THE UNIVERSITY OF READING

Forecast Calibration and Combination: Bayesian Assimilation of Seasonal Climate Predictions

Caio Augusto dos Santos Coelho

A thesis submitted for the degree of Doctor of Philosophy

Department of Meteorology

January 2005

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Caio Augusto dos Santos Coelho

Abstract

The ultimate aim of this study is to produce improved probability forecasts of seasonal rainfall for South America. Such forecasts allow local governments plan their actions prior to the occurrence of climate anomalies such as those observed during El Niño-Southern Oscillation (ENSO) events. ENSO is one of the most important modes of climate variability affecting precipitation over South America. Therefore, the improvement of ENSO seasonal forecasts can improve the quality of rainfall forecasts for South America.

This study establishes a unified framework for the production of calibrated probability forecasts of observable variables based on information from ensembles of climate model predictions. In the same way that data assimilation is needed to get observed information into climate models, an analogous assimilation is required to convert multi-model climate predictions into well-calibrated forecasts of real-world observable variables. This Bayesian combination/calibration procedure is referred to as forecast assimilation.

The methodology, which allows the combination of coupled model with empirical predictions, is developed and tested in three stages. First, the Bayesian procedure is developed for the calibration of forecasts of an ENSO index (Niño-3.4) obtained from an individual coupled model. Next, the method is extended for the calibration and combination of equatorial Pacific sea surface temperature (SST) anomaly forecasts from seven DEMETER coupled models. Hence, in the second stage the method deals with multi-model forecasts and acquires the first spatial (longitudinal) component. Finally, in the third stage the Bayesian multi-model method is applied to the calibration and combination of spatial field forecasts of rainfall over South America. Results show that Bayesian combined forecasts are better calibrated and more reliable than both raw and bias-corrected coupled model forecasts.

Acknowledgements

A large number of people, who I am gratefully thankful, has supported and helped me during my PhD. First of all I would like to thank my supervisor David Stephenson for introducing me to the fascinating world of statistics and probability applied to clima te research. David has always been enthusiastic about my work and constantly challenged me with new ideas and concepts, encouraging me to always move away from excesses of details in order to have a broader view about academic climate research. I am also grateful to David for polishing my English communication skills, with are now much better than when I started my PhD, and also for helping me to develop a more accurate and critical scientific thinking.

I would also like to thank my co-supervisor Francisco Doblas-Reyes (Paco) for introducing me to the concept of multi-model ensemble forecasting and also for constantly asking me very thoughtful questions, which made me think more critically about my work. Paco has always been enthusiastic about my work and taught me how to acknowledge the real value of my research. I am also thankful to Paco for frequently encouraging me to make new contacts with researchers of correlated areas, and for stimulating me to get involved in other research projects.

Next, I would like to thank all members of my thesis committee, Alan O'Neil, Ellie Highwood and George Craig, for their constructive comments and suggestions that helped to improve my PhD work and kept me on track. I am also thankful to Sergio Pezzulli an d Magdalena Balmaseda, who substantially contributed for the development of part of this work. The help and support of all members of the Climate Analysis Group is also gratefully appreciated. Particularly, I am grateful to Chris Ferro, Ian Jolliffe and M att Sapiano for their valuable statistical and computational insights. Thoughtful comments by Geert Jan van Oldenborgh and Alexandre Pezza on the South American rainfall work are also gratefully appreciated.

I am thankful to Tim Palmer and David Anderson, who provided the model data used in this study, and Alexandre Guetter, who provided the river flow data. I am particularly thankful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (C NPq) for granting me a scholarship for the development of this PhD research.

Finally, I would like to express my gratitude to Simone Coelho, my wife, for her support during these years of hard work and for accepting the challenge of marrying me and moving to England to study and construct a better future for both of us and our fut ure descendents in Brazil. In fact, I am gratefully thankful to my whole family for all the support and encouragement always provided, especially from my parents João Luiz Coelho and Neuza Coelho and my brother Paulo Coelho. To conclude, I could not f orget to thank all my friends, who I am not going to list all here, that followed all this process, for their companionship and for all those relaxing happy moment we have experienced together.

Contents

Chapter 1

Introduction

1.1 Motivation

The need to produce timely forecasts of the ever-changing environment is a major driving factor in weather and climate research. The challenge of producing and improving forecasts has generated many fascinating areas of research in atmospheric and oceanic science such as numerical weather prediction, data assimilation, statistical down-scaling, single-model ensemble prediction, and more recently multi-model ensemble prediction. Much effort has gone into producing reliable and informative forecasts of future observable variables given knowledge of observable variables at earlier times. Such forecasts are inherently probabilistic because of the uncertainties in environmental data (e.g. measurement errors and sparse coverage) and structural and parametric uncertainty in numerical prediction models.

Recent studies have focussed on how best to combine multi-model climate predictions (Kharin and Zwiers 2002; Krishnamurti *et al.* 1999, 2000a, 2000b and 2001; Coelho *et al.* 2003, 2004, etc.). There are many possible methods for combining forecasts (see section 2.3 in chapter 2 for a review) but no unique method can be prescribed that is ideal for all the types of weather/climate forecasting problems. In other words, there is no universally good method for all forecasting problems. However, there is a need to develop a framework that can incorporate the different approaches for combining weather and climate predictions in order to provide the most informative forecasts of future observables. In addition to the issue of how best to combine multiple predictions, there is also an important issue as to how best to calibrate the predictions. There are many reasons why model predictions should not be taken on face value as forecasts of observable variables (e.g. drift in coupled models that leads to bias) and so calibration is a necessary step in obtaining useful information for forecast users and risk/impact assessments. This thesis addresses both the issue of calibration and the issue of combination of seasonal climate predictions.

This study establishes a unified framework for the production of calibrated probability forecasts of observable variables based on information from ensembles of climate model predictions. In the same way that data assimilation is needed to get observed information into climate models, it is argued that an analogous assimilation is required to convert climate model predictions into well-calibrated forecasts of real-world observable variables. This Bayesian combination/calibration procedure is referred to here as *forecast assimilation* (see section 3.2 of chapter 3). Forecast assimilation is a unified framework that extends previous approaches such as statistical down-scaling and model output statistics (Glahn and Lowry 1972) for use with multi-model ensemble predictions.

1.2 Aim and strategy

The aim of this study is to produce improved and well-calibrated seasonal probability forecasts of rainfall for South America through the appropriate application of statistical modelling methods. Statistical modelling methods have already been developed for producing empirical predictions and also for the calibration of deterministic predictions produced by physically derived dynamical models. The calibration of deterministic predictions against past observation is known as model output statistics. However, with the availability of ensemble climate predictions from different coupled models there is clearly the need for development of statistical modelling methods for the calibration and combination of these predictions. Bayesian forecast assimilation is the method that is here developed and proposed for dealing with multi-model ensemble predictions. Rather than directly address South American rainfall, the methodology is developed progressively in three stages. First, the Bayesian procedure is developed for the calibration of forecasts of a single ENSO index (Niño-3.4) obtained from an individual coupled model. Therefore, in this first stage the method is developed for a single point in space. Next, the method is extended for the calibration and combination of equatorial Pacific sea surface temperature (SST) anomaly forecasts from seven distinct coupled models. Hence, in the second stage the method deals with multi-model forecasts and acquires the first spatial (longitudinal) component. Finally, in the third stage the Bayesian multi-model method is applied to the calibration and combination of spatial field forecasts of rainfall over South America.

1.3 Thesis plan

Chapter 2 provides the background of the thesis: an overview of seasonal forecasting; a brief description of the DEMETER project [1] (Palmer *et al.* 2004) and its datasets that are used in this study; and a literature review on forecast calibration and combination. Chapter 3 establishes a unified framework for forecasting, defines the concept of forecast assimilation, and describes the Bayesian methodology applied in this study. Chapter 4 presents the results of the first application (univariate zero-dimensional Niño-3.4 forecasts). Chapter 5 shows the results of the second application (multivariate one-dimensional equatorial Pacific SST forecasts). Chapter 6 provides a brief literature review of South America seasonal forecasting and presents the results of the third and final application (multivariate two-dimensional South American rainfall forecasts). Finally, chapter 7 concludes the thesis with a summary of findings and a discussion of possible future areas of research.

[1] The DEMETER acronym stands for Development of a European Multi-model Ensemble system for seasonal to inTERannual prediction. See website http://www.ecmwf.int/research/demeter

Chapter 2

Background

2.1 Aim

This chapter presents the background necessary for understanding the two scientific issues addressed in this thesis: forecast calibration and combination. An overview on seasonal forecasting and a literature review on forecast combination are presented here.

2.2 Seasonal forecasting

2.2.1 Overview

Traditional weather forecasts provide information about the weather that is likely during the next few days. Although it is not possible to precisely predict daily weather changes beyond about a week in advance (Lorenz 1963), it is possible to make inferences about likely future conditions averaged over periods of a few months. Seasonal forecasts provide information on these long-term time averages (usually greater than one month but less than one year).

Day-to-day weather is, however, largely unpredictable on seasonal time-scale. Therefore seasonal forecasts are probabilistic, with probabilities that can be estimated from ensembles of predictions obtained from climate models. Given the unpredictable chaotic nature of daily weather, it is natural to question the feasibility of seasonal forecasting. Seasonal forecasting is feasible because the atmospheric variability on seasonal time scale is modulated by slowly varying boundary forcing (Charney and Shukla 1981; Brankovic *et al.* 1994). Conditions at the Earth's surface, such as slow fluctuations in surface temperature, soil moisture and snow cover, are able to influence weather patterns. These influences are hardly noticed in daily weather events but become noticeable in long-term weather averages. In particular, SSTs can effectively modulate seasonal climate variability. Slow fluctuations in SST can be predicted, with various degrees of accuracy depending on the ocean basin, up to one year ahead (Davis 1976; Balmaseda *et al.* 1994; Davey *et al.* 1994; Wu *et al.* 1994) and the relationship between SST and weather can be represented in computer models of the atmosphere and ocean. These models form the dynamical basis for seasonal forecasting.

The strongest links between SST patterns and seasonal climate are found in the tropics, and it is here that seasonal forecasts are most successful (Goddard *et al.* 2001). The most well-known predictability is associated with the El Niño-Southern Oscillation (ENSO), which is a coupled ocean-atmosphere phenomenon that occurs on average every three to seven years (Philander 1990). ENSO has two phases, depending on the conditions of the upper tropical Pacific ocean: a warm phase, known as El Niño, when warmer than normal conditions are observed in the tropical Pacific; and a cold phase, known as La Niña, when colder than normal conditions are observed in the tropical Pacific. During its warm phase sea level pressure (SLP) is below normal in the central-east tropical Pacific and above normal in the west tropical Pacific/Indonesian region. During its cold phase this pressure pattern reverses, with the central-east tropical Pacific presenting above normal SLP and the west tropical Pacific/Indonesian region experiencing below normal SLP. ENSO can change weather/climate not only in the tropics, but also in several other regions around the world via atmospheric teleconnection (Wallace and Gutzler 1981; Trenberth *et al.* 1998). It can affect seasonal rainfall in remote locations away from the tropics leading to droughts in some regions and floods in others (Ropelewski

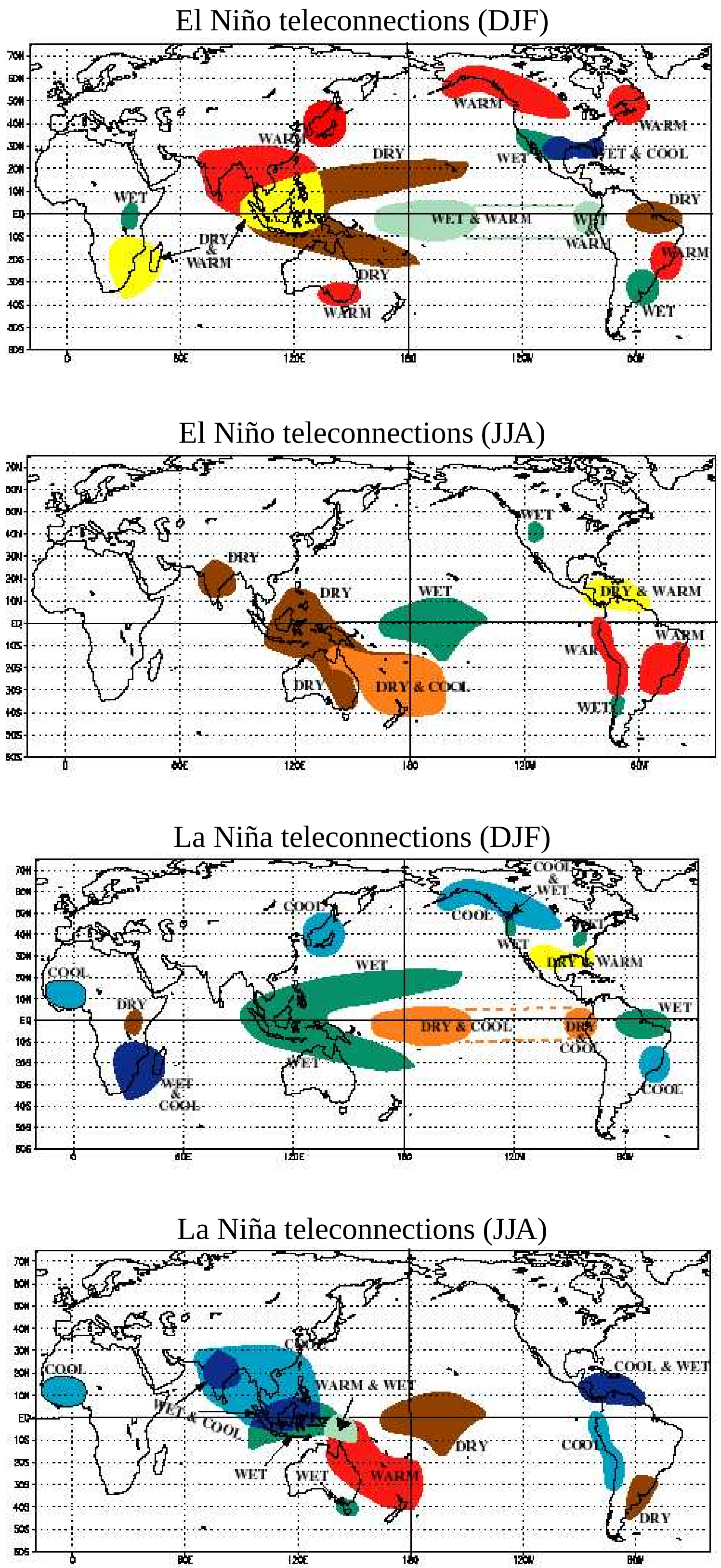


Figure 2.1: El Niño) and La Niña observed teleconnections during December, January and February (DJF) and June, July and August (JJA). Source: Climate Prediction Center (http://www.cpc.ncep.noaa.gov, 2004).

and Halpert 1987, 1989). Figure 2.1 shows El Niño and La Niña observed teleconnections during December, January and February (DJF) and June, July and August (JJA). Some climate models are able to predict these patterns. However, the current generation of climate models provides forecasts that are only marginally skillful in the extra-tropics (Graham *et al.* 2000) and slightly more skillful in the tropics (Goddard *et al.* 2001). This is because tropical areas have a moderate amount of predictable signal, whereas in the mid-latitudes random weather fluctuations are generally larger than the predictable component of the weather.

Variations in the Pacific SST, however, are not the only sources of predictability for weather patterns. Warm or cold SSTs in the tropical Atlantic or Indian oceans can influence seasonal climate in nearby continents. For example, the SST in the western Indian ocean has a strong effect on precipitation in tropical eastern Africa (Rocha and Simmonds 1997; Goddard and Graham 1999), and the SST in the tropical Atlantic affect rainfall in Northeast Brazil (Moura and Shukla 1981). In addition to the tropical oceans, snow cover and soil wetness also influence seasonal climate (Douville and Royer 1996; Douville 2003; Dirmeyer 2003). When snow cover is above average in a specific region for a given season, it has a greater than usual cooling effect on the air of this region. Soil wetness, which comes into play most strongly during warm seasons, also has a cooling influence. In summary, all these factors that affect the atmospheric circulation (SST, snow cover and soil wetness) constitute the basis of long-term predictions.

Zebiak and Cane (1987) first demonstrated the feasibility of seasonal prediction of ENSO using an idealised dynamical model. Operational seasonal forecasting started experimentally in the mid-1990s. One of the first experimental seasonal forecasts was produced at the European Centre for Medium-range Weather Forecasts (ECMWF) in 1995. Since then several other centres around the world have also developed seasonal forecasting systems (e.g. Stockdale *et al.* 1998; Mason *et al.* 1999; Kanamitsu *et al.* 2002; Alves *et al.* 2002). This successful development can be attributed to: a) improvements in our understanding of the coupled ocean-atmosphere system during the second half of the 20^{th} century (Neelin *et al.* 1998),

b) the development and use of buoys to observe and measure the evolution of near-surface waters in the tropical Pacific (McPhaden *et al.* 1998), and c) successful predictions of El Niño by prototype coupled ocean-atmosphere models (Zebiak and Cane 1987). Nowadays, although still experimentally, more than a dozen systems are routinely used to make predictions and their results are published in the Experimental Long-Lead Forecast Bulletin (http://www.iges.org/ellfb/). Most attention is given to ENSO because it has such a large global impact (Fig. 2.1) and much seasonal forecast skill derives from ENSO, with non-ENSO years presenting minimal skill (Barnston *et al.* 1999a; Landman and Mason 1999).

In terms of climate variables of interest, until recently, most emphasis had been placed on the prediction of anomalous geopotential heights at 700 hPa (important level for moisture transport), 500 hPa (important steering level for mid-latitude storms) and 200 hPa (important diagnostic of upper level divergence and thus convective activity in the tropics). Today, the typical variables examined for prediction purposes are those that concern society most, i.e. near-surface air temperature, precipitation, wind and atmospheric pressure.

Seasonal forecasts are potentially valuable to several sectors of society, such as agriculture (Challinor *et al.* 2005), tropical health (Morse *et al.* 2005) and electricity generation (Palmer 2002). Seasonal forecasts would be more valuable if their skill were improved. Such an improvement would be particularly beneficial for the people of developing countries such as those in South America. Brazil, in particular, has more than 90% of the electricity produced by hydropower stations (http://www.ons.org.br). The installed hydroelectric capacities in Paraguay, Brazil, Uruguay, Colombia, and Bolivia constitute more than 60% of the electricity production of these countries (http://encarta.msn.com). Hydroelectric power is also important in Peru, Chile, Ecuador, Suriname and Argentina, where installed hydroelectricity generating capacity accounts for more than 40 percent of all generating capacity. Good quality seasonal forecasts allow local governments to make more effective plans prior to the occurrence of climate anomalies such as those observed during ENSO events. ENSO is one of the most important modes of climate variability affecting precipitation over South America (Ropelewski and Halpert 1987, 1989). Therefore, the improvement of ENSO seasonal forecasts indirectly helps improve the quality of precipitation forecasts for South America. This thesis focuses on both, improvement of ENSO and South American precipitation seasonal forecasts. A review of South American rainfall seasonal forecasting is presented in Chapter 6, section 6.2.

2.2.2 Current forecast approaches

Seasonal forecasts are currently produced in five different ways. The simplest methods for seasonal climate prediction are purely based on observations of past and present climate. The use of the climatological average as the prediction is a valid starting point. This method is known as *climatological forecast* or simply *climatology*. Alternatively, one can assume that a recent seasonal climate will persist through the upcoming season (Huang *et al.* 1996). This is the so-called *persistence forecast*. The development of prediction methods that are superior to these two simple approaches is the goal of most climate prediction research. Three distinct modelling approaches are currently used: a) empirical (statistical) relationships based on historical data, b) two-tier dynamical forecast systems, i.e. using atmospheric general circulation models (AGCM) forced with predicted or persisted SSTs, or c) single-tier dynamical forecast systems, i.e. using fully coupled ocean-atmosphere general circulation models (CGCM) that predict the joint evolution of SST and atmospheric flow.

The majority of the empirical forecasting systems use SSTs as predictors (e.g. Pezzi *et al.* 2000 and Folland *et al.* 2001), which are also the fundamental source of predictability in two-tiered forecast systems (Colman and Davey 2003). Some empirical modelling studies claim to have improved forecast skill when, in addition to SSTs, atmospheric predictors are included (Hastenrath *et al.* 1995; Francis and Renwick 1998). The comparative skill of forecasts provided by physically-derived dynamical climate models and empirical models is a subject of much debate (Berliner *et al.* 2000b). Recent forecast comparisons suggest that empirical models perform at least as well as physically-derived dynamical climate models (Barnston *et al.* 1999b; Anderson *et al.* 1999). Some studies argue that empirical models perform better (e.g. Landsea and Knaff 2000), while other studies claim that dynamical climate models can provide better forecasts (e.g. Trenberth 1998).

