Mearns (2005), Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles, *Journal of Climate*, *18*(10), 1524–1540.

Terink, W., R. Hurkmans, P. Torfs, and R. Uijlenhoet (2010), Evaluation of a bias correction method applied to downscaled precipitation and temperature reanalysis data for the Rhine basin, *Hydrology and Earth System Sciences*, *14*(4), 687–703.

Teutschbein, C., and J. Seibert (2012), Is bias correction of Regional Climate Model (RCM) simulations possible for non-stationary conditions?, *Hydrology and Earth System Sciences Discussions*, *9*(11), 12,765–12,795.

Thorne, P. W., D. E. Parker, J. R. Christy, and C. A. Mears (2005), Uncertainties in climate trends: Lessons from upper-air temperature records, *Bulletin of the American Meteorological Society*, *86*(10), 1437–1442.

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

Van den Besselaar, E., M. Haylock, G. Van der Schrier, and A. K. Tank (2011), A European daily high-resolution observational gridded data set of sea level pressure, *Journal of Geophysical Research*, *116*(D11), D11,110.

Van der Linden, P., and J. Mitchell (2009), ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project, *Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK*, *160*.

van Oldenborgh, G. J., S. Drijfhout, A. v. Ulden, R. Haarsma, A. Sterl, C. Severijns, W. Hazeleger, and H. Dijkstra (2009), Western Europe is warming much faster than expected, *Climate of the Past*, *5*(1), 1–12.

van Oldenborgh, G. J., F. D. Reyes, S. Drijfhout, and E. Hawkins (2013), Reliability of regional climate model trends, *Environmental Research Letters*, *8*(1), 014,055.

Van Vuuren, D. P., J. Edmonds, M. Kainuma, K. Riahi, A. Thomson, K. Hibbard, G. C. Hurtt, T. Kram, V. Krey, J.-F. Lamarque, *et al.* (2011), The representative concentration pathways: an overview, *Climatic change*, *109*, 5–31.

Vannitsem, S. (2011), Bias correction and post-processing under climate change, *Nonlinear Processes in Geophysics*, *18*(6), 911–924.

Vautard, R., A. Gobiet, D. Jacob, M. Belda, A. Colette, M. Déqué, J. Fernández, M. García-Díez, K. Goergen, I. Güttler, *et al.* (2013), The simulation of European heat waves from an ensemble of regional climate models within the EURO-CORDEX project, *Climate Dynamics*, *41*(9-10), 2555–2575.

Wang, M., and J. E. Overland (2009), A sea ice free summer Arctic within 30 years?, *Geophysical Research Letters*, *36*(7), L07,502.

Wang, Y., L. R. Leung, J. L. McGregor, D.-K. Lee, W.-C. Wang, Y. Ding, and F. Kimura (2004), Regional Climate Modeling: Progress, Challenges, and Prospects, *Journal of the Meteorological Society of Japan, 82 (6): 1599-1628*, *82*(PNNL-SA-41328).

Watterson, I. (1996), Non-dimensional measures of climate model performance, *International Journal of Climatology*, *16*(4), 379–391.

Waugh, D., V. Eyring, *et al.* (2008), Quantitative performance metrics for stratospheric-resolving chemistry-climate models, *Atmospheric Chemistry and Physics Discussions*, *8*(3), 10,873–10,911.

Weaver, A., J. Sedláček, M. Eby, K. Alexander, E. Crespin, T. Fichefet, G. Philippon-Berthier, F. Joos, M. Kawamiya, K. Matsumoto, *et al.* (2012), Stability of the Atlantic meridional overturning circulation: A model intercomparison, *Geophysical Research Letters*, *39*(20), L20,709.

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

Weigel, A., M. Liniger, and C. Appenzeller (2008), Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?, *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 241–260.

Whetton, P., I. Macadam, J. Bathols, and J. O'Grady (2007), Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models, *Geophysical Research Letters*, *34*(14), L14,701.

Wilby, R. L., T. Wigley, D. Conway, P. Jones, B. Hewitson, J. Main, and D. Wilks (1998), Statistical downscaling of general circulation model output: a comparison of methods, *Water Resources Research*, *34*(11), 2995–3008.

Williams, K., and M. Webb (2009), A quantitative performance assessment of cloud regimes in climate models, *Climate Dynamics*, *33*(1), 141–157.

WMO (1992), Manual on the Global Data Processing System,World Meteorological Organisation, section III, Attachment II.7 and II.8, (revised in 2002).

Xu, Y., X. Gao, and F. Giorgi (2010), Upgrades to the reliability ensemble averaging method for producing probabilistic climate-change projections, *Climate Research*, *41*, 61–81.

Yin, J. (2012), Century to multi-century sea level rise projections from CMIP5 models, *Geophysical Research Letters*, *39*(17), L17,709.

Zhang, H., and Z. Pu (2010), Beating the uncertainties: ensemble forecasting and ensemble-based data assimilation in modern numerical weather prediction, *Advances in Meteorology*, *2010*.

Zhang, X. (2010), Sensitivity of arctic summer sea ice coverage to global warming forcing: towards reducing uncertainty in arctic climate change projections, *Tellus A*, *62*(3), 220–227.

Zhang, X., G. Hegerl, F. Zwiers, and J. Kenyon (2005), Avoiding inhomogeneity in percentile-based indices of temperature extremes, *Journal of Climate*, *18*(11), 1641–1651.