Physically-derived dynamical models have an intrinsic but yet unrealised advantage. They can be improved in order to acquire a more comprehensive representation of real world physical processes. Therefore, there is possibility for further improvements in prediction skill. Additionally, they are not limited by non-stationarity of climate or by extreme or unusual (rare) outcomes that may not have occurred in the available historical record. Empirical models are less complex and have the advantage of being computationally much cheaper to run. These models, however, generally assume stationarity of climate and their improvements are limited to the increase of volume of observational data, a task fundamentally more limited, although linear trends in time can be empirically modelled. Thus, considerable energy is now being devoted for improving physically-derived dynamical models at many centres around the world. Physically-derived dynamical models, however, are not entirely independent of empirical assumptions. Because of the relatively coarse spatial resolution used in AGCMs and CGCMs (approximately 200 km resolution), physical processes occurring on scales smaller than the model's grid can resolve are parameterised empirically. Mathematical relationships based on observed data describe the large-scale aggregate behaviour of fundamentally small-scale processes such as convection and radiative transfer.

A primary drawback of the current generation of CGCMs is that the SST field tends to drift away from realistic values as the integration proceeds, thus forcing unrealistic patterns of atmospheric anomalies. Drift can occur rapidly because of an imbalance in initial conditions, or slowly, because of faults in internal parameterisation in one of the components of the model or from the coupling itself (i.e., flux errors) (Délécluse *et al.* 1998). Drift in coupled models leads to systematic biases in the pattern (i.e., shift of the forecast compared to the observed pattern), amplitude (i.e., underestimation or overestimation of the intensity of the climate signal) and variance (i.e., underestimation or overestimation of the spread of the ensemble) of climate simulations (Smith and Livezey 1999; Anderson 1996).

Goddard *et al.* (2001) suggest two components of the current climate models that need to be improved: physics/parameterisation schemes and data assimilation. For the atmosphere, parameterisation of convection and marine stratus is particularly important for atmospheric energetics and needs to be improved. Improvements are also necessary for boundary layer parameterisations, which control fluxes that govern coupling between ocean and atmosphere. For the oceans, the main concern is the improvement of the parameterisation of mixing processes in surface layers (first 150 metres). In terms of data assimilation, Goddard *et al.* (2001) suggest that most efforts should be concentrated on ocean data assimilation because the ocean contains the primary memory for time-scales longer than a few weeks. Furthermore, this is particularly important for seasonal forecasting because the observations are much more limited in oceanic regions than over land. Therefore, data assimilation methods are expected to optimise their information content. Methods for assimilating additional data sources that are presently becoming available (e.g., new satellite measurements and in situ monitoring of upper-ocean temperatures and salinity via programmable buoys) also need to be developed/improved.

2.2.3 Ensemble forecasting

For both medium-range (from 3-10 day ahead) and seasonal forecasts, it is common practice to use the ensemble technique to produce probabilistic forecasts (e.g. Tracton and Kalnay 1993, Molteni *et al.* 1996; Stockdale *et al.* 1998). An ensemble is a set of forecasts that verify at the same time (Sivillo *et al.* 1997). Each forecast in the ensemble is referred to as a member. The ensemble members can differ in their initial boundary conditions and initial atmospheric conditions. The ensemble technique aims to sample uncertainties in the initial conditions used to produce the forecasts. Ensemble-based predictions lead naturally to probabilistic forecasts.

However, forecast uncertainty derives not only from model weakness, but also from inherent unpredictability of the precise state of the atmosphere (Lorenz 1963). Even with a perfect model, atmospheric internal variability would still impose uncertainty on the most likely climate outcome. Probabilistic forecasting provides a way of addressing both sources of forecast uncertainty (i.e. uncertainty in initial conditions and uncertainty due to atmospheric internal variability) by indicating the probability distribution of expected possible outcomes (Kumar *et al.* 2000). Statistical and physically-based dynamical methods have been developed for estimating the probability density function (p.d.f.) of climate outcomes for a coming season, and this is the ultimate aim of this study.

Figure 2.2 shows an example of ECMWF coupled model ensemble forecasts of the mean SST anomaly for the Niño-3 region, bounded by 90^{o}W-150^{o}W and 5^{o}S-5^{o}N in the equatorial Pacific, which is indicated in the top left corner of the figure. Forecasts were made at weekly intervals from November 1996. Each forecast is for 6 months. The thick dark blue line shows the observed SST anomalies. Although three forecasts have been made for each week, only one forecast of each week is shown in order to avoid clutter (red plumes). The onset of El Niño 1997/98 in April 1997 was well predicted by this coupled system. The subsequent evolution was also well forecast, although a tendency to underpredict the amplitude of the event was noticeable. These forecast plumes can be calibrated against observations in order to produce reliable forecast probabilities. At this point it is worth stressing the distinction between climate model outputs and observed climate/weather. Climate model outputs should not be treated as observed climate because they contain model structural and parametric errors, which must be corrected by calibration against observations. In other words, climate model outputs are variables in model space that should be calibrated against observations in order to better approximate the observed climate in real-world observation space. The calibration issue is addressed in chapter 3.

Given the two distinct approaches to seasonal forecasting, i.e., empirical and physically-derived dynamical modelling, it is natural to ask whether combining

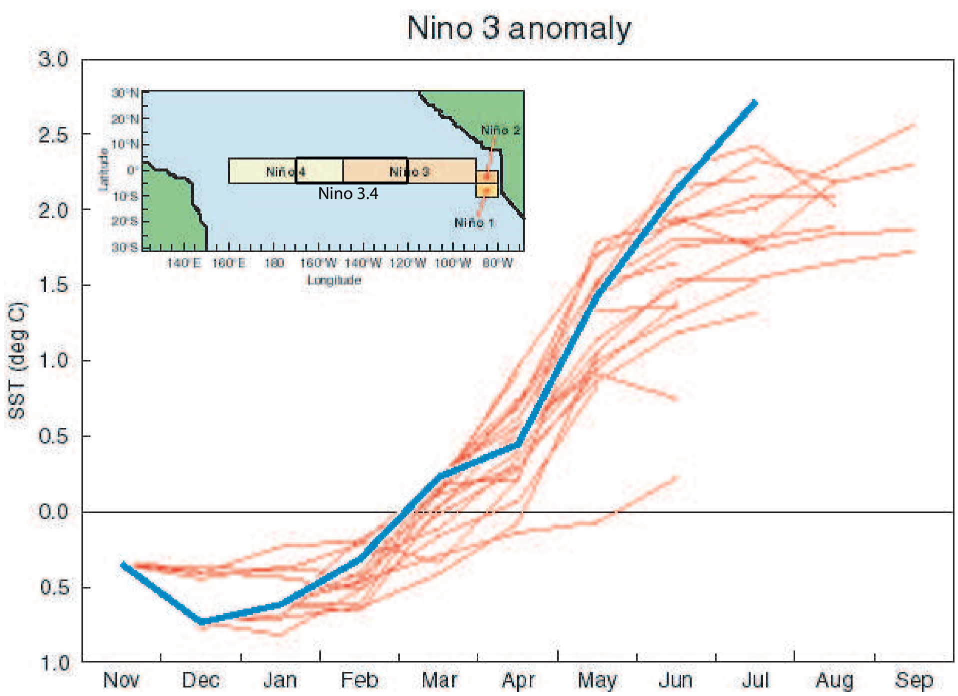


Figure 2.2: ECMWF coupled model Niño-3 SST anomaly forecasts (red plumes) made at weekly intervals from November 1996. Each forecast is for 6 months . The thick dark blue line shows the observed SST anomalies. See text for further information. Source: ECMWF web page (http://www.ecmwf.int, 2004).

them may produce a forecast with more skill than either forecast considered separately. Additionally, the use of ensembles of forecasts not only from a single climate model but from several different climate models helps to sample uncertainty in model formulation (Palmer 2000). Different climate models may be skillful in different regions. Therefore, the calibration and combination of ensembles of forecasts from distinct climate models with empirical forecasts (i.e., using all available information) is likely to better sample forecast uncertainty and, if the individual forecasts are skillful, to improve forecast quality. This thesis addresses both calibration and combination of ensemble forecasts provided by physically-derived dynamical and empirical models in order to produce a reliable p.d.f. for long-range climate predictions.

2.2.4 The DEMETER project

DEMETER[1] is the acronym of the European Union (EU) funded project entitled Development of a European Multi-model Ensemble system for seasonal to inTER-annual prediction (Palmer *et al.* 2004). The aim of the project was to develop a well-validated European coupled multi-model ensemble forecasting system for reliable seasonal to inter-annual prediction.

The fundamental idea behind the multi-model approach used in DEMETER is that climate prediction imperfections are caused by two main sources of errors: a) initial conditions and b) model formulation. An ensemble forecast of an individual coupled model samples uncertainties in the initial conditions used to produce the forecast. Uncertainties in the model formulation are not sampled by this single-model ensemble approach. However, different models use different numerical/parameterisation schemes to represent mathematically the same physical processes. Although these models use the same basic equations for the evolution of climate, the numerical representation of these equations on digital computers differs from model to model. The multi-model ensemble, consisting of ensemble forecasts produced by different climate research institutions, allows the estimation of uncertainties in model formulation.

The ability of multi-model ensembles of AGCM forecasts to produce more reliable probability forecasts of seasonal climate than single-model ensembles has been addressed by the EU-funded PROVOST (Prediction of Climate Variations on Seasonal to Inter-annual Time-scales) project and also by the DSP (Dynamical Seasonal Prediction) project in the United States (Palmer and Shukla 2000). PROVOST had several AGCMs integrated over four months with prescribed observed SSTs. Nine ensemble members were generated for each model from each starting date. Despite the use of identical prescribed SSTs, results showed large model-to-model variability in the predictions of SSTs and other variables such as precipitation and geopotential height at 500 hPa (Pavan and Doblas-Reyes 2000). Similar results

[1] Refer to http://www.ecmwf.int/research/demeter for more information about this project.

have also been found in DSP (Straus and Shukla 2000). This inter-model variability made single-model ensemble forecasts generally unreliable. The multi-model ensemble approach, on the other hand, has produced more reliable forecasts than any single-model. The skill of both single and multi-model forecasts has been assessed using probabilistic skill scores (Doblas-Reyes *et al.* 2000; Graham *et al.* 2000; Palmer *et al.* 2000). The results showed that multi-model ensemble forecasts produced higher scores than any of the single-model ensembles. These results stimulated the development of the DEMETER project.

The major aim of DEMETER was to advance the concept of multi-model ensemble predictions. Seven state-of-the-art global coupled ocean-atmosphere models were used to produce series of multi-model ensemble hindcasts (i.e., retrospective forecasts made after the events are observed) of common climate variables. The seven modelling partners are listed in Table 2.1. All partners produced hindcasts for a common period from 1980 to 2001 (22 years). Three of the seven coupled models (CNRM, ECMWF and UKMO) produced hindcasts for the period 1959 to 2001 (43 years). All coupled models were run four times per year, starting the first day of February, May, August and November at 00:00 GMT. Nine ensemble forecasts (i.e., nine members) were produced for each coupled model for the next six months including the starting month. Wind stress and SST perturbations were used to generate the ensemble for each model. Atmospheric and land-surface initial conditions were taken from the ERA-40 project (ECMWF 40 years re-analysis project). Additional information about ensemble generation and atmospheric/oceanic initial conditions can be found in Palmer *et al.* (2004)

Chapter 3 of this thesis presents statistical techniques for the correction (calibration) of coupled model systematic-errors and the combination of hindcasts provided by the participants of DEMETER. Chapter 4 examines the ability of ECMWF coupled model in forecasting the mean SST for the Niño-3.4 region (Fig. 2.2) in the central Pacific. Chapters 5 and 6 investigate the ability of the DEMETER multi-model system in reliably forecasting equatorial Pacific SSTs and South American rainfall anomalies, respectively. The feasibility of applying DEMETER multi-

Institution	Acronym	Country
Météo-France (Centre National de Recherches Météorologiques)	CNRM	France
European Centre for Research and Advanced Training in Scientific Computation	CERFACS	France
Laboratoire d' Océanographie Dynamique et de Climatologie	LODYC	France
Istituto Nazionale de Geofisica e Vulcanologia	INGV	Italy
European Centre for Medium-Range Weather Forecasts	ECMWF	International organisation
Max-Plank Institut für Meteorologie	MPI	Germany
Met Office	UKMO	U.K.

Table 2.1: DEMETER modelling partners.

model hindcasts to scales smaller than the scale that can be resolved by global climate models is also assessed in chapter 6, with a statistical down-scaling application example of river flow forecast for the Tocantins river in the north of Brazil.

2.3 Calibration of forecasts

Physically-derived dynamical model predictions are never perfect and therefore calibration against observations is necessary. The simplest possible method of calibration is *bias-correction*. This method consists in determining the historical mean forecast error, which is then subtracted from the forecast value in order to produce a calibrated forecast. Calibration can then be considered as a way of obtaining predictions with average statistical properties similar to those of a reference observational dataset.

Forecast calibration has first been addressed for the calibration of deterministic weather forecasts by Glahn and Lowry (1972), who introduced the model output statistics (MOS) technique. MOS is an objective technique that consists of determining a statistical relationship between an observable predictand (e.g. temperature at a particular location) and variables forecast by a physically-derived dynamical model at some projection time(s) in the future. The MOS technique linearly corrects systematic model errors via the regression of observed predictands on model forecast variables. A reasonably long calibration period (for seasonal forecasts this period is usually more than 10 years), where both predictand observations and retrospective model forecasts are available, is required to construct the calibration equations. When a new set of forecasts is issued these regression equations are then used to produce a calibrated forecast. MOS systems have been implemented at a number of national meteorological centres, including those in Italy (Conte *et al.* 1980), Britain (Francis *et al.* 1982), Canada (Brunet *et al.* 1988), the Netherlands (Lemcke and Kruizinga, 1988), and the United States (Carter *et al.* 1989).

Spatial MOS techniques, based on multivariate statistical methods, have also been developed in order to correct systematic misplacements in model predictions (Feddersen *et al.* 1999). In such systems, calibration is achieved by adjusting model simulations taking into account the linear relationship between the predicted and the observed spatial patterns. This approach allows spatial predicted patterns to be shifted around in order to better reproduce the observed patterns. Spatial MOS calibration has also been used for down-scaling of seasonal predictions (Feddersen 2003; Feddersen and Andersen 2005).

Only the calibration of deterministic forecasts has been discussed so far. However, as previously mentioned in section 2.2, well-calibrated probabilistic forecasts are more desirable than deterministic forecasts. Therefore, not only well-calibrated estimates of the forecast mean values are required but estimates of forecast uncertainty are also required. Forecast uncertainty can be estimated from ensemble of predictions produced by physically-derived dynamical models (see section 2.2.3). The spread of the ensemble is usually used to estimate forecast uncertainty. The mean and the spread of the ensemble can be used to produce probabilistic forecasts. Generally it is found that dynamical models underestimate the forecast uncertainty (Atger 2003), making the forecast unreliable in the sense that the range of forecast values does not contain the observed values. Unrealistic predictions may be due to biases linked to either model errors or weakness of the method used for ensemble generation (Atger 2003). Therefore calibration by inflation of the ensemble spread is usually performed (e.g. Hamill and Colucci 1998; von Storch 1999) in order to improve the reliability of the forecasts. Reliability refers to the correspondence between the forecast probability of an event and the relative frequency of the event, conditioned upon the forecast probability (see Appendix D for a mathematical definition of reliability). Reliability is a measure of forecast uncertainty correctness and can be used as a synonym for calibration.

This brief review on forecast calibration reveals two main concerns: a) adjustment of predicted patterns to correct systematic errors and misplacements in model predictions; and b) ensemble spread correction to improve the reliability of the predictions. However, these two problems are generally tackled separately. For instance, a recent paper by Doblas-Reyes *et al.* (2005) has used a spatial MOS technique to calibrate seasonal climate model predicted patterns and an inflation technique to improve forecast reliability. Chapter 3 of this thesis proposes a method that is able to deal with these two problems simultaneously. This method is tested in three different application examples in chapters 4, 5 and 6.

2.4 Combination of forecasts

2.4.1 Introduction

The rapid development of computers and numerical methods during the last century stimulated the production of a variety of forecasts in several areas of science. Given the increasing availability of forecasts it is natural to ask whether combining them may produce a forecast with more skill than each forecast considered separately. However, the combining of forecasts is still a subject of much debate and controversy in the literature (see Palm and Zellner 1992; Yang 2004). Although it seems plausible that combined forecasts are better than individual forecasts (e.g. Clemen 1989; Krishnamurti *et al.* 1999; Pavan and Doblas-Reyes 2000), some issues have not yet been completely resolved. Among these issues are questions such as: What is the “best” method for combining? Some forecasting systems contain larger biases (errors) than others. The question of whether or not it is worth combining unbiased forecasts with biased forecasts when compared with individual unbiased forecasts has not yet been completely answered. Trenkler and Gotu (2000) compiled a list of approximately 600 publications on this area during the period from 1970 to 2000. This brief review does not intend to compile all possible methods. It summarises the most common combination methods used in economics and atmospheric sciences.

The fact that forecasts need to be combined suggests that all the models are mis-specified (Chatfield 2001, Section 4.3.5). This indicates that it may be better to devote more effort to refining the best of the models so that it encompasses the rest, or alternatively to find a more theoretically satisfying way of combining several plausible models, such as for example Bayesian Model Averaging (BMA) - a methodology that accounts for model uncertainty (Hoeting *et al.* 1999). BMA is the weighted average of posterior distributions (see definition in section 3.3) of different models. It is suggested that BMA based prediction intervals are better calibrated than single-model based prediction intervals, tending the latter to be usually too narrow (i.e. predictions are overconfident). A comprehensive review on BMA is presented by Hoeting *et al.* (1999). Some guidance regarding the assessment of prediction intervals for combined forecasts is given by Taylor and Bunn (1999).

The combination issue was addressed long ago by Laplace (1818). As described by Stigler (1973), Laplace was interested in comparing the properties of two estimators of the slope parameter in a problem of linear regression through the origin. By examining the joint distribution of these estimators he deduced a combining formula, which was a linear combination of the two estimators. However, his main conclusion was that the method would only be feasible if the error distribution of the response variable of his linear regression problem was known.

According to Hoeting *et al.* (1999), one of the earliest model combination studies in the statistical literature was performed by Barnard (1963) in a paper investigating airline passenger data. However, the majority of the early work in model/forecast combination has not been published in statistical journals. In fact, forecast combination has been most widely applied in economics and atmospheric sciences as summarised below.

2.4.2 Forecast combination in economics

Reid (1968, 1969) and Bates and Granger (1969) are often referred to as the seminal works on combining forecasts in economics. They were the first to develop a general analytical linear model for combining forecasts in an optimal way and to apply their techniques to real world situations. Their work provided the initial impetus to the development of theory in the combination of forecasts. A great number of articles have been published since then, and as suggested by de Menezes *et al.* (2000) these numbers still continue to grow substantially. Clemen (1989), Granger (1989), Diebold (1998, Section 12.3) and de Menezes *et al.* (2000) all provide comprehensive reviews on combining economic forecasts.

Clemen (1989) lists more than 200 studies in the form of an annotated bibliography, which contributed to knowledge regarding the combination of forecasts, either through theory or application. The main conclusion of this paper is that forecast accuracy can be substantially improved through the combination of multiple individual forecasts. Furthermore, it also suggests that the simple arithmetic mean of individual forecasts often works reasonably well when compared to more complex combinations. De Menezes *et al.* (2000) review recent studies on forecast combination and also provide guidance on the use of combined forecasts not only based on the accuracy of forecasts but also on three other properties of the forecast error: variance, asymmetry and serial correlation. In an attempt to answer the question of which was the "best" method for combining forecasts, they concluded that the most appropriate combination method depends on the choice of the forecast error property (variance, asymmetry or serial correlation) and sample size.

A summary of the seven well-established forecast combination methods most commonly used in economics is given by de Menezes *et al.* (2000). All these methods adopt the linear combination of forecasts in which each forecast has a pre-estimated weight that all add up to one. These seven weighting methods are as follows:

a) Simple average.

All forecasts have the same weights equal to $1/m$, where m is the total number of forecasts to be combined. As suggested by Clemen (1989), this forecast often performs as well as more sophisticated methods.

b) Outperformance.

This method was proposed by Bunn (1975). The weights are probabilities assessed and revised in a Bayesian manner. Each individual weight is interpreted as the probability that its respective forecast will perform the best (in the smallest absolute error sense) on the next occasion. Each probability is estimated as the fraction of occurrences in which its respective forecasting model has performed the best in the past.

c) Optimal.

This seminal method for combining forecasts was proposed by Bates and Granger (1969). The weights are determined in order to minimize the combined forecast error variance. Diebold and Lopez (1996) refer to this method as the "variance-covariance" method because the weights are obtained using the covariance matrix of forecast errors. Granger and Ramanathan (1984) showed that the method is equivalent to a least squares regression in which the constant is suppressed and the weights are constrained to sum to one. This approach requires the covariance matrix of forecast errors to be properly estimated. In practice this matrix is often not stationary, in which case it is estimated on the basis of a short history of forecasts and thus the method becomes an *adaptive* approach to combining forecasts.

d) Optimal (*adaptive*) with independence assumption.

The covariance matrix of forecast errors is restricted to be diagonal, comprising just the individual forecast error variances (Bunn 1985).

e) Optimal (*adaptive*) with restricted weights.

As well as the diagonal restriction, individuals weights are restricted not to be outside the interval [0,1] (Newbold and Granger 1974).

f) Regression.

The combined forecast is obtained via ordinary least squares (OLS) regression with the inclusion of a constant (Granger and Ramanathan 1984).

g) Regression with restricted weights.

A least squares regression with the inclusion of a constant is performed but the weights are constrained to sum to one (Holden *et al.* 1990).

2.4.3 Forecast combination in atmospheric sciences

Sanders (1963) was one of the first studies in the meteorological literature to discuss the possibility of combining forecasts to produce probability forecasts. By averaging probability forecasts of two individual forecasts, Sanders (1963) demonstrated using the Brier score (Brier 1950) that the combined probability forecast performed better than either individual forecast. Similar results were obtained by Stael von Holstein (1971) and Winkler *et al.* (1977) while studying the performance of various probability consensus approaches. Clemen (1985) and Clemen and Murphy (1986a, b) developed Bayesian forecast combination techniques to determine the contributions of the different forecasts used for the combination. More recently Rajagopalan *et al.* (2002) and Robertson *et al.* (2004) used Bayesian methods for combining climate forecasts from three different climate models, with the aim of producing better categorical climate forecasts. They concluded that the skill of the multimodel combined forecast was significantly better than the climatological forecast in only a few regions of the globe. However, the skill of the Bayesian combined forecast was found to be better than the skill of individual model forecasts and also better than the simple average of forecasts from different models.

Other studies have linearly combined deterministic forecasts. Thompson (1977) was the first to show that a simple linear combination of two independent 24 hour weather predictions with optimal weights, obtained when the mean square error of the combined forecast was minimised, could reduce the forecast error variance by about 20%. Fraedrich and Leslie (1987) also noted that by linearly combining stochastic short-range forecasts with numerical model weather predictions it was possible to obtain significantly better prediction skill. Fraedrich and Smith (1989) then extended this approach to the seasonal time scale, with a forecast lead-time of up to three months. They linearly combined an empirical forecast with a deterministic model forecast for predicting tropical Pacific SST anomalies. It was shown that by minimising the combined forecast mean square error (MSE) considerable improvement in MSE skill can be obtained. More recently, Metzger *et al.* (2004) extended the Fraedrich and Smith (1989) combination scheme to predict Niño-3 index anomalies for lead times up to 24 months. It was found that the linear combination of empirical and deterministic forecasts provided improvement in prediction skill if the predictions of individual schemes were independent and of comparable skill. However, only modest skill improvements were found in practice.

Krishnamurti *et al.* (1999, 2000a,b, 2001) and Stefanova and Krishnamurti (2002) have recently introduced the multi-model super-ensemble method for combining numerical weather and climate forecasts. This method, which linearly combines ensemble forecasts from different models by minimising the mean square error of the combined forecast, has also been used by Pavan and Doblas-Reyes (2000). It has been demonstrated that the multi-model super-ensemble invariably performs better than any independent model. A simplified version of the multi-model super-ensemble method has been proposed by Ziehmann (2000) for medium-range weather forecasts. In this method deterministic ensemble forecasts from different models are assigned exactly the same weight in the combination procedure. In other words, the forecast is given by the simple multi-model ensemble mean. Mylne *et al.* (2002) refers to this method as the “poor man’s ensemble”. Ziehmann (2000) has shown that the “poor man’s ensemble” performs better than the ensemble of an individual model.

Kharin and Zwiers (2002) summarise several methods commonly used in atmospheric science for combining forecasts from different models. These methods are based on the standard multiple linear regression method with the regression coefficients being subject to some constraints. They are unbiased since the mean forecast biases are removed by the regression. Forecast biases are well-known features of climate models (e.g. Stockdale 1997) and their removal is essential for any statistical forecast improvement scheme. According to Kharin and Zwiers (2002) the three most often used unbiased combination methods are as follows:

a) The bias-removed multimodel ensemble mean forecast.

The combined forecast is produced by taking the difference between the multi-model ensemble mean and the multi-model mean bias. The multi-model ensemble mean is the mean of a collection of forecasts, each of which is produced with a different model. The multi-model mean bias is the (past) historical mean forecast error, estimated with all forecast that compose the multi-model. This method is equivalent to the simple average method used in economics.

b) The regression-improved multimodel ensemble mean forecast.

The combined forecast is obtained by linearly regressing the multimodel ensemble mean against the observations.

c) The regression-improved multimodel forecast.

The combined forecast is produced by the OLS regression of different model forecasts with no constraints on the regression coefficients. This method is equivalent to the regression method used in economics.

More recently Goddard *et al.* (2003) introduced the multi-model ensembling approach for refining (calibrating) and combining probability forecasts from different atmospheric general circulation models. They found that improved forecast reliability can be achieved by calibrating the model probabilities, estimated from an ensemble of forecasts, prior to combining. The calibration is based on model history. Using ensemble means, conditional probabilities are determined from past model performance. After calibration the combination of forecasts is performed using two methods:

d) Pooled Multi-Model ensemble.

Each model is weighted equally. This is again an equivalent method to the simple average used in economy.

e) Bayesian Multi-Model ensemble.

Optimal weights are determined for each model relative to climatology by maximizing a likelihood measure. This Bayesian method is described in the paper by Rajagopalan *et al.* (2002).

Finally, as described by Goddard *et al.* (2001), climate forecasts can still be combined subjectively using a method known as

f) Consensus forecasting.

Consensus forecasting consists in subjectively blending forecasts for adjacent areas and combining different forecast information for the same regions. Subjective forecast combination has become an important area of development with the advent of regional fora (e.g. Drought Monitoring Centre 1998), and is used in the construction of the International Research Institute for Climate Prediction (IRI) “net assessments” (Mason *et al.* 1999), and the National Centers for Environmental Prediction (NCEP) seasonal forecasts (van den Dool *et al.* 1998). Consensus forecasts are probabilistic, with probabilities being usually attributed to three categories (below normal, normal and above normal). Although the subjective process can be improved by implementing objective combination techniques such as the one proposed in chapter 3 of this thesis, the consensus forecast has some practical value. It permits a simple combination of all available forecasts when full or compatible verification data is unavailable. The risk is that forecasts with minimal or no skill can affect consensus adversely, but the approach is supportable on the basis that the simple average of forecasts is often an improvement on any single forecast product.

2.5 Summary

This chapter has presented an overview on seasonal forecasting, a brief description of the DEMETER project, and a literature review on forecast calibration and combination focussed in economics and atmospheric sciences. The main findings are that:

- Forecasts need calibration against observations to correct systematic errors in model predictions

- Forecasts can be combined in many different ways
- Combined forecasts are often more skillful than individual forecasts
- There is an overlap of methods used in economics and atmospheric sciences
- The multiple linear regression is widely used in both areas
- Although not as popular as multiple linear regression, another well-established method is the Bayesian combination of forecasts. As revealed by Clemen (1989) this method has been widely used in statistics and economics since the mid-1960s, when the paper by Geisser (1965) was published. In atmospheric sciences, the Bayesian approach for combining forecasts was introduced in the mid-1980s by Epstein (1985) and Clemen (1985). Nevertheless, it has not been much explored after that until recent papers by Berliner *et al.* (2000a), Berliner *et al.* (2000b) and Rajagopalan *et al.* (2002).

Chapter 3

Methodology

3.1 Aim

The aim of this chapter is to provide the technical knowledge necessary for the implementation of the Bayesian method, which is used in chapters 4, 5 and 6 of this thesis for the calibration and combination of seasonal climate predictions of an ENSO index (Niño-3.4), Equatorial Pacific SST and South American rainfall.

3.2 A unified framework for forecasting

3.2.1 The forecasting process

Figure 3.1 shows a highly simplified (low-dimensional) schematic of the forecasting process. It is important to recognise that observable variables (e.g. temperature at a particular location) are not the same mathematical quantities as model grid point variables. The state vector of the real atmosphere moves dynamically around q-dimensional *observation state space* whereas the model state vector moves around p-dimensional *model state space*. To initialise models with information from observations, observations in observation state space have to be mapped into model state space using a procedure known as *data assimilation* (Daley 1991; Courtier 1997; Bouttier and Courtier 1999). A set of numerical model predictions can then be

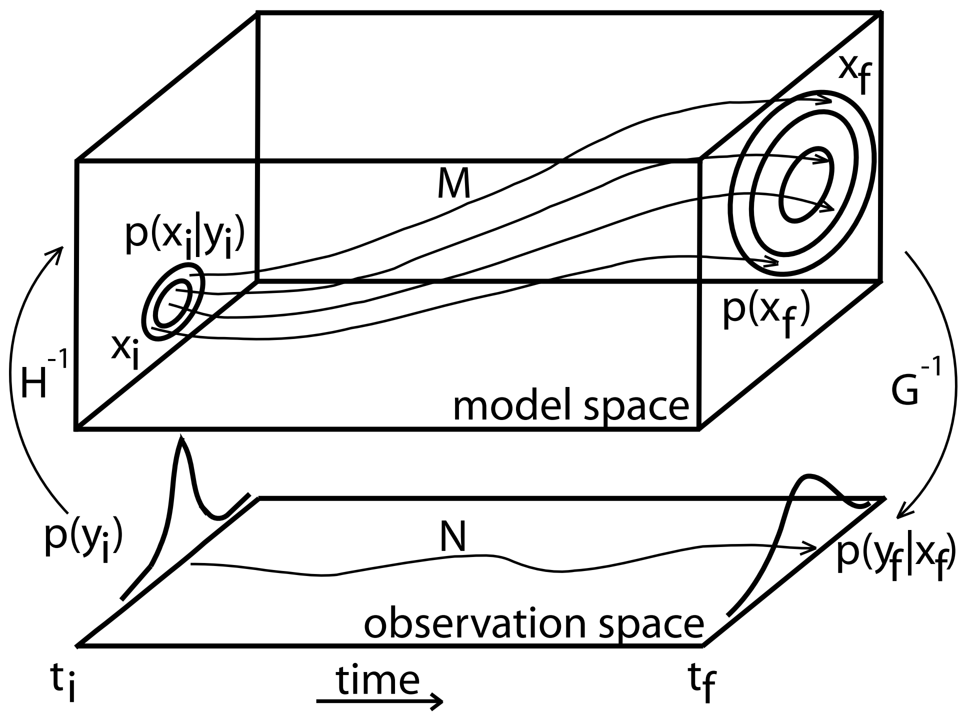


Figure 3.1: Schematic showing the forecasting process. Time t_i is the initial time and time t_f is the final forecast target time. The evolution operator (N) in observation space is not known and so numerical forecasting approximates it by mapping observations into model space (H), evolving model states in time in model space via the model operator (M), and then mapping model predictions back into observation space (G). Environmental forecasting is particularly challenging because of the complexity and high dimensionality of the model and observation spaces.

made to produce an ensemble of possible future model states – a procedure known as *ensemble prediction* (Molteni *et al.* 1996; Stephenson and Doblas-Reyes 2000; Palmer 2002; and references therein).

It is often naïvely assumed that ensembles of model predictions are probability forecasts of the real world. Model variables are generally neither representative nor unbiased estimates of site-specific observable variables. Instead, model predictions should be considered as proxy information that can be used to *infer* the probability of future observables (Wilks 2000; Glahn 2004). The skill of forecasts depends on their ability to discriminate between observable outcomes (known as *forecast resolution*; Jolliffe and Stephenson 2003) rather than their ability to match identically observations. For example, temperature forecasts that distinguish between hot and cold days but that are always 2°C too warm are more skillful than unbiased forecasts that can not distinguish between hot and cold days. To make inferences, one needs a probability model that can give the probability of observable quantities when provided with model forecast data (e.g. regression). There needs to be a procedure for mapping the model predicted state back into observation space. To recognise its analogous role to data assimilation, this important final step will be referred to here as *forecast assimilation*. Forecast assimilation incorporates a wide diversity of previous methods such as bias-correction, statistical down-scaling, model output statistics, perfect prognosis, etc. (Wilks 1995). As apparent in Fig. 3.1, there is a strong analogy/duality between data assimilation and forecast assimilation, which will be elaborated mathematically in the following sections of this chapter.

To summarise, three important steps are needed in order to find the probability density function $p(y_f|y_i)$ of a future observable variable y_f: data assimilation to find $p(x_i|y_i)$, model ensemble prediction to find $p(x_f|x_i)$, and forecast assimilation to find $p(y_f|x_f)$. The standard statistical symbol '|' denotes "given" (conditional upon). The desired probability density $p(y_f|y_i)$ is obtained by integrating over model states using a Monte Carlo approximation (a carefully chosen ensemble). For this to be a good approximation, the initial ensemble states should be sampled from the distribution $p(x_i|y_i)$ – a condition not always satisfied in the design of operational ensemble systems (Stephenson and Doblas-Reyes, 2000).

3.2.2 The analogy with data assimilation

Whereas data assimilation is concerned with how best to estimate the probability density function of model state x_i given observational data y_i, the dual problem of forecast assimilation is concerned with how best to estimate the probability density function of a future observable y_f given model prediction data x_i. Both these activities involve the estimation of conditional probabilities – $p(x_i|y_i)$ for data assimilation and $p(y_f|x_f)$ for forecast assimilation. The resulting distributions are conditioned on the available data such as observations y_i for data assimilation and model predictions x_i for forecast assimilation. The identity known as Bayes' theorem shows how to obtain these conditional probabilities from the unconditional probability distributions $p(x_i)$ and $p(y_f)$ (Lee 1997; Gelman *et al.* 1995). For example, data assimilation uses Bayes' theorem in the form

$$p(x_i|y_i) = \frac{p(y_i|x_i)p(x_i)}{p(y_i)} \tag{3.1}$$

to update the background distribution $p(x_i)$ to obtain the conditional distribution $p(x_i|y_i)$ (Lorenc 1986; Bouttier and Courtier 1999). In other words, data assimilation consists of updating the initial state (first guess) distribution from $p(x_i)$ to $p(x_i|y_i)$ when new observational data y_i become available. Similarly, forecast assimilation uses Bayes' theorem in the form

$$p(y_f|x_f) = \frac{p(x_f|y_f)p(y_f)}{p(x_f)} \tag{3.2}$$

to update the climatological distribution $p(y_f)$ to obtain the conditional distribution $p(y_f|x_f)$. Initial and final time subscripts are suppressed in subsequent equations. For perfect models with no prediction errors or for forecasts with very short lead times, the data assimilation equations can be used to perform forecast assimilation (i.e. perfect prognosis). However, for all other model predictions containing errors then forecast assimilation is required in addition to data assimilation. From this it can be seen that the true role of ensemble model predictions is to provide data x that can be used to update the probability distribution of the observable variable from $p(y)$ to $p(y|x)$ rather than to provide an estimate of the marginal distribution $p(x)$.

3.3 The Univariate Normal Model

This section introduces the Bayesian method for the calibration and combination of single point in space (index) forecasts that is used in chapter 4. The Bayesian approach has been discussed for decision making in applied meteorology by Epstein (1962) and for statistical inference and prediction in climatology by Epstein (1985).

It has also successfully been used in other areas such as hydrology (e.g. Krzysztofowicz 1983; Krzysztofowicz and Herr 2001) and recently in climate studies (e.g. Berliner *et al.* 2000a,b and Rajagopalan *et al.* 2002). The Bayesian method is a consistent probabilistic approach that can be used for combining historical (climatological) information y with physically-derived dynamical model ensemble mean forecasts $\bar{x}$. It is based on rigorous probability theory and so can provide well-calibrated probability forecasts.

With no access to a physically-derived coupled model ensemble mean forecast $\bar{x}$, the only possible probabilistic assessment about the observable variable y has to be based on the assumption that future values of y will behave like they did in the past. For example, the probability distribution of y can be estimated by using the climatological probability density function $p(y)$ estimated from historical observations. In Bayesian theory $p(y)$ is known as the *prior distribution* and encapsulates *prior knowledge* about likely possible values of y – from past experience not all values of y were found to occur equally likely. More informative prior distributions than the climatological distribution, derived for example from empirical models can also be defined.

However, when a particular ensemble mean forecast $\bar{x} = x_m$ (obtained as the mean of m members of an ensemble forecast) is known for the future, it is then possible to update the prior $p(y)$ to obtain the conditional *posterior distribution* $p(y|\bar{x} = x_m)$. In other words, this is the probability distribution of y given that the forecast $\bar{x} = x_m$ is known. Conditioning on forecasts helps to reduce the uncertainty about future values of y (Jolliffe and Stephenson, 2003; their chapter 9). This procedure is illustrated schematically in Fig. 3.2 for normal p.d.f.'s. The normal prior probability density (short-dashed line) when combined with a normal likelihood probability density (dashed line) yields a normal posterior probability density (solid line). The posterior distribution $p(y|\bar{x} = x_m)$ is found from the prior $p(y)$ by making use of Bayes' theorem in the form

$$\overbrace{p(y_t|\bar{x}_t = x_m)}^{posterior} = \frac{\overbrace{p(\bar{x}_t = x_m|y_t)}^{likelihood}\,\overbrace{p(y_t)}^{prior}}{p(\bar{x}_t = x_m)} \tag{3.3}$$

where y_t is the observable variable at time *t* and x_m is a particular value of ensemble mean forecast at time *t*. Note that both the posterior distribution and the likelihood function are considered to be functions of y_t. Finally, $p(\bar{x}_t = x_m)$ does not depend on y_t and therefore only plays the role of a normalising constant (Lee 1997).

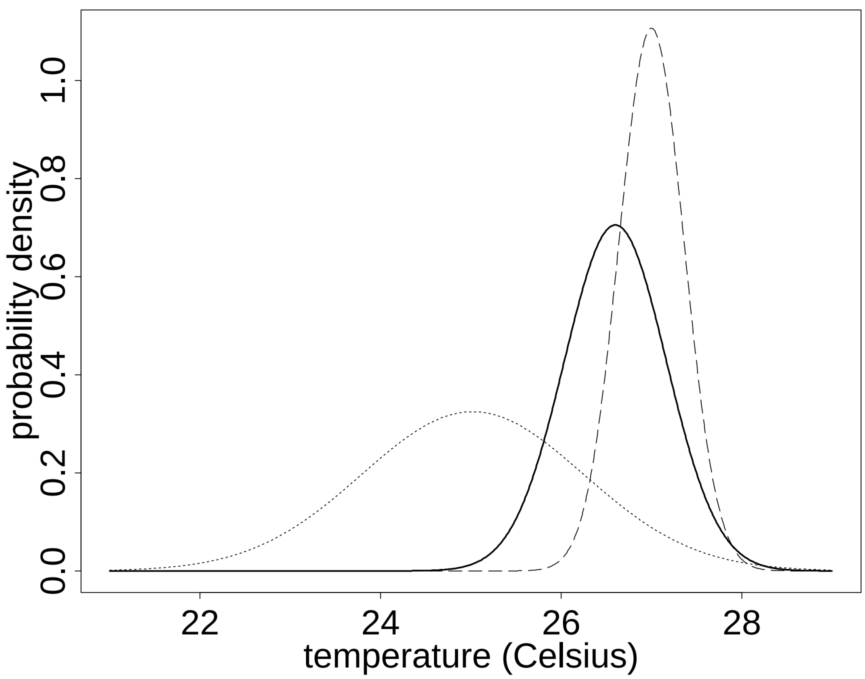


Figure 3.2: Prior distribution (short-dashed line), likelihood (dashed line) and posterior distribution (solid line). Source: Coelho *et al.* 2004.

The likelihood $p(\bar{x}|y)$ of obtaining an ensemble mean forecast $\bar{x}$ given observations y is an essential ingredient in the Bayesian updating procedure that can be estimated by stratifying past ensemble mean forecasts (hindcasts) on past observations. The likelihood provides a convenient summary of the calibration and resolution of past forecasts (Jolliffe and Stephenson 2003).

This Bayesian approach has several important advantages over approaches that rely *solely* on sampling ensembles of physically-derived coupled model forecasts (e.g. Stockdale *et al.* 1998 and Taylor and Buizza 2003). Firstly, the Bayesian approach appropriately incorporates prior information about the distribution contained in historical observations (i.e., it allows combination). Secondly, the likelihood provides a natural way of correcting for biases in the model forecasts that often occur in physically-derived coupled models (i.e., it allows calibration). Thirdly, the resulting well-calibrated posterior distribution allows one to then generate an arbitrarily large sample (a *mega-ensemble*) of possible climate realizations, of use for example in scenario studies of risk and forecast value (Jolliffe and Stephenson 2003; their chapter 8).

The Bayesian method has three main steps: a) choice of the prior distribution; b) modelling of the likelihood function; and c) determination of the posterior distribution.

a) Choice of the prior distribution.

For simplicity, it has been assumed that the prior distribution is normal (Gaussian)

$$y_t \sim N(\mu_{ot},\, {\sigma_{ot}}^2) \tag{3.4}$$

where μ_{ot} is the mean and ${\sigma_{ot}}^2$ is the variance of a normal (N) probability distribution. The standard statistical symbol $\sim$ denotes "is distributed as". The mean μ_{ot} and the variance ${\sigma_{ot}}^2$ can be estimated using for example climatological (historical) data of y. More sophisticated empirical models can be used to estimate these two parameters (see an example in section 4.2.2 of chapter 4). When no information about y is available one can assume a uniform distribution with infinite ${\sigma_{ot}}^2$ for the prior.

b) Modelling of the likelihood function.

As for the prior distribution, for simplicity, it has been assumed that the likelihood distribution is normal (Gaussian). Figure 3.3 illustrates how the likelihood is modelled. It shows a scatter plot of raw (uncorrected) ECMWF coupled model ensemble forecasts versus the observed December Niño-3.4 index for the period 1987-1999. Ensemble means x_m of $m = 9$ ensemble members are depicted using large open circles. The dashed line is what one expects for perfect forecasts in which the forecast values are identical to the observed values. The likelihood $p(\bar{x}_t \mid y_t)$ is modelled by performing a linear regression of the ensemble mean forecasts ($\bar{x}_t$) on matching observations (y_t):

$$\bar{x}_t \mid y_t \sim N(\alpha + \beta y_t, \delta) \tag{3.5}$$

where α and β are the intercept and slope parameters, respectively. The constant variance δ is estimated by the mean of the squared regression residuals

$$\hat{\delta} = \frac{1}{n-2} \sum_{t=1}^{n} (\bar{x}_t - \hat{\alpha} - \hat{\beta} y_t)^2 \tag{3.6}$$

where *n* is the length of the time series and the hat symbol denotes an estimated parameter.

The solid line in Fig. 3.3 is the best fit linear regression between raw ensemble mean values $\bar{x}_t$ and observations y_t, corresponding to estimates of $\hat{\alpha}$ = 6.65^oC, $\hat{\beta}$ = 0.73, and $\hat{\delta}$ = 0.09$^oC^2$. This fitted line is the likelihoood function used for the calibration of the forecasts. Note, however, that rather than regress the forecasts x on the observations y one might perform the *inverse regression* of the observation y on the forecasts x in order to calibrate the forecasts. This inverse regression is discussed in Appendix A.

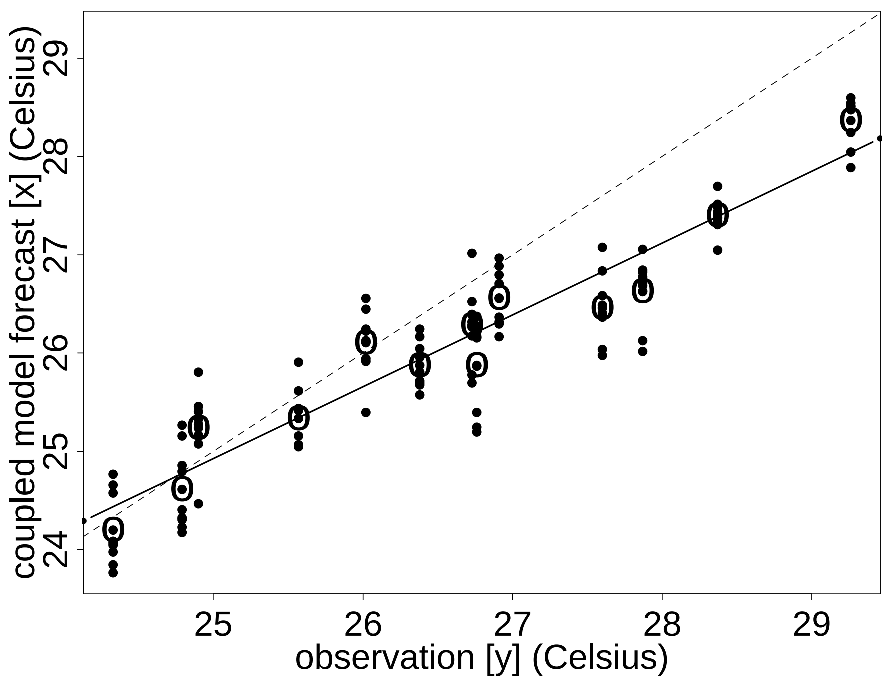


Figure 3.3: December Niño-3.4 index likelihood model assuming constant variance δ for the ensemble mean ($\bar{x}$) as formulated in Eqn. 3.5. Parameter estimates are $\hat{\alpha} = 6.65^oC$, $\hat{\beta} = 0.73$ and $\hat{\delta} = 0.09^oC^2$ ($R^2 = 0.93$). Each black dot is one of the $m = 9$ ensemble members. Big open circles are ensemble means $\bar{x} = x_m$. The solid line is the linear regression fit between ensemble means $\bar{x} = x_m$ and observations y. The dashed line is what would be obtained for perfect forecasts. Adapted from Coelho *et al.* 2003.

c) Determination of the posterior distribution

From Bayes' theorem (Eqn. 3.3) it can be shown that for a normal prior distribution $y_t \sim N(\mu_{ot}, \sigma^2_{ot})$ and normal likelihood $\bar{x}_t | y_t \sim N(\alpha + \beta\, y_t, \delta)$, the posterior distribution is also normal (Lee 1997). The resulting normal posterior distribution is given by

$$y_t \mid \bar{x}_t \;\sim\; N(\mu_t, \sigma^2_t) \tag{3.7}$$

with the mean μ_t and the variance σ^2_t equal to

$$\frac{1}{\sigma^2_t} = \frac{1}{{\sigma_{ot}}^2} + \frac{\beta^2}{\delta} \tag{3.8}$$

$$\frac{\mu_t}{\sigma^2_t} = \frac{\mu_{ot}}{{\sigma_{ot}}^2} + \frac{\beta^2}{\delta}\left(\frac{\bar{x}_t - \alpha}{\beta}\right). \tag{3.9}$$

A derivation of Eqns. (3.8) and (3.9) is presented in Appendix B. The inverse of the variance is known in statistics as the *precision*. Equation (3.8) states that the precision of the posterior distribution $\left(\frac{1}{\sigma^2_t}\right)$ is exactly equal to the precision of the prior distribution $\left(\frac{1}{{\sigma_{ot}}^2}\right)$ plus the precision of the ensemble system $\left(\frac{\beta^2}{\delta}\right)$. Unbiased forecasts would have precision $\frac{1}{\delta}$. However, forecasts are not unbiased and so the precision is instead given by the term $\frac{\beta^2}{\delta}$.

Equation (3.9) gives the posterior combined mean (μ_t) as the precision weighted sum of the prior mean (μ_{ot}) and the physically-derived raw (uncorrected) coupled model ensemble mean $(\bar{x}_t)$. Note that the precision of the prior distribution and the precision of the ensemble system are weights for the prior mean and uncorrected ensemble mean, respectively. The mean bias of the ensemble system is corrected by the difference between $\bar{x}_t$ and α divided by the re-scaling factor β.

In practice, in order to avoid artificial skill, all estimated parameters are obtained here using a cross-validation "leave one out" method (section 6.3.6, Wilks 1995). To produce a forecast for time *t*, only data at other times (years) different from *t* are used to estimate model parameters and errors.

3.4 The Multivariate Normal Model

3.4.1 Data assimilation and forecast assimilation equations

This section presents the equations of the Bayesian multivariate model that is used in chapters 5 and 6 for the calibration and combination of one-dimensional and two-dimensional forecasts. This model is referred here to as forecast assimilation. Both variational data assimilation and forecast assimilation are based on multivariate normal models. Given the analogy/duality between them, as described in section 3.2.2 of this chapter, it is useful to start by briefly reviewing the key equations in data assimilation. As noted in the previous sections of this chapter, the practical implementation of Bayes' theorem requires the specification of a suitable probability model. For the sake of generality, consider a p-dimensional model space and a q-dimensional observation space where $p, q \geq 1$. The least-squares estimation used in variational assimilation is equivalent to maximum likelihood estimation when the $p \times 1$ model state x and the $q \times 1$ observations y given a model state x are (multivariate) normally distributed:

$$x = x_b + \epsilon_B \qquad (3.10)$$

$$y = Hx + \epsilon_R \qquad (3.11)$$

where the $p \times 1$ vector x_b is the background model state (the first guess) and ϵ_B and ϵ_R are (multivariate) normally distributed errors with zero mean and covariances B and R, respectively. The matrix H is known as the *observation* or *interpolation operator* that predicts observables (e.g. satellite-measurable radiances) from model states (e.g. vertical temperature profiles). These equations can be written more informatively in probability notation as follows:

$$x \sim N(x_b, B) \qquad (3.12)$$

$$y|x \sim N(Hx, R) \qquad (3.13)$$

where $\sim N(\mu, \Sigma)$ means distributed as a multivariate normal distribution with mean μ and covariance Σ. By use of Bayes' theorem (Eqn. 3.1), it can then be shown (e.g. Section 4, Bouttier and Courtier 1999) that these equations lead to

$$x|y \quad \sim \quad N(x_a, A) \tag{3.14}$$

with the analysis model state x_a and the analysis error covariance A given by

$$x_a \quad = \quad x_b + K(y - Hx_b) \tag{3.15}$$

$$A \quad = \quad (I - KH)B. \tag{3.16}$$

The $(p \times q)$ matrix $K = BH^T(HBH^T + R)^{-1}$ is known as the gain/weight matrix. Therefore, observation data y can be used to update the background model state x_b to give an improved analysis estimate x_a provided one can estimate matrices B, R, and H. The analysis state x_a is the *maximum a posteriori* (MAP) estimate (Robert 2001) that maximises the probability density $p(x|y)$. In other words, the MAP estimate x_a can be found by minimising $-2\log p(x|y)$ which is given up to a constant by

$$J_{x|y} = (x - x_b)^T B^{-1}(x - x_b) + (y - Hx)^T R^{-1}(y - Hx). \tag{3.17}$$

The quantity $J_{x|y}$ is known in variational (e.g. 3-d VAR) data assimilation as the *cost function* and various sparse matrix algorithms can be used to find the value of x that minimises this function.

The equations for forecast assimilation of normally distributed predictions are the dual of those for data assimilation with x and y interchanged. One assumes that the observable state and the model predictions given an observable state are (multivariate) normally distributed:

$$y \quad = \quad y_b + \epsilon_C \tag{3.18}$$

$$x \quad = \quad G(y - y_0) + \epsilon_S \tag{3.19}$$

where y_b is the *background observable state* (e.g. the climatological mean value or a persistence forecast) and ϵ_C and ϵ_S are (multivariate) normally distributed errors with zero mean and *background observable covariance* C and *forecast error covariance* S, respectively. For generality, a bias term y_0 has been included to take account of the mean bias often found in model predictions [1]. The $(p \times q)$ matrix G is the *forecast operator* (or *likelihood*) that can be estimated by regression of the model predictions on the observed values. The equations can be rewritten more succinctly as the following probability models:

$$\begin{aligned} y &\sim N(y_b, C) && (3.20)\\ x|y &\sim N(G(y - y_0), S). && (3.21)\end{aligned}$$

Then Bayes' theorem (Eqn. 3.2) can be used to show that

$$y|x \sim N(y_a, D) \qquad (3.22)$$

with the *forecast observable state* y_a and the *forecast error covariance* D given by

$$\begin{aligned} y_a &= y_b + L(x - G(y_b - y_0)) && (3.23)\\ D &= (I - LG)C = (G^T S^{-1} G + C^{-1})^{-1}. && (3.24)\end{aligned}$$

The $(q \times p)$ matrix $L = CG^T(GCG^T + S)^{-1}$ is the forecast gain/weight matrix. Therefore, model prediction data x can be used to update the background observable state y_b (e.g. the climatological mean) to give an improved forecast of the observable y_a provided one can estimate matrices C, S, and G and bias vector y_0. The forecast observable state y_a is the MAP estimate that maximises the probability

[1] a bias term is not required for variational data assimilation because the non-linear observation operator $\mathcal{H}(x)$ is linearised about $x = x_b$ to obtain the unbiased equation $y' = Hx' + O(x'^2)$ where $y' = y - \mathcal{H}(x_b)$ and $x' = x - x_b$.

$p(y|x)$ or alternatively minimises $-2\log p(y|x)$ given up to a constant by

$$\begin{aligned} J_{y|x} &= (y-y_b)^T C^{-1}(y-y_b) \\ &+ (x-G(y-y_0))^T S^{-1}(x-G(y-y_0)). \end{aligned} \tag{3.25}$$

The quantity $J_{y|x}$ is the cost function that needs to be minimised for forecast assimilation. It is the sum of two penalty terms: one that penalises departures $y-y_b$ from the background observable state and one that penalises departures $x-G(y-y_0)$ from calibrated model predictions.

3.4.2 Estimation and the need for dimension reduction

For multivariate normal forecast assimilation one needs estimates of vectors y_b and y_0 and matrices C, S, and G. The vector y_b and matrix C are parameters of the prior observable distribution $y \sim N(y_b, C)$. Reliable estimates of these parameters can be most simply obtained by calculating the climatological mean and sample covariance of past observations. More precise estimates of the prior can be obtained using empirical forecasts (if available) and so empirical forecasts can be elegantly merged with numerical model predictions (Coelho *et al.* 2004). The other parameters are estimated by performing a multivariate linear regression of the model predictions on the observations for a period when both predictions and observations are available (the calibration period). For model predictions in chapter 5 the seven ensemble mean forecasts obtained from the seven DEMETER models are used rather than 63 (7×9) individual forecasts. In chapter 6 three ensemble mean forecasts from ECMWF, CNRM and UKMO models are used. The slope G, bias vector y_0, and the prediction error covariance S can be estimated using ordinary least squares estimation:

$$G = S_{xy}S_{yy}^{-1} \tag{3.26}$$

$$y_0 = -(\overline{x} - \overline{y}G^T)G(G^TG)^{-1} \tag{3.27}$$

$$S = S_{xx} - S_{xy}S_{yy}^{-1}S_{xy}^T \tag{3.28}$$

where S_{xx} is the $(p \times p)$ covariance matrix of the model predictions, S_{yy} is the $(q \times q)$ covariance matrix of the observables, and S_{xy} is the $(p \times q)$ cross-covariance matrix.

Calibration of gridded forecasts is particularly difficult because of the large dimensionality of gridded data sets compared to the number of independent forecasts and the strong dependency between values at neighbouring grid points. When the matrix S_{yy} is poorly conditioned (or even rank deficient, e.g. when $q > n$, where *n* is the length of the time series that is used for calibration) then the estimation of $G = S_{xy}S_{yy}^{-1}$ becomes problematic (or impossible). Furthermore, for L to be well estimated then $GCG^T + S$ also has to be well-conditioned. Put simply, one cannot simultaneously calibrate many predictions if one has only a small historical record of calibration data. This problem becomes even worse for multi-model predictions where the number of grid points is multiplied by the number of models (e.g. in the example of chapter 5, $p = 7 \times 56 = 392$). To avoid this problem one can use various multivariate dimension reduction techniques to reduce the dimensionality of the data sets. Instead of considering grid point variables, one can project the data onto a small set of spatial patterns to obtain a small number of indices. For example, one could perform principal component regression by using the leading principal components of the model predictions and the observations (Derome *et al.* 2001; Jolliffe 2002, his section 8.1). Alternatively, one can use either maximum covariance analysis (MCA) [sometimes referred to as SVD] or Canonical Correlation Analysis (CCA) to extract leading co-varying modes from the model prediction and observation data (von Storch and Zwiers 1999). An MCA based regression approach has been used in previous studies to improve single model seasonal forecasts (Feddersen *et al.* 1999, Feddersen 2003). In the examples of equatorial Pacific SSTs and South American rainfall of chapters 5 and 6 both MCA and CCA dimension reduction approaches with up to 8 retained modes have been tested. It has been found that MCA with 3 modes gave the best cross-validated forecast results, which are shown in these two chapters.

In practice, forecast assimilation is performed as follows:

1. In order to produce cross-validated forecasts on data not used in the estima-

tion, the year to be forecast is removed from the data set.

2. The time mean is subtracted from the remaining observations and the model predictions to make anomalies stored in a $(n \times q)$ data matrix Y of observations and a $(n \times p)$ data matrix X of model predictions (in the example of chapter 5 $n = 21$ $q = 56$ and $p = 392$).

3. An SVD analysis is performed of the matrix $Y^T X = U \Sigma^* V^T$ to determine the leading MCA modes.

4. In order to estimate the prior distribution, the background observable covariance matrix C is calculated for the $k = 3$ leading MCA modes of the observations. The mean of the prior distribution $y_b = 0$ since we are treating anomalies about the long-term climatological mean.

5. A multivariate regression of the k-leading MCA model prediction modes on the k-leading MCA observation modes is performed in order to estimate G, y_0, and S.

6. The estimated quantities C, y_b, G, y_0, and S are then used to forecast the observations of the removed year using model predictions from that year.

3.4.3 The special case of no additional prior information

In the special case when the prior is estimated over the same calibration period used to estimate the prediction error covariance S (i.e. no extra information is used to estimate y_b or C) then forecast assimilation gives the same forecasts as obtained by ordinary least squares multivariate regression of the observations on the predictions

$$y \;=\; \overline{y} + S_{yx} S_{xx}^{-1} (x - \overline{x}) \tag{3.29}$$

The proof of this can easily be derived by noting that $S = S_{xx} - G S_{yy} G^T$ and so when $C = S_{yy}$ one obtains $L = S_{yy} G^T (G S_{yy} G^T + S_{xx} - G S_{yy} G^T)^{-1}$ which simply becomes $L = S_{yx}S_{xx}^{-1}$. Multivariate regression of observations on predictions is the basis of the MOS approach to calibration that has been used in several previous multi-model studies (Pavan and Doblas-Reyes 2000; Kharin and Zwiers 2002; Doblas-Reyes *et al.* 2005). Forecast assimilation incorporates this approach as a special case when the prior is estimated using only the calibration data. An important difference in forecast assimilation is that one models the likelihood by regression of the model predictions on the observables rather than directly modelling the observables as functions of model predictions. The likelihood regression in forecast assimilation minimises uncertainty in the model predictions for given observations whereas the MOS approach minimises uncertainty in observations for given model predictions. A discussion about these two different ways of performing the regression (classical and inverse) is presented in Appendix A.

3.5 Summary

This chapter has presented a unified framework for forecasting. This framework has been named forecast assimilation due to its analogy/duality with data assimilation. The chapter still introduced the univariate Bayesian model that is used in chapter 4 for the calibration and combination of empirical and physically-derived coupled model forecasts of Niño-3.4 index forecasts (i.e., forecasts for a single point in space). Additionally, this chapter has generalised the single point in space Bayesian model for the calibration and combination of multivariate forecasts [i.e., one-dimensional forecasts and spatial field two-dimensional forecasts]. The multivariate Bayesian model is used in chapters 5 and 6 for the calibration and combination of multi-model forecasts of equatorial Pacific SSTs and South American rainfall.

Chapter 4

A zero-dimensional example: Niño-3.4 forecasts

4.1 Aim

The aim of this chapter is to test the use of the univariate Bayesian normal model introduced in chapter 3 (section 3.3) for the calibration and combination of empirical and coupled model Niño-3.4 index forecasts for December at a 5-month lead time. This method merges valuable past (historical) information with physically-derived dynamical coupled model ensemble forecasts to produce well-calibrated estimates of the mean forecast value and its respective uncertainty.

4.2 Introduction

This chapter examines forecasts of an ENSO index (Niño-3.4), which is the mean SST inside the area delimited by the parallels of $5^o N$ and $5^o S$ and the meridians of $120^o W$ and $170^o W$ in the equatorial Pacific. ENSO is an important large-scale ocean-atmosphere coupled phenomenon that has large impacts on the climate of many regions around the world (Horel and Wallace 1981; Stoeckenius 1981; Ropelewski and Halpert 1986, 1987 and 1989). Since the strong El Niño episode in 1982/83, many efforts have been made to produce routine forecasts of tropical Pacific SSTs. Long-lead forecasts help local governments and industries plan their actions prior to the occurrence of the phenomenon (Patt 2000).

ENSO forecasts are currently produced using either physically-derived dynamical climate models or empirical (statistical) relationships based on historical data. For a comprehensive review of ENSO forecasting studies developed during the last two decades see Mason and Mimmack (2002). As previously discussed in chapter 2 (section 2.2.2), the comparative skill of these two approaches is a subject of much debate in the literature. Given these two distinct approaches to forecasting, it is natural to question the feasibility of combining them in order to produce forecasts more skillful than either forecast considered separately.

As pointed out by Mason and Mimmack (2002), ENSO forecasts are usually issued in deterministic terms and very little attention has been directed to careful estimation of forecast uncertainty. This study treats ENSO forecasts in probabilistic terms, with particular attention directed to the estimation of prediction uncertainty. For this particular application, Niño-3.4 index forecasts are summarized by the mean and the variance of a normal distribution. These two numbers, which fully describe the entire forecast probability distribution, are used to produce interval forecasts.

The next section of this chapter demonstrates the use of the univariate Bayesian method for the calibration and combination of empirical and physically-derived coupled model forecasts. This demonstration uses DEMETER ECMWF 5-month lead December mean Niño-3.4 index forecasts started from conditions at the end of the preceding July. Empirical and physically-derived coupled model ensemble forecasts available over the n=44-year period (1958-2001) are used. Details concerning datasets and forecast lead time are given in Appendix C.

4.3 Empirical forecasts

Figure 4.1 shows the historical (1950-2001) July and December Niño-3.4 index time series obtained from Reynolds optimum interpolation version 2 SST dataset[1] (Reynolds *et al.*, 2002). The two time series are positively correlated ($r = 0.87$), illustrating the importance of persistence in predictability of the Niño-3.4 index. The largest El Niño (1972, 1982 and 1997) and La Niña (1970, 1973, 1988 and 1998) events can clearly be seen.

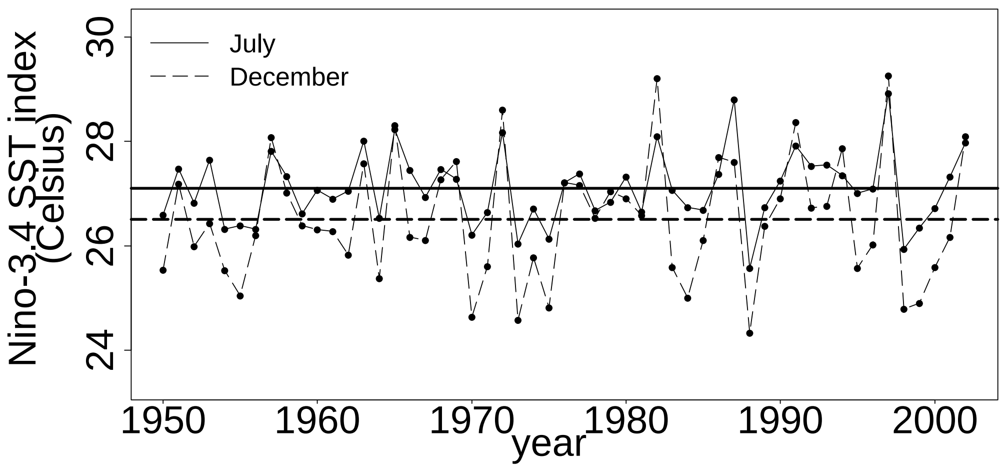


Figure 4.1: Observed July (solid line) and December (dashed line) Niño-3.4 time series (1950-2001) in ^oC. The horizontal thick solid line is the July climatological mean of 27.1^{o}C and the horizontal thick dashed line is the December climatological mean of 26.5 ^{o}C for this period

The simplest 5-month lead empirical model for forecasting December mean Niño-3.4 index uses linear regression with the preceding July mean Niño-3.4 index historical time series as the linear predictor. Mathematically, $y_t = \beta_0 + \beta_1\psi_t + \epsilon'_t$, where y_t and ψ_t are the December and July Niño-3.4 monthly mean values, respectively, β_0 and β_1 are the intercept and slope parameters, respectively, ϵ'_t is a "Normal (Gaussian)" random variable with zero mean and variance σ_o^2 [i.e., $\epsilon'_t \sim N(0, \sigma_o^2)$], and *t* is the year being forecast. This model can be written more explicitly in

[1]http://www.cpc.noaa.gov/data/indices/index.html

probabilistic notation as

$$y_t \,|\, \psi_t \sim N(\mu_{ot},\, {\sigma_o}^2) \tag{4.1}$$

with the mean of y_t given by

$$\mu_{ot} = \beta_0 + \beta_1 \psi_t \tag{4.2}$$

that is, a linear function of the predictor ψ_t. The Niño-3.4 index is known to be well approximated by the normal distribution and so the normal regression model is appropriate (Burgers and Stephenson 1999; Hannachi *et al.* 2003; Hannachi *et al.* 2004).

Figure 4.2 shows a scatter plot of the December versus the preceding July Niño-3.4 index for the period 1950-2001 ($N = 52$ observations). The linear regression fit is indicated on Fig. 4.2 as a solid line. A large amount of the total variance of December values is explained by the preceding July Niño-3.4 index (R^2=0.76). This emphasises the importance of persistence for forecasting the Niño-3.4 index.

To avoid artificial skill, this empirical model has been evaluated using a cross-validation "leave one out" method (Wilks 1995, Section 6.3.6). Figure 4.3a shows empirical forecasts for the target period 1958-2001 (solid line), observed values (dashed line) and the December climatological mean of 26.5^oC (short-dashed line). The 95$\%$ prediction interval (P.I.) for y_t "given" ψ_t is also shown (grey shading). The 95$\%$ prediction interval is defined by

$$\hat{\mu}_{ot} \pm 1.96\, \hat{\sigma}_{ot} \tag{4.3}$$

where $\hat{\mu}_{ot} = \hat{\beta}_0 + \hat{\beta}_1 \psi_t$ is the Niño-3.4 index predicted mean for a particular December and $\hat{\sigma}_{ot}$ is the prediction standard deviation given by

$$\hat{\sigma}_{ot} = \hat{\sigma}_o \left(1 + \frac{1}{n} + \frac{(\psi_t - \overline{\psi}_t)^2}{n\, S_t^2} \right)^{\frac{1}{2}} \tag{4.4}$$

where $n = N - 1$ is the total number of years used in the cross-validation, $\overline{\psi}_t =$

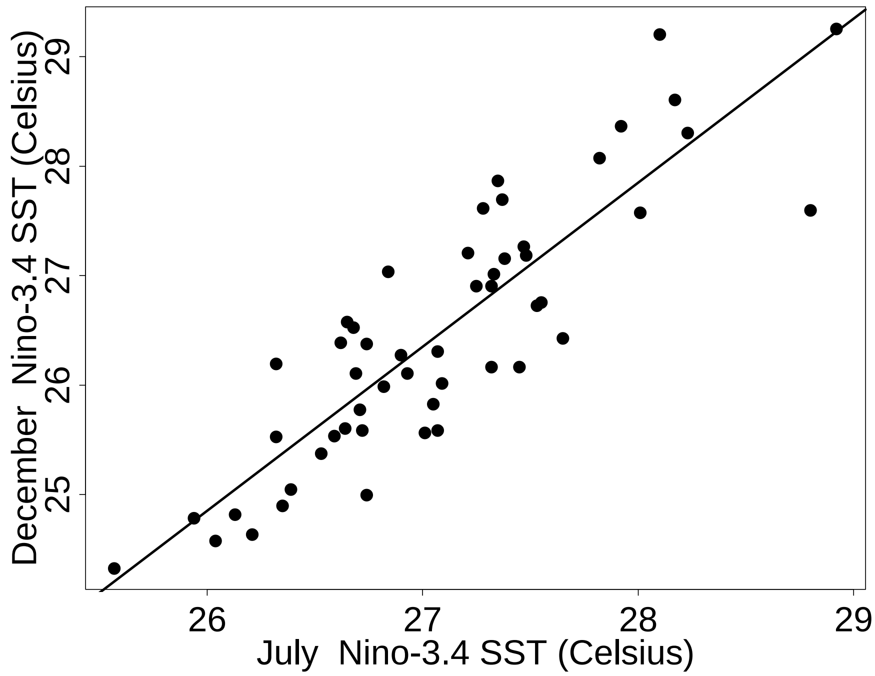


Figure 4.2: Scatter plot of July versus December Niño-3.4 index (^oC). The solid line is the 1950-2001 linear regression model ($\hat{\beta}_0$ = -14.14^oC, $\hat{\beta}_1$ = 1.50, R^2 = 0.76)

$\frac{1}{n}\sum\limits_{i\neq t}\psi_i$ is the long-term climatological mean of the July Niño-3.4 index, $S_t^2 = \frac{1}{n}\sum\limits_{i\neq t}[\psi_i - \overline{\psi}_t]^2$ and $\hat{\sigma}_o = \left[\frac{1}{n-2}\sum\limits_{i\neq t}(y_i - \hat{\mu}_{oi})^2\right]^{\frac{1}{2}}$ is the estimated empirical model standard deviation (see Draper and Smith 1998, Section 3.1).

Eqns. (4.3) and (4.4) show that the smallest prediction interval is obtained when the predictor equals its mean value $\psi_t = \overline{\psi}_t$. On the other hand, by moving away from $\overline{\psi}_t$ in either direction the prediction interval increases. The greater distance a particular July Niño-3.4 index (ψ_t) is from the climatological mean value ($\overline{\psi}_t$), the larger is the extrapolation error made when predicting the following December Niño-3.4 index (y_t). However, the use of Eqn. (4.4) compared to $\hat{\sigma}_{ot} = \hat{\sigma}_o$ leads to only small changes in practice in the prediction interval, since the S_t^2 term in the denominator is proportional to the sum of n terms of the same magnitude as the term $(\psi_t - \bar{\psi}_t)^2$. The most precise predictions are obtained for July Niño-3.4 index values in the "middle" of the observed range of ψ_t, while for more extreme

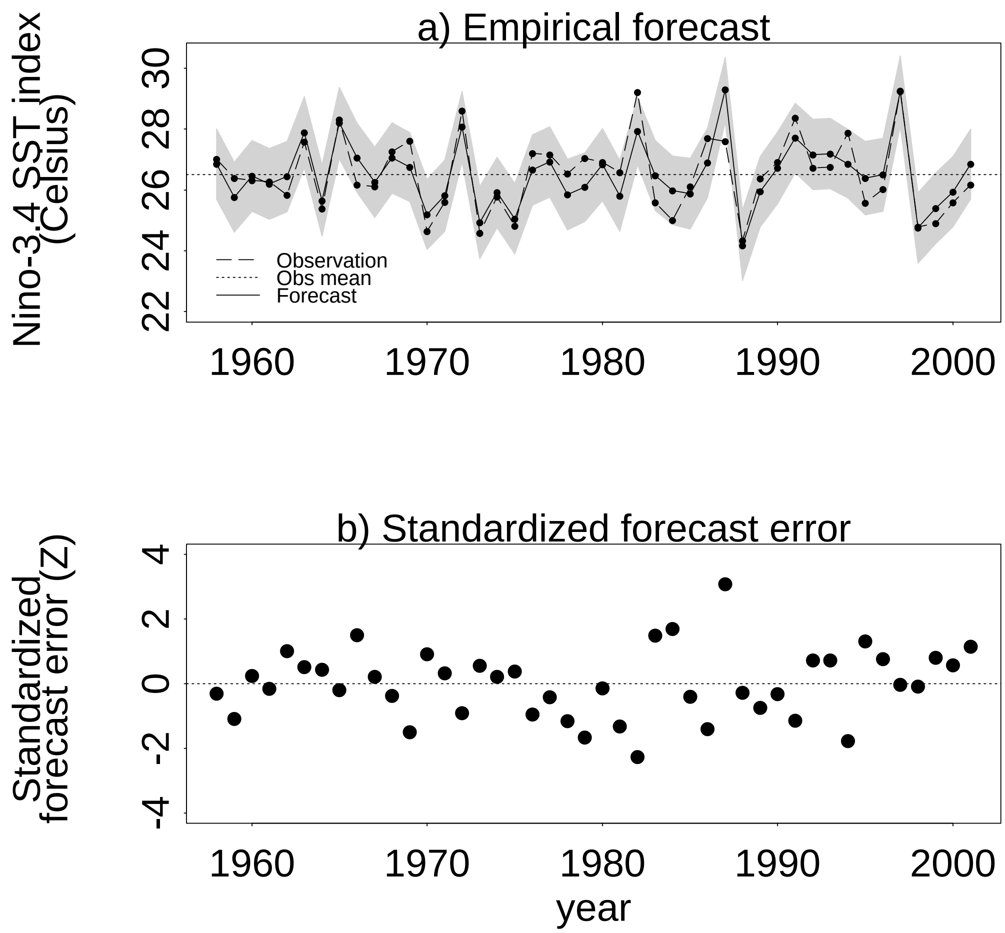


Figure 4.3: a) December 1958-2001 Niño-3.4 index empirical forecasts (^oC). Observed values (dashed line), forecasts (solid line) and the 95$\%$ prediction interval (grey shading). The short-dashed line is the December 1950-2001 climatological mean (26.5 ^oC). b) Standardized forecast error.

values further away from the climatological mean, predictions are less precise.

Figure 4.3a shows that the empirical forecast prediction interval does not vary much from year to year, indicating stability of estimates such as $\hat{\sigma}_o$. This simple model shows surprisingly accurate results, especially for the 1988 and 1998 La Niña episodes and for the 1997 El Niño episode. Within the 44 years of December hindcasts the model has only forecast the Niño-3.4 index outside the 95$\%$ P.I. in 1982 and 1987. This is in agreement with the expected error rate of 5$\%$. Persistence works well for the July to December forecasts that occur after the spring barrier

(Balmaseda *et al.* 1995; van Oldenborgh *et al.* 2003). Measures of forecast skill and uncertainty will be discussed in more detail in section 4.6 of this chapter.

Figure 4.3b shows the time series of the standardized forecast errors

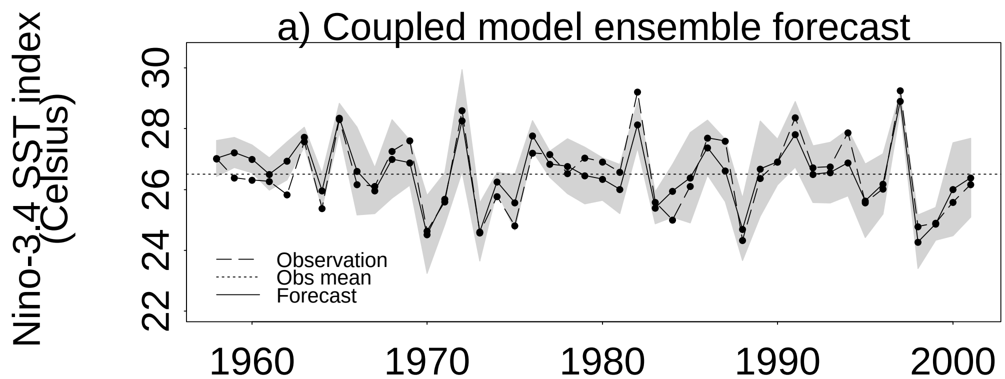
$$Z_t = \frac{\hat{\mu}_{ot} - y_t}{\hat{\sigma}_{ot}} \tag{4.5}$$

where $\hat{\mu}_{ot}$ is the forecasted mean, y_t is the observed value and $\hat{\sigma}_{ot}$ is the prediction standard deviation at time *t*. If this empirical model is appropriate, the standardized forecast errors should be distributed as independent normally distributed random variables with zero mean and unit variance. This appears to be the case from Fig. 4.3b. The standardised forecast errors appear to have constant variance and are well centred on zero with no obvious large outliers. The largest standardized forecast error occurred in 1987.

This empirical model defined by Eqns. (4.1) and (4.2) provides an informative and straightforward prior distribution for the univariate Bayesian procedure described in section 3.3 of chapter 3, and therefore will be used here to estimate the prior (Eqn. 3.4).

4.4 Coupled model ensemble forecasts

Figure 4.4a shows DEMETER ECMWF coupled model ensemble forecasts for the period 1958-2001 (same period as in Fig. 4.3 for the empirical forecasts). The forecasts have been bias corrected by removing the historical mean forecast bias (error) from the raw ensemble mean forecasts. The bias-corrected ensemble mean forecast, which has been obtained from the ensemble of $m = 9$ forecasts, is shown as a solid line. The 95$\%$ P. I., given by the bias-corrected ensemble mean plus or minus 1.96 times the standard deviation of the raw (uncorrected) ensemble forecasts (s_x), is represented by the grey shading. The dashed line shows the observed values of Niño-3.4 and the short-dashed line is the December climatological mean of 26.5^oC. In general, the interannual variability of the bias-corrected forecasts follows that of



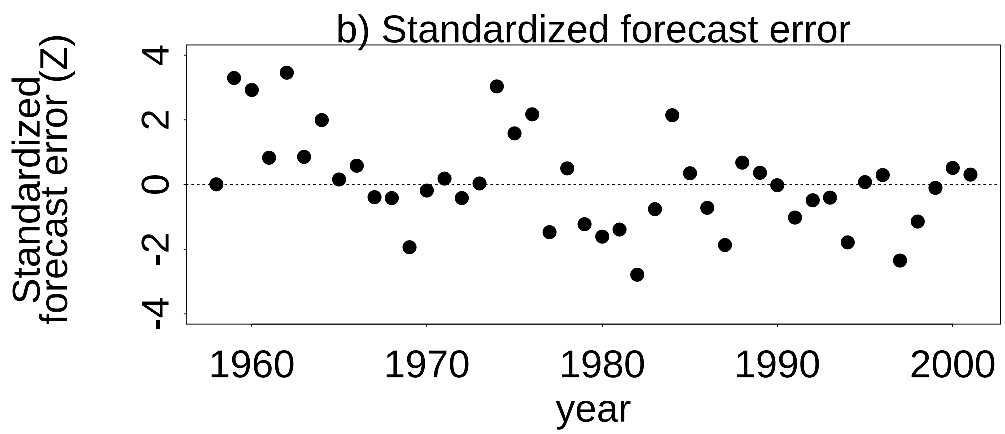


Figure 4.4: a) December Niño-3.4 index bias-corrected coupled model ensemble forecasts (^oC). Observed values (dashed line), forecasts (solid line) and the 95$\%$ prediction interval, given by the bias-corrected ensemble mean plus or minus 1.96 times the standard deviation of the ensemble forecasts (s_x), represented by the grey shading. The short-dashed line is the December 1950-2001 climatological mean (26.5 ^oC). b) Standardized forecast error.

the observations. However, in several cases the observations lie outside the prediction interval given by the ensemble spread indicating that these forecasts are poorly calibrated. Section 4.6 of this chapter will discuss quantitative comparisons of the skill and prediction uncertainty of the uncorrected coupled model, bias-corrected and empirical forecasts.

Figure 4.4b shows standardized forecast errors for the ECMWF bias-corrected coupled model ensemble forecast. Standardized forecast errors (Eqn. 4.5) were obtained by dividing the forecast error by the standard deviation of the 9 coupled model forecasts (s_x) for each year. Even after being bias-corrected these forecasts still show predominantly positive biases towards warmer Niño-3.4 values in the first half of the period and predominantly negative biases towards cooler Niño-3.4 values in the second half. The years 1962 and 1982 produced the largest standardized forecast errors due to having both large forecast errors and small ensemble standard deviations. The variance of the standardized forecast errors is found to be 2.23, which is much greater than the value of one expected for forecasts that have larger and more realistic estimates of forecast uncertainty (*cf.* the value of 1.15 obtained for the empirical forecasts).

The following section estimates the likelihood function (Eqn. 3.5) used to calibrate coupled model ensemble forecasts of the Niño-3.4 index against observations. To demonstrate the method of calibration, instead of bias-corrected forecasts issued by ECMWF, uncorrected (raw) coupled model outputs will be used.

4.5 Combined and calibrated forecasts

Figure 4.5 shows a scatter plot of raw (uncorrected) coupled model ensemble forecasts versus the observed December Niño-3.4 index for the period 1958-2001. Ensemble means are depicted using large open circles. The dashed line is what one expects for perfect forecasts in which the forecast values are identical to the observed values. The likelihood $p(\bar{x}_t \mid y_t)$ is modelled by performing a linear regression between the ensemble mean forecasts ($\bar{x}_t$) and matching observations (y_t) as indicated in Eqn. (3.5).

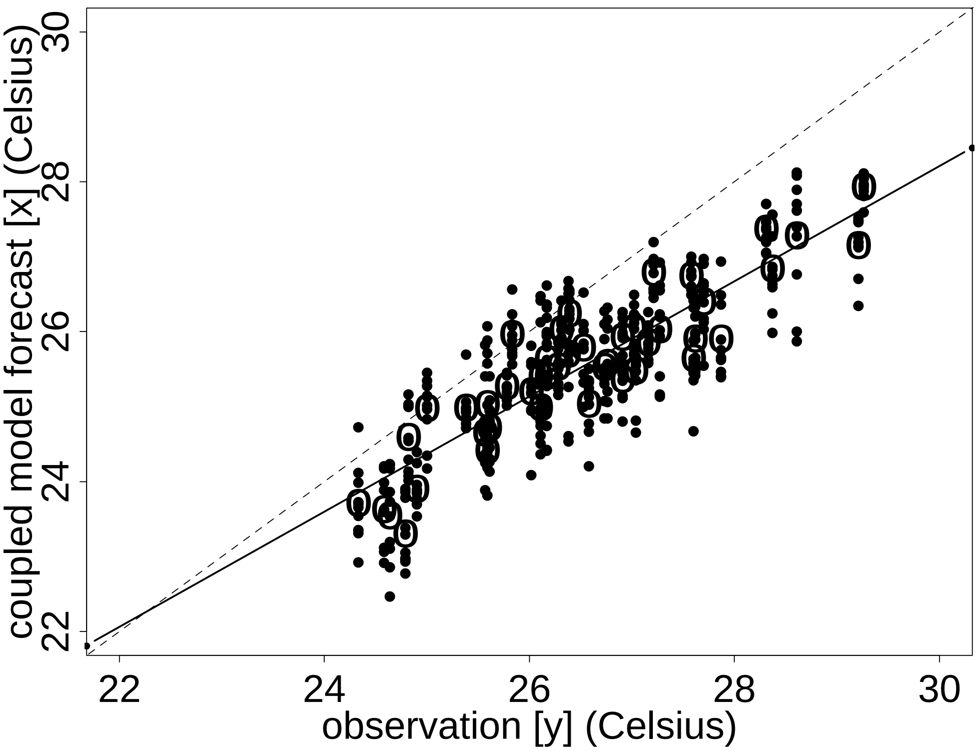


Figure 4.5: December Niño-3.4 index likelihood model. $\hat{\alpha} = 5.14^oC$, $\hat{\beta} = 0.77$, $\hat{\delta} = 0.19\,[^oC]^2$ and $R^2 = 0.83$. Each black dot is one ensemble member. Big open circles are ensemble means. The solid line is the regression between raw (uncorrected) ensemble mean forecasts and observations. The dashed line is what would be obtained for perfect forecasts.

The solid line in Fig. 4.5 is the best fit linear regression between raw (uncorrected) ensemble mean values $\bar{x}_t$ and observations y_t, corresponding to estimates for the whole period of $\hat{\alpha}$ = 5.14^{o}C, $\hat{\beta}$ = 0.77, $\hat{\delta} = 0.19\,[^oC]^2$ and $R^2 = 0.83$. It can be clearly seen that the uncorrected coupled model ensemble forecast is biased. These values and Fig. 4.5 indicate that: a) the variance in Niño-3.4 explained by the coupled model is underestimated [i.e. $Var(\bar{x}_t) < Var(\hat{y}_t)$ since $\beta < 1$]; and b) the coupled model generally underestimates the mean SST in the Niño-3.4 region [solid line below dashed line in Fig. (4.5)].

To avoid introducing artificial skill, likelihood distribution parameters are estimated using cross-validation by leaving out the year being forecast. The mean cross-validated likelihood estimated parameters are: $\hat{\alpha}$ = 5.14 (1.47) [^{o}C]; $\hat{\beta}$ = 0.77 (0.06); and $\hat{\delta}$ = 0.20 (0.01) $[^oC]^2$, where the values in brackets are the mean of the standard errors obtained for each of the cross-validated estimates.

The Bayesian procedure described in section 3.3 of chapter 3 allows the calibration of coupled model ensemble forecasts against observations by the likelihood function (solid line in Fig. 4.5). It also allows the combination of these calibrated forecasts with the empirical forecasts of Fig. 4.3, which are used to estimate the prior distribution parameters. Combined/calibrated forecasts are obtained from the posterior distribution (Eqn. 3.7), which has mean and variance given by Eqns. (3.8) and (3.9), respectively.

Figure 4.6a shows the mean of the combined forecast (solid line), observations (dashed line), the 95$\%$ P.I. (grey shading) and the December climatological mean of 26.5^oC (short-dashed line). Comparison of this forecast with the empirical forecast alone (Fig. 4.3a) and bias-corrected coupled model ensemble forecast alone (Fig. 4.4a) shows that the combined forecasts are in closer agreement with the observations. The 95$\%$ P.I.'s are also reduced compared to those of the empirical forecasts indicating a reduction in forecast uncertainty due to combination with coupled model forecasts. Unlike the bias-corrected coupled model forecasts, only three forecast years (1962, 1984, and 1994) fall outside the 95$\%$ P.I., indicating that the forecasts are better calibrated and more reliable than the bias-corrected coupled model forecasts.

Figure 4.6b shows the combined forecast standardized errors. The smallest errors were found within the period 1970-1978. The largest errors were in 1962, 1969, 1979, 1982, 1984 and 1994. Unlike the bias-corrected coupled model forecast standardized errors (Fig. 4.4b), combined forecast standardized errors (Fig. 4.6b) are evenly distributed and centred on zero.

Figure 4.7 shows plots of the standardized forecast error versus forecast values for the three types of forecasts presented so far. Figure 4.7b shows that the bias-corrected coupled model ensemble standardized forecast error is slightly positively biased and has large spread for forecast values between 26^oC and 28^oC than empir-

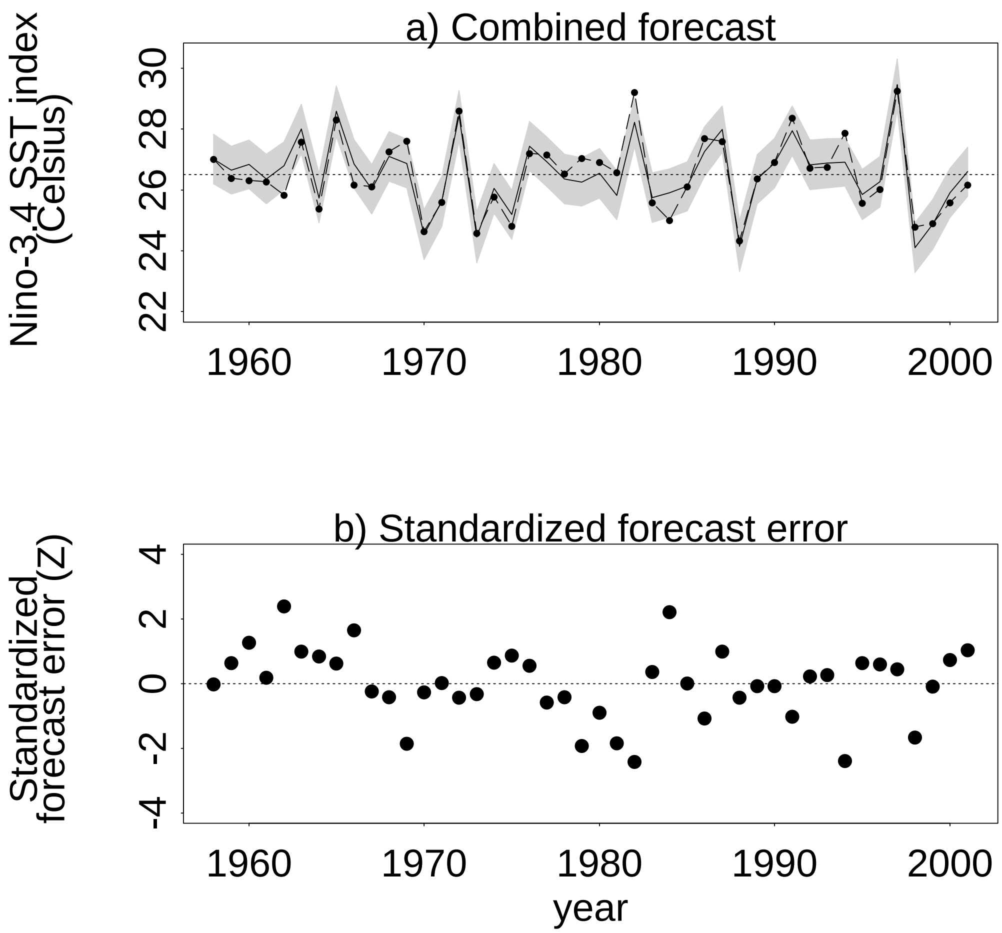


Figure 4.6: a) December Niño-3.4 index combined forecasts (^oC). Observed values (dashed line), forecasts (solid line) and the 95$\%$ prediction interval (grey shading). The short-dashed line is the December 1950-2001 climatological mean (26.5 ^oC). b) Standardized forecast error.

ical (Fig. 4.7a) and combined (Fig. 4.7c) forecasts. The standardized errors for the empirical forecast (Fig. 4.7a) and for the combined forecast (Fig. 4.7c) are evenly spread around the zero line.

4.6 Skill assessment

Table 4.1 gives some deterministic verification scores and a measure of forecast uncertainty for seven different December Niño-3.4 index predictions for the period

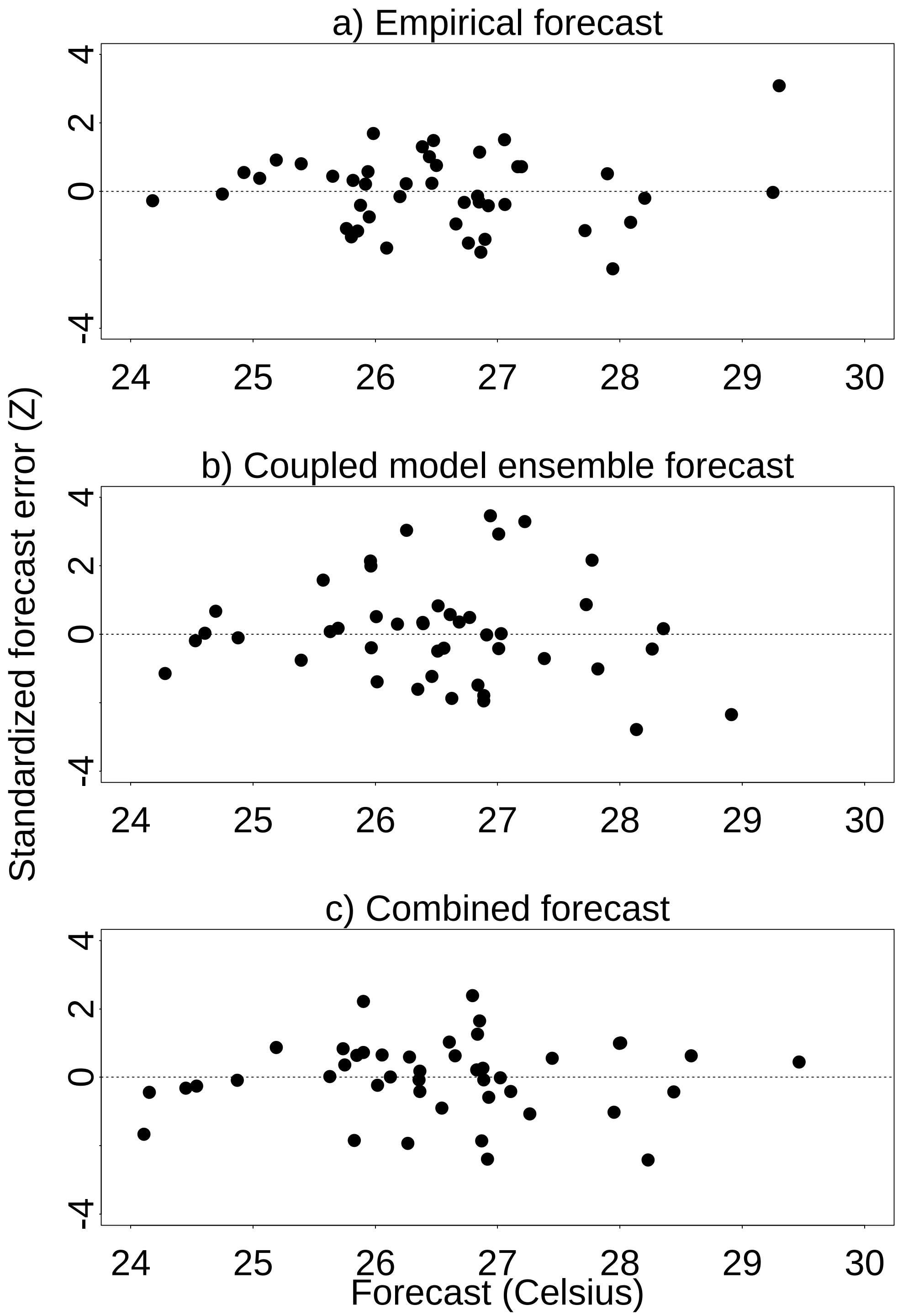


Figure 4.7: Standardized forecast error versus forecast in ^oC for (a) the empirical forecast, (b) the bias-corrected coupled model ensemble forecast and (c) the combined forecast

1958-2001. All the forecasts were produced using the cross-validation "leave one out" method. Table 4.1 summarises the skill of these forecasts in the 44-year period.

Forecast	Mean μ	Std. dev. σ	MSE $[^oC]^2$	MAE [^{o}C]	Skill Score [%]	Uncertainty [^{o}C]
Climatology	$\bar{y}$	s_y	1.47	0.98	0	1.20
Empirical	$\hat{\mu}_{ot}$	$\hat{\sigma}_{ot}$	0.37	0.49	74 (+58)	0.59
Raw ensemble	$\bar{x}_t$	s_x	1.23	0.99	16 (0)	0.44
Bias-corrected	$\bar{x}_t - \bar{\bar{x}} + \bar{y}$	s_x	0.26	0.41	82 (+66)	0.44
Uniform Prior	$\frac{\bar{x}_t-\alpha}{\beta}$	$\sqrt{\frac{\delta}{\beta^2}}$	0.34	0.49	77 (+61)	0.58
Regression	$a + b\bar{x}_t$	$\lambda^{\frac{1}{2}}$	0.28	0.43	81 (+65)	0.53
Combined	$\hat{\mu}_t$	$\hat{\sigma}_t$	0.20	0.34	87 (+71)	0.41
Perfect forecast	-	-	0	0	100	0

Table 4.1: Forecast symbols, verification scores, skill score and mean forecast uncertainty. MSE and MAE are the mean squared error and mean absolute error of the mean forecast, respectively. The skill is measured by the MSE skill score (see text for more details)- values in brackets indicate the percentage improvement compared to the raw ensemble system skill score. Forecast uncertainty is given by the mean of the prediction standard deviation over the period 1958-2001.

The seven forecasts of Table 4.1 are describe below:

• The climatological forecast is given by the historical Niño-3.4 index December mean value $\mu_t = \bar{y}$ of 26.5^oC and the historical December standard deviation $\sigma_t = s_y$ of 1.20^oC.

• The empirical forecast is given by $\mu_t = \hat{\mu}_{ot}$ and $\sigma_t = \hat{\sigma}_{ot}$, as defined in Section 4.2.2 of this chapter.

• The raw coupled model ensemble forecast is given by the mean $\mu_t = \bar{x}_t$ and the standard deviation $\sigma_t = s_x$ of the uncorrected ensemble forecast.

• The bias-corrected forecast is given by $\mu_t = \bar{x}_t - \bar{\bar{x}} + \bar{y}$ and $\sigma_t = s_x$, where $\bar{x}_t$ is the raw (uncorrected) ensemble mean forecast at time *t*, and $\bar{\bar{x}}$ and $\bar{y}$ are the time means of the raw (uncorrected) ensemble mean forecast and the observed mean values over the forecast period 1958-2001, respectively. This is a special case of a Bayesian forecast with uniform prior (defined below) and simplified likelihood [$\beta = 1$ in Eqn. (3.5)]. The simplified likelihood models the bias of the ensemble mean as a constant ($\alpha = \bar{\bar{x}} - \bar{y}$) and the sample variance of the ensemble forecast as $\delta = s_x^2$.

• The combined forecast with uniform prior is given by $\mu_t = \frac{\bar{x}_t - \alpha}{\beta}$ and $\sigma_t = \sqrt{\frac{\delta}{\beta^2}}$. It is obtained by setting ${\sigma_{ot}}^{-2}$ to zero in Eqns. (3.8) and (3.9), that is, *all* values of the index are equally likely. This prior characterises a "no-previous-information" reference case. The combined forecast with uniform prior can be seen as a Bayesian bias-correction in the raw (uncorrected) ensemble mean and it is useful for comparison with the bias-corrected forecast. Note, however, that the standard deviation $\sqrt{\frac{\delta}{\beta^2}}$ is not the same as s_x of the bias-corrected forecast.

• The regression forecast is given by $\mu_t = a + b\bar{x}_t$ and $\sigma_t = \lambda^{1/2}$, where *a*, *b* and λ are constant parameters estimated from the linear regression between observed values y and raw (uncorrected) coupled model ensemble-mean forecasts $\bar{x}$.

• The combined forecast is given by μ_t and σ_t, as defined by Eqns. (3.8) and (3.9).

Mean Squared Error (MSE) and Mean Absolute error (MAE) have been used as verification scores for the forecast means. The MSE skill score given by $SS = 1 - (MSE/MSE_c)$, where MSE_c is the climatological MSE, was used to measure forecast skill. Forecast uncertainty was summarised by the time mean of the prediction standard deviations over the forecast period 1958-2001.

Forecast	Likelihood	Prior
$a)$ bias-corrected	$\bar{x}_t \mid y_t \sim N(\alpha + y_t, s_x^2)$	Uniform (i.e. ${\sigma_{ot}}^2 \rightarrow \infty$)
$b)$ uniform prior	$\bar{x}_t \mid y_t \sim N(\alpha + \beta y_t, \delta)$	Uniform (i.e. ${\sigma_{ot}}^2 \rightarrow \infty$)
$c)$ Regression	$\bar{x}_t \mid y_t \sim N(\alpha + \beta y_t, \delta)$	$y_t \sim N(y_c, \sigma_c^2)$ (*)
$d)$ combined	$\bar{x}_t \mid y_t \sim N(\alpha + \beta y_t, \delta)$	$y_t \sim N(\beta_0 + \beta_1 \psi_t, {\sigma_{ot}}^2)$

Table 4.2: Likelihood and prior distributions for the correction methods applied to December Niño-3.4 index forecasts. (*) y_c and σ_c^2 are the climatological mean and the climatological variance of y, respectively, which are obtained with the same dataset used to estimate the likelihood.

The four correction methods applied to December Niño-3.4 index forecasts (i.e., bias-corrected forecast, combined forecast with uniform prior, regression and combined forecast) can all be seen as combined forecasts with particular likelihood and prior distributions as indicated in Table 4.2. Each likelihood model provides a different way of calibrating coupled model forecasts against observations. The incorporation of previous knowledge is obtained by the use of the prior distribution. When the prior is taken to be uniform, no previous knowledge is incorporated in the correction method. Note that by using the likelihood model $\bar{x}_t \mid y_t \sim N(\alpha + \beta y_t, \delta)$ and the normal prior $y_t \sim N(y_c, \sigma_c^2)$, where y_c and σ_c^2 are the climatological mean and the climatological variance of y, respectively, which were obtained with the same dataset used to estimate the likelihood, one gets a posterior distribution that is exactly the same as the linear regression between observed values y and coupled model ensemble-mean forecasts $\bar{x}$. This posterior distribution is given by the inverse regression model $y_t \mid \bar{x}_t \sim N(a + b\bar{x}_t, \lambda)$ that is here referred to as regression forecast (forecast c) of Table 4.2). The proof of this equivalence is shown in Appendix A for standardized variables. Hoadley (1970) provides the general proof for any continuous variable and section 3.4.3 of chapter 3 gives the general proof of this equivalence for the multivariate regression of observations on forecasts. The regression model of Table 4.2 is a particular case of the general multivariate regression model presented in section 3.4 of chapter 3. However, instead of having multivariate variables for forecasts and observations here both are univariate variables.

Table 4.1 shows that raw (uncorrected) coupled model and climatological forecasts have the largest forecast MSE and MAE and the poorest skill. The empirical forecasts outperform uncorrected and climatological forecasts. However, bias-corrected coupled model forecasts outperform the empirical, uncorrected and uniform prior forecasts. Bias-corrected and uniform prior forecasts incorporate no previous information, i.e. both assume infinite variance for the prior distribution ($\sigma_{ot}{}^2 \to \infty$ as indicated in Table 4.2). This means that the resulting forecast improvement is due to the calibration, which is given by the likelihood. The fact that bias-corrected forecasts outperform uniform prior forecasts suggests that the calibration used to produce bias-corrected forecasts, which does not have the scaling factor β in the likelihood model, is better than the calibration used to produce uniform prior forecasts. Regression forecasts (forecast c) of Table 4.2), which use the same likelihood model as uniform prior forecasts but a more informative prior given by a normal distribution with mean and variance obtained from the same historical values of y used to build the likelihood, outperform uniform prior forecasts and have comparable MSE, MAE and MSE skill score to bias-corrected forecasts. This indicates that the use of a better prior helped to reduce forecast error. Bayesian combined forecasts, which use the same likelihood model as regression and uniform prior forecasts and a more refined prior, outperform all other forecasts. This indicates that these forecasts are better calibrated due to the use of a better and more informative prior than those used in the other forecasts. Bayesian combined forecasts show an impressive improvement of 71$\%$ in skill when compared to the raw (uncorrected) coupled model forecasts, indicating that the use of a more informative prior led to additional improvement in forecast skill.

Table 4.1 still shows that combined, raw and bias-corrected coupled model forecasts give the smallest forecast uncertainty estimates. Both raw (uncorrected) and bias-corrected mean prediction uncertainties have exactly the same values because bias-correction does not correct biases in the ensemble variance. Climatological, empirical, uniform prior and regression forecasts give larger uncertainty estimates than raw (uncorrected) and bias-corrected forecasts. Combined forecasts give the smallest uncertainty estimates of all forecasts. The width of the 95$\%$ P.I. in Figs. (4.3a), (4.4a) and (4.6a), which is proportional to the mean uncertainty estimate of Table 4.1, shows that combined forecasts (Fig. 4.6a) provide more realistic and reliable uncertainty estimates compared to the other forecasts, with only a few observations outside the 95$\%$ P.I.

Table 4.3 summarizes the standardized forecast errors. Well-calibrated forecasts should have standardized forecast errors with zero mean and unit variance. The mean standardized forecast error shows that the raw (uncorrected) coupled model forecast is negatively biased, with the largest mean error of all the forecasts. Bias-corrected forecasts are slightly positively biased. The climatological, empirical, uniform prior, regression and the combined forecast have the smallest mean errors (close to zero), indicating that these forecasts are unbiased. The raw (uncorrected) coupled model ensemble and the bias-corrected ensemble forecasts have the largest variances of the standardized forecast errors. All other forecasts have variances slightly larger than one suggesting that the prediction uncertainty of these forecasts is being marginally underestimated.

Forecast	Mean	Variance
Climatology	-0.01	1.05
Empirical	0.02	1.15
Raw ensemble	-2.57	3.11
Bias-corrected	0.11	2.23
Uniform Prior	0	1.06
Regression	0	1.08
Combined	0.01	1.21
Perfect forecast	0	1

Table 4.3: The mean and variance of standardized forecast errors.

All forecasts presented here have been assessed using deterministic verification scores. Note, however, that the forecast mean μ_t and the forecast standard deviation σ_t of each of these forecasts can be used to produce probabilistic forecasts under the assumption that the forecasted variable follow a normal (Gaussian) distribution, i.e., $y_t \sim N(\mu_t, \sigma_t^2)$. To demonstrate this utility, let's consider forecast probabilities p_k for the event Niño-3.4 index y_k less than its climatological value of 26.5 ^oC (i.e. $p_k = Pr(y_k < 26.5^oC)$, where $Pr(E)$ denotes the probability of the event E). The corresponding binary observations are stored in the variable o_k, i.e., $o_k = 1$ when the event $y_k < 26.5^oC$ occurred and $o_k = 0$ when the event $y_k < 26.5^oC$ has not occurred. The index k denotes a numbering of the n forecast/observation pairs.

The skill of forecast probabilities p_k for the event $p_k = Pr(y_k < 26.5^oC)$ can be assessed using the Brier score (Brier 1950) given by

$$BS = \frac{1}{n}\sum_{k=1}^{n}(p_k - o_k)^2. \qquad (4.6)$$

The Brier score (BS) is analogous to the MSE, but instead of averaging squared differences between pairs of mean forecast values μ_t and observed values y_t, it averages the squared differences between pairs of forecast probabilities p_k and the subsequent binary observations o_k. The Brier score can take values in the range $0 \leq BS \leq 1$. The Brier score is negatively oriented, with perfect forecast exhibiting $BS = 0$. Therefore, the smaller the Brier score the better is the quality of the forecast.

Forecast	BS	BSS (%)
Climatology	0.25	0
Empirical	0.13	49
Raw ensemble	0.28	-11
Bias-corrected	0.15	39
Uniform Prior	0.15	42
Regression	0.14	42
Combined	0.12	52
Perfect forecast	0	100

Table 4.4: Brier score (BS) and Brier Skill Score (BSS) in percentage.

Table 4.4 shows the Brier score of all forecasts presented so far for the event Niño-3.4 index less than its climatological value of 26.5^oC ($y_t < 26.5^oC$). The climatological probability forecast ($p_k = 0.52$) is obtained counting the number of times the observed Niño-3.4 index was less than its climatological value of 26.5^oC and dividing this count by the total number of events ($n = 44$). Table 4.4 also shows the Brier skill score (BSS) defined as $BSS = (1 - BS/BS_c) \times 100\%$, where BS_c is the Brier score of the climatological forecast. By definition the Brier skill score of the climatological (reference) forecast is 0%, delimiting the no-skill bottom line from which other forecasts must improve. Therefore, the Brier skill score indicates how much better a particular forecast is compared to the climatological no-skill reference forecast. Table 4.4 shows that raw (uncorrected) forecasts have the largest Brier score of all forecasts, even larger than the Brier score of the reference climatological forecast, leading to a negative BSS. Bias-corrected, uniform prior and regression forecats are better calibrated than raw (uncorrected) coupled model forecasts, having comparable values of Brier score and BSS. Empirical forecasts have the second smallest Brier score and consequently the second highest BSS, suggesting that these forecasts are better calibrated than climatological, raw (uncorrected), bias-corrected, uniform prior and regression forecasts. Finally, combined forecasts have the smallest Brier score and the highest BSS, indicating that these forecasts are better calibrated and more skillful than all other forecasts.

4.7 Summary

The univariate Bayesian normal model introduced in chapter 3 (section 3.3) has been used for calibrating and combining empirical and raw (uncorrected) coupled model ensemble forecasts of the Niño-3.4 index at a 5-month lead time. The combined forecast has been shown to have greater forecast skill than either of the forecasts individually. This indicates that both empirical and raw (uncorrected) coupled model ensemble forecasts contain mutually useful information. In other words, neither forecast is sufficient for the other forecast and so increased forecast skill can be obtained by combining both types of forecast. In order to produce improved interval forecasts of the Niño-3.4 index, empirical and coupled model forecasts should be combined together. The combined forecast also provides a more reliable prediction error estimate because it is based on a well-founded calibration approach that incorporates valuable historical information.

Good quality forecasts are expected to have both small prediction errors (good accuracy) and reliable forecast uncertainty estimates. It has been shown that, although the ECMWF raw (uncorrected) coupled model ensemble forecast is able to simulate the inter-annual variability of the Niño-3.4 index reasonably well 5 months in advance, it underestimates the mean SST value in the Niño-3.4 region and provides unreliable forecast uncertainty estimates. The simple empirical model, on the other hand, provides more skillful forecasts compared to the raw (uncorrected) coupled model ensemble forecast. These forecasts are less biased and present larger and more reliable uncertainty estimates. When the Bayesian approach was used to combine these two forecasts together, more skillful forecasts were obtained (both in deterministic and probabilistic terms) having more accuracy and reliability.

Chapter 5

A one-dimensional example:

Equatorial Pacific SST forecasts

5.1 Aim

The aim of this chapter is to test the use of forecast assimilation (section 3.4) for the calibration and combination of physically-derived dynamical climate model equatorial Pacific SST forecasts produced at a 6-month lead time. Forecast assimilation is a probabilistic Bayesian approach that converts climate model predictions into well-calibrated probability forecasts of real-world observable variables. Grid point forecasts of all seven DEMETER climate models are used in this demonstration.

5.2 Introduction

Forecasting equatorial Pacific SSTs associated to ENSO appropriately is the first step for successful seasonal predictions. ENSO is characterised by a strong coupling between oceanic and atmosphere physical processes. During warm (El Niño) episodes, when SSTs are warmer than normal in the equatorial Pacific, trade winds (near surface winds at the equator) are weaker than normal in the central and western equatorial Pacific (Bjerknes 1966; Rasmusson and Carpenter 1982; Philander 1983; Philander 1985; McPhaden *et al.* 1998). Conversely, during cold (La Niña) episodes, when SSTs are colder than normal in the equatorial Pacific, trade winds are stronger than normal in the central and western equatorial Pacific.

This relationship between surface winds and SST is an important aspect of ENSO. Deep atmospheric convection typically occurs over the warmest SSTs in the tropical Pacific (Graham and Barnett 1987). Warm SSTs (above 30^{o}C) in the equatorial region near the date line, in a region of strongly convergent surface winds, help to activate deep atmospheric convection. Converging winds act to sustain both deep convection (via moisture convergence) and warm SSTs (via ocean dynamics) (Bjerknes 1966; Philander *et al.* 1984). These processes locally reinforce one another, and representing them properly in coupled ocean-atmosphere models is a challenge for ENSO modelling (e.g. Zebiak and Cane 1987; Battisti 1998; Délécluse *et al.* 1998).

The zonal redistribution of warm surface layer water masses is an important oceanic feature of ENSO (Wyrtki 1975; Philander 1990; McPhaden 1995). In the western equatorial Pacific the thermocline (identified by the depth of the 20^{o}C isotherm) shallows by 20-50 metres during El Niño, whereas in the eastern equatorial Pacific the thermocline deepens by a comparable amount. These thermocline depth variations are correlated with changes in the strength of equatorial oceanic currents. As a result of these changes, an anomalous eastward mass transport of warm water by the equatorial surface currents is observed during the onset of El Niño events.

Changes in the zonal distribution of the upper ocean heat content are reflected in sea level variations (Wyrtki 1975; Rebert *et al.* 1985) because of the vertically coherent structure of the upper ocean thermal field. In other words, an anomalously deep thermocline tends to be associated with anomalously high sea levels and vice-versa. Adjustments of the upper ocean heat and mass are strongly influenced by excitation and propagation of equatorial oceanic waves. These are the primary mechanisms by which the winds communicate their influence to other parts of the ocean basin. Westerly wind bursts, usually observed in the western equatorial Pacific, can generate easterly propagating oceanic Kelvin waves (Miller *et al.*

1988; McPhaden *et al.* 1988). These waves can depress the thermocline in the eastern Pacific, reducing upwelling near the west coast of South America. Westerly wind bursts can also affect surface currents, causing zonal advection of water. This zonal advection is evident in thermocline variations, as well as in time series of sea level. Kelvin waves generated by westerly wind bursts can either initiate or terminate an ENSO event in the equatorial Pacific, although triggering of ENSO events by westerly bursts is still a subject that deserves further investigation (Ineson and Davey 1997; Vitart *et al.* 2003).

The importance of simulating coupled ocean-atmosphere interactions in the equatorial Pacific is clear from the preceding exposition. The next sections of this chapter present equatorial Pacific SST forecasts produced by the seven DEMETER climate coupled models listed in Table 5.1. The quality of these forecasts gives an indication of how well these models can simulate the complex non-linear physical interactions of the coupled ocean-atmosphere system at the equator. Both the simple multi-model ensemble forecasts and Bayesian combined and calibrated probabilistic and deterministic forecasts obtained with forecast assimilation are shown and discussed.

5.3 Coupled model forecasts

The DEMETER project has produced an invaluable multi-model ensemble of global coupled model seasonal hindcasts. Hindcasts were produced four times a year using ERA-40 reanalysis initial conditions starting at 00GMT on the first day of February, May, August and November, as described in section 2.2.4 of chapter 2. This section will focus on the longest lead 6-month ahead predictions of equatorial Pacific SST for the four target months of July, October, January and April. The results presented here are based on the 63 hindcasts created by the 9-member ensembles from the seven DEMETER coupled models over the common period 1980-2001. Predictions are verified against the ERA-40 reanalysis SSTs (see appendix C for further information).

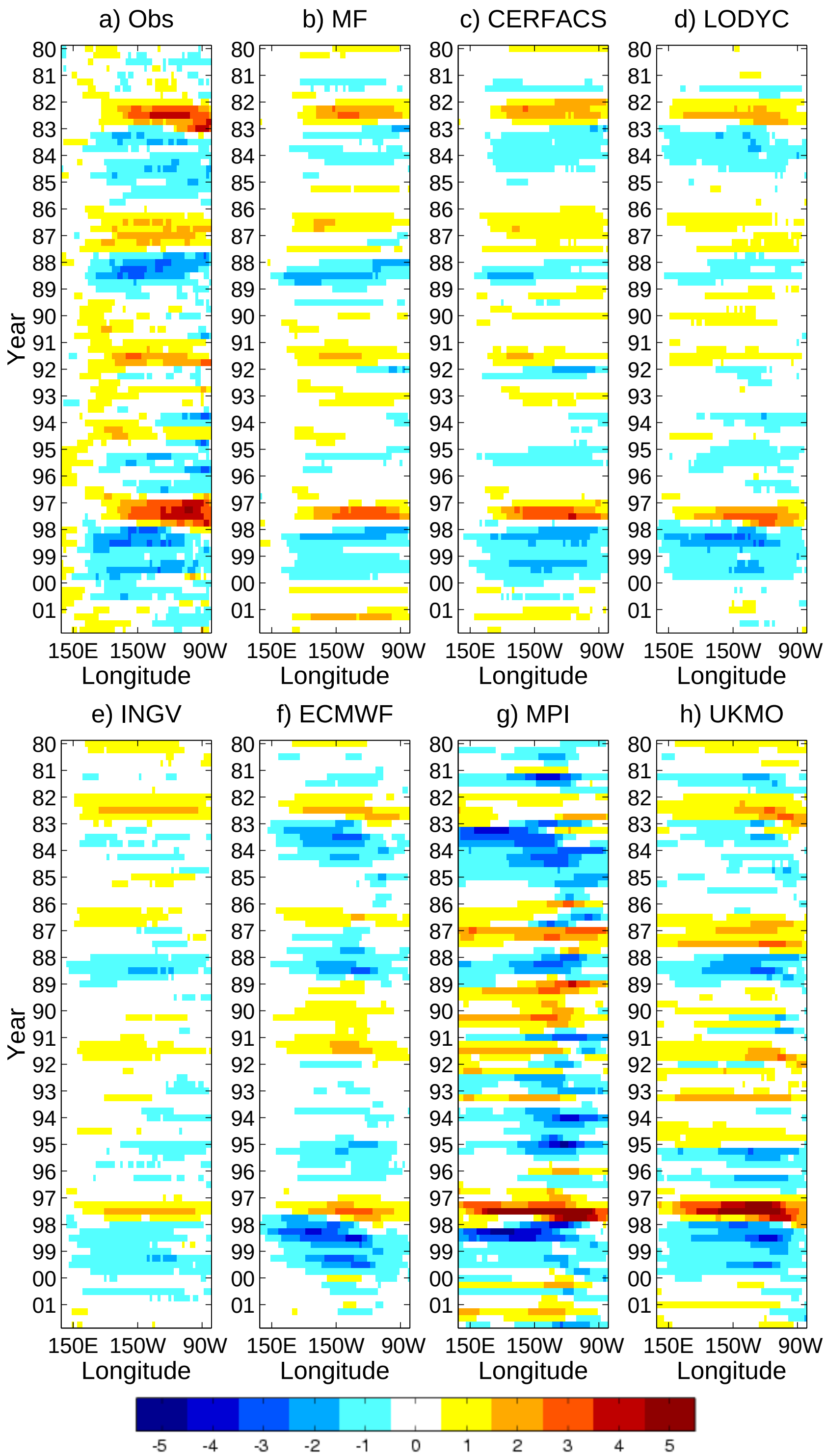


Figure 5.1: Hovmuller plots of Pacific SST anomalies (^oC) along the equator (140^{o}E-82.5^{o}W) from July 1980 to July 2001: a) observed anomalies, and 6-month lead forecasts from b) Météo-France, c) CERFACS, d) LODYC, e) INGV, f) ECMWF, g) MPI, and h) UKMO.

Model	MIN (^oC)	MAX (^oC)	RMSE (^oC)	BS
MF	-2.5	3.4	0.84	0.20
CERFACS	-2.5	3.5	0.82	0.20
LODYC	-3.0	3.0	0.82	0.22
INGV	-1.7	2.3	0.88	0.22
ECMWF	-3.8	3.2	0.89	0.21
MPI	-5.2	9.8	1.46	0.29
UKMO	-4.1	5.7	0.90	0.22
Ens. mean	-3.4	4.4	0.77	0.19
FA 1980-2001	-3.5	4.8	0.75	0.17
FA 1958-2001	-3.4	4.8	0.75	0.17
Obs. 1980-2001	-2.7	4.5	1.13	0.25

Table 5.1: Forecast statistics for the various models based on all the gridded anomalies shown in Fig. 5.1: MIN is the minimum forecast value, MAX is the maximum forecast value, RMSE is the Root Mean Squared Forecast Error and BS is the Brier Score. The Brier score is given for probabilistic forecasts of cold events defined by anomalies less than or equal to zero. The statistics were calculated by pooling over all the space-time points in the hovmuller plots. The Brier score for the observations is obtained by forecasting the climatological value of $p = 0.5$ for each event at each grid point. The RMSE for the observations is that obtained by forecasting an anomaly of zero.

Figure 5.1 shows longitude-time hovmuller plots of SST anomalies along the equator in the Pacific sector from Indonesia to the west coast of South America (sampled four times a year: January, April, July, October). The equatorial Pacific section contains 56 grid points along the equator running from 140^{o}E to 82.5^{o}W. The observed SST anomalies (Fig. 5.1a) clearly reveal four major positive anomaly El Niño events (1982/83, 1986/87, 1991/92, 1997/98) separated by negative anomaly La Niña episodes. The individual coupled model DEMETER forecasts are shown in panels b)-h) of Fig. 5.1 – anomalies were produced by subtracting the long-term mean for each calendar month for each of the models. All of the 6-month lead model forecasts capture the main features of the observed ENSO events in the Pacific SSTs. However, more careful inspection reveals that all the models except MPI and UKMO tend to underestimate the peak magnitude of the ENSO events. MPI also has more interannual variation than the observations and the other models. Furthermore, most of the model forecasts are rather similar and appear to overestimate the spatial extent of certain El Niño events, such as 1997/98.

Table 5.1 summarizes the forecasts using statistics calculated by pooling over all the anomaly data shown in the hovmuller plots of Fig. 5.1. The RMSE and BS were constructed by averaging over all 56 grid points and 88 time points shown in the hovmuller plots. With the exception of the MPI model, the model predictions have similar RMSE scores (0.82-0.90). The MPI predictions have more variance and lower correlations with the observations and this leads to a much larger RMSE score. Apart from the MPI and UKMO predictions, the models tend to underestimate the maximum values and overestimate the minimum values compared to observations. The RMSE scores are similar to those obtained for ENSO forecasts using empirical regression methods (Coelho *et al.* 2003; Coelho *et al.* 2004).

5.4 Combined and calibrated forecasts

This section presents and compares combined forecasts obtained with forecast assimilation (as described in section 3.4.1) with simple multi-model forecasts of equatorial Pacific SST anomalies. Figures 5.2a-d show longitude-time hovmuller plots of equatorial Pacific SST anomalies for observations and three different types of combined forecast. One of the simplest and most naïve ways to combine multi-model predictions is to calculate the ensemble mean of all the model predictions (Doblas-Reyes *et al.* 2003; Kharin and Zwiers 2002; Krishnamurti *et al.* 2001). The comparison of the ensemble mean (Fig. 5.2b) with the observations (Fig. 5.2a) shows that the ensemble mean captures the main ENSO events. However, more careful examination reveals that the peak magnitudes of El Niño events in the ensemble mean forecast are slightly smaller than those in the observations and the peak magnitudes of the La Niña events are slightly larger. With the exception of MPI and UKMO, the models generally underestimate the magnitude of the maximum anomalies whereas only one model (INGV) underestimates the magnitude of the minimum anomalies when compared to observations (Table 5.1). This leads to

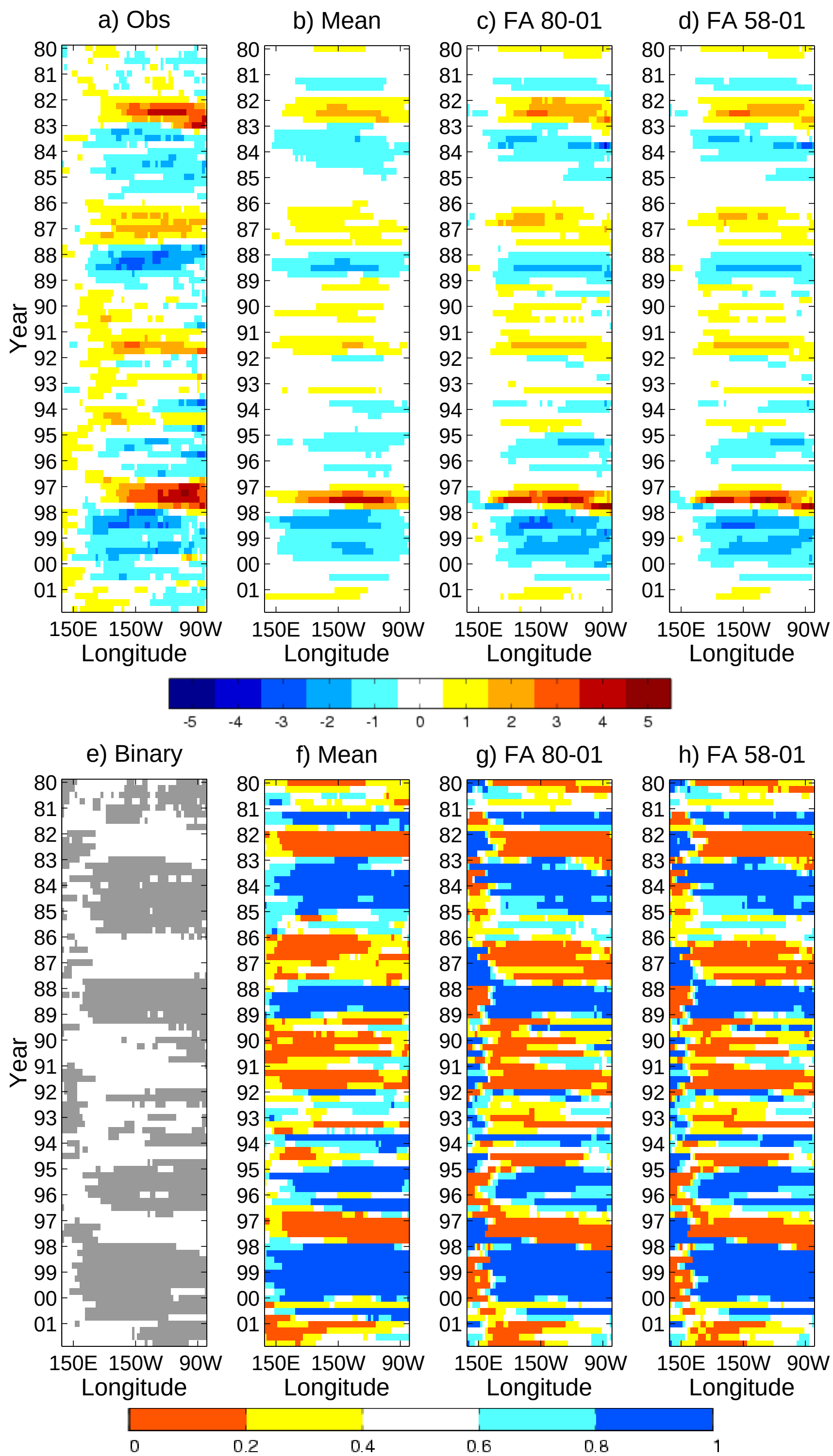


Figure 5.2: Hovmuller plots of SST anomalies (^oC) along the equator (July 1980 to July 2001): a) observations, b) multi-model ensemble mean 6-month lead forecast, c) the forecast assimilation forecast with prior estimated over 1980-2001, d) the forecast assimilation forecast with prior estimated over 1958-2001. The lower four panels show the binary event defined by when the observed anomaly is less than or equal to zero (grey shading in panel e), and the corresponding probability forecasts for the binary event based on the multi-model ensemble mean and standard deviation (panel f), forecast assimilation with prior from 1980-2001 (panel g), and forecast assimilation with the prior from 1958-2001.

the ensemble mean forecast underestimating the observed interannual variance.

Figure 5.2c shows the combined forecast obtained with forecast assimilation, where the likelihood and prior were estimated over the common period 1980-2001 (FA 1980-2001). As explained in section 3.4.3, this special case of forecast assimilation is identical to the more traditional MOS multivariate regression of observations on forecasts. The FA 1980-2001 combined forecast (Fig. 5.2c) qualitatively resembles the ensemble mean forecast (Fig. 5.2b) although there are some important differences in sign in the western Pacific. For comparison, Figure 5.2d shows the forecast assimilation forecast where the prior has been estimated over the extended period 1958-2001 (FA 1958-2001). Figure 5.2d closely resembles Fig. 5.2c implying that prior information about observed SSTs does not make a large difference to the results in this particular example. There is sufficient information in the calibration period 1980-2001 to provide a good estimate of the probability density of observed SSTs without the need for more prior information about the observations. However, prior observational information will improve the final probability forecasts in applications where the prior is more informative relative to the model predictions. For example, in applications where the model predictions are less skillfull or where the prior is more skillfull (e.g. an empirical forecast instead of climatology). From Table 5.1 it can be seen that all the combined forecasts have smaller RMSE than the individual model predictions and that the forecast assimilation forecasts give slightly smaller RMSE than the ensemble mean forecasts. Furthermore, the combined forecasts give minimum and maximum values that are in closer agreement with the observations than those obtained for individual models.

Optimal decision-making requires an estimate of prediction uncertainty in addition to a forecast of the mean. Prediction uncertainty $\hat{\sigma}$ is the before-the-event prediction of the root mean squared error of the forecast. When accurately estimated, the long-term mean prediction uncertainty should equal the standard deviation of the forecast errors. The prediction uncertainty for the ensemble mean forecasts is the standard deviation of the 7 model ensemble mean forecasts. This naïve approach will be adopted here although it is likely to give an underestimate of the true prediction uncertainty due to dependency between different model forecasts. The prediction uncertainty for the forecast assimilation forecasts takes account of some of the model dependency by using the square root of the diagonal elements of the matrix D (Eqn. 3.24). One of the advantages of the forecast assimilation approach is that it is capable of giving more realistic prediction intervals and hence (as will be shown) more reliable probability forecasts.

The skill of probabilistic forecasts is tested by forecasting the simple binary event of observed SST anomalies at each grid point being less than or equal to zero (Fig. 5.2e). Grey shaded areas illustrate events when SST anomalies were observed to be less than or equal to zero. White areas illustrate events when SST anomalies were observed to be above zero. For each type of combined forecast, the predicted forecast mean and uncertainty have been used to calculate a probability for the anomaly being less than or equal to zero

$$p = Pr(Y_p \leq 0) = \Phi(\frac{0 - y_p}{\hat{\sigma}}) \tag{5.1}$$

where $Y_p \sim N(y_p, \hat{\sigma}^2)$ and $\Phi(y_*)$ is the area under the standard normal curve to the left of $y_* = -y_p/\hat{\sigma}$. Note that the probability forecast depends on both forecast of the mean y_p and the prediction uncertainty $\hat{\sigma}$. These probability forecasts are shown for the combined forecasts in Figs. 5.2f-h. Note that the forecast assimilation but not the multi-model ensemble mean probabilities forecast correctly the reversed events visible to the west of the date line at 180^o in Fig. 5.2e.

The skill of the probability forecasts has been assessed using the Brier score (Eqn. 4.6). Brier scores for the three combined forecasts have been calculated by pooling over all the space-time points in the hovmuller plots and are given in Table 5.1. With the exception of the MPI model, all forecasts have smaller Brier score than 0.25 and so are more skillful than climatological forecasts that always issue the probability of 0.5. The relative improvement of the Brier scores compared to the no-skill Brier score of 0.25 for climatology appears small but this is a common feature of the Brier score known to occur for even quite skillful forecasting systems.

The three combined forecasts have smaller Brier score than any of the individual models and so combining has improved the skill of the probability forecasts. The forecast assimilation forecasts have the smallest Brier scores and so provide the most skillful probability forecasts. The use of prior information from 1958-2001 rather than 1980-2001 has little effect on the Brier score of the forecast assimilation forecasts.

Murphy (1973) has shown that the Brier score can be decomposed in three components: reliability, resolution and uncertainty (see decomposition in Appendix D). As described in Appendix E, these components can be interpreted geometrically using a reliability diagram. Figure 5.3 shows the reliability diagrams for the three combined forecasts of the event that SST anomalies are less than or equal to zero. The s-shaped curve of Fig. 5.3a indicates that the multi-model ensemble is overconfident. In other words, it forecasts high probabilities p_i when smaller frequencies $\bar{o}_i$ are observed, and low probabilities when larger frequencies are observed. The FA 1980-2001 and FA 1958-2001 forecasts (Figs. 5.3b and 5.3c, respectively) also show s-shaped curves but are much closer to the diagonal $\bar{o}_i = p_i$ line, indicating better calibrated forecasts and smaller Brier scores. The area between the dashed line and the diagonal is smaller in Figs. 5.3b and 5.3c than in Figs. 5.3a, indicating that FA 1980-2001 and FA 1958-2001 forecasts have better reliability than the multi-model ensemble forecast (see Appendix E for further explanation). These forecasts also have better resolution than the multi-model ensemble forecast since the area between the dashed line and the horizontal dotted line is larger for FA 1980-2001 and FA 1958-2001 forecasts (Figs. 5.3b and 5.3c) than for the multi-model ensemble forecast (Figs. 5.3a). This improvement in resolution can also be noted in the histograms plotted in the bottom right corner of each panel of Fig. 5.3, which shows the forecast probability frequency for the 10 equally spaced probability bins from 0 to 1 used to construct the reliability diagrams. The higher frequency in the first and last bins of the histograms of Figs. 5.3b and 5.3c compared to Fig. 5.3a indicate that FA 1980-2001 and FA 1958-2001 forecasts have better resolution than the multi-model ensemble forecast.

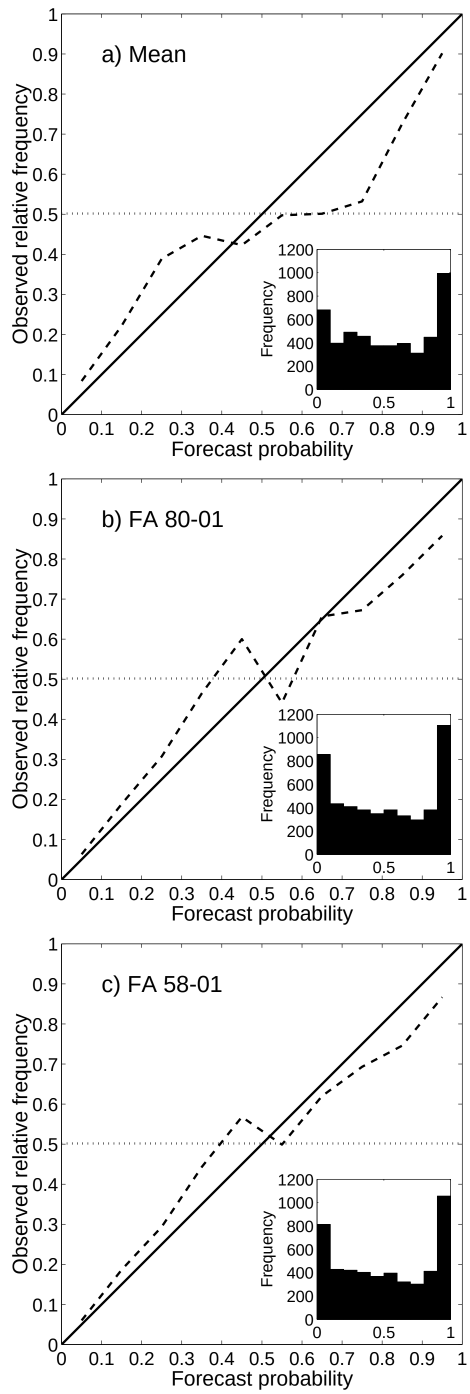


Figure 5.3: Reliability diagram for the event 'SST anomalies less than or equal to zero' a) for the multi-model ensemble forecast, b) for FA 1980-2001 forecasts, and c) for FA 1958-2001 forecasts. Probabilities have been ordered and grouped in 10 equally spaced probability bins from 0 to 1 and are plotted in the centre of each bin. Perfectly calibrated forecast should have all points falling on the diagonal solid line. The horizontal dotted line is the climatological observed frequency of the event. Histogram plots in the bottom right corner of each panel show the forecast probability frequency for these 10 equally spaced probability bins.

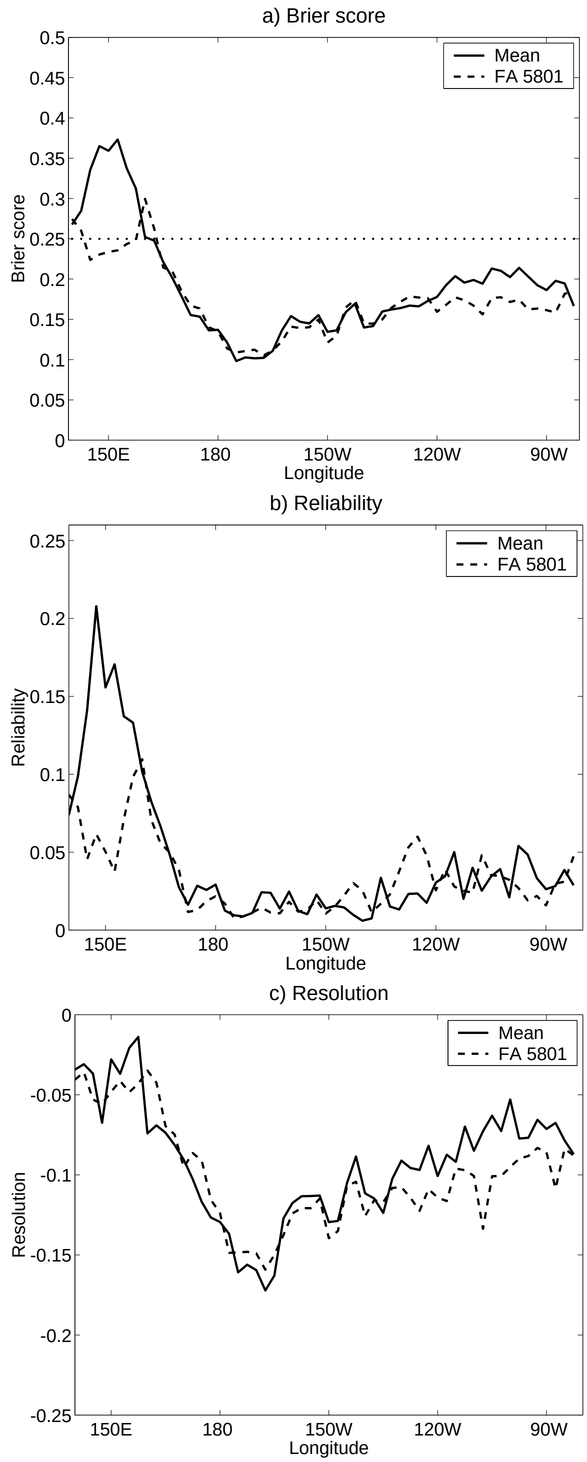


Figure 5.4: The Brier score and its components as a function of longitude for the multi-model ensemble (solid line) and forecast assimilation FA 1958-2001 (dashed line) probability forecasts of SST anomalies less than or equal to zero. Panel: a) the Brier score, b) the reliability component, and c) the negated resolution component. The reliability and negated resolution components were estimated at each longitude using 10 equally spaced probability bins from 0 to 1. The Brier score is the sum of the reliability, negated resolution, and uncertainty (close to 0.25 at all longitudes) terms. Smaller values indicate more forecast skill.

Given the similarities of FA 1980-2001 and FA 1958-2001 the remaining of this chapter will focus on FA 1958-2001. Figure 5.4 shows the Brier score and its resolution and reliability components as a function of longitude. Fig. 5.4a shows that the forecasts are most skillful in the central Pacific, around 170^{o}W, in the region that has the smallest Brier score of around 0.1. In this region the Brier scores are smaller than the score of 0.25 obtained from the climatological forecasts that always forecast the probability of 0.5. The climatological forecast is indicated in Fig. 5.4a by the horizontal dotted line. To the west of the date line, the forecast assimilation score (dashed line) is markedly smaller than the multi-model ensemble mean score (solid line), which is worse here than the score for climatology (dotted line). The forecast assimilation score is also slightly less than that of the multi-model ensemble mean forecast eastwards of 120^{o}W. Figure 5.4b shows that the improvement in the western Pacific is due to an improvement in the reliability of the forecasts. As can be noted from Figs. 5.2e-h, the multi-model ensemble mean fails to capture the reversed sign of the events in the western Pacific, whereas the forecast assimilation forecasts are able to capture this behaviour. A possible cause for the westward displacement of the forecast anomalies is that four (LODYC, ECMWF, MPI and UKMO) of the seven models have cold SST biases in the central-western equatorial Pacific (not shown), suggesting that these models may have excessively strong easterly winds over this region. Similar behaviour has been noticed by Davey *et al.* (2002) in an intercomparison study of 23 coupled models. Figure 5.4c shows that the resolution of the combined forecasts is similar although there is evidence of slightly improved resolution in the forecast assimilation forecasts eastwards of 135^{o}W.

As previously mentioned in section 3.4.2, forecast assimilation requires dimension reduction to deal with the large dimensionality of gridded datasets compared to the number of independent forecasts and colinearity problems at neighbouring grid points. In order to show how dimension reduction affects the dependency of the skill of the combined forecast, Fig. 5.5 shows the root mean squared errors and Brier scores for cross-validated combined forecasts as a function of the number

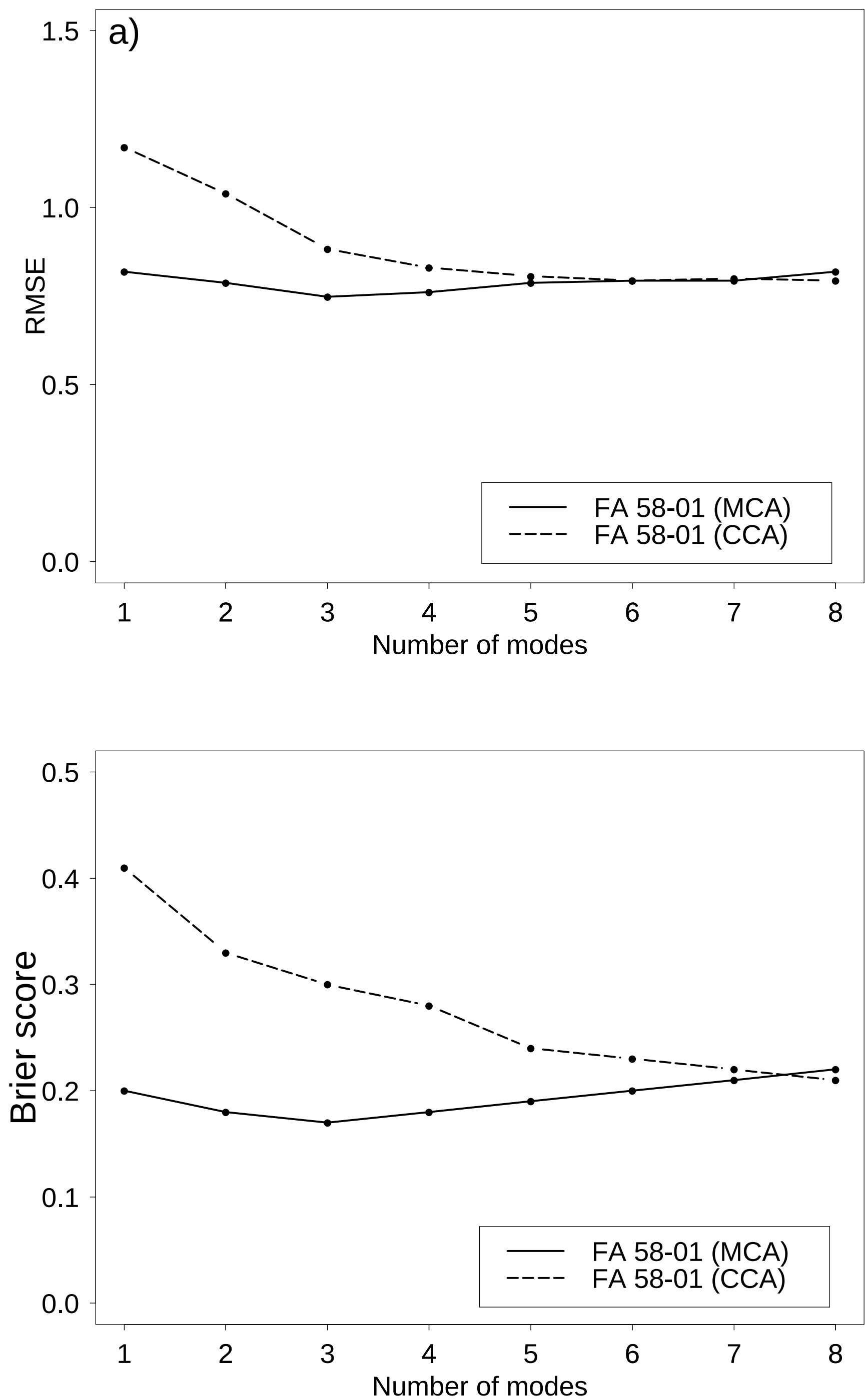


Figure 5.5: Scores of FA 1958-2001 cross-validated forecasts versus the number retained modes for MCA (solid line) and CCA (dashed line) data reduction. Panel a) Root Mean Squared Error (^oC), b) Brier score. Scores are calculated for each number of retained modes using all longitude and time values of hovmuller plots similar to those shown in Figs. 5.2d and 5.2h.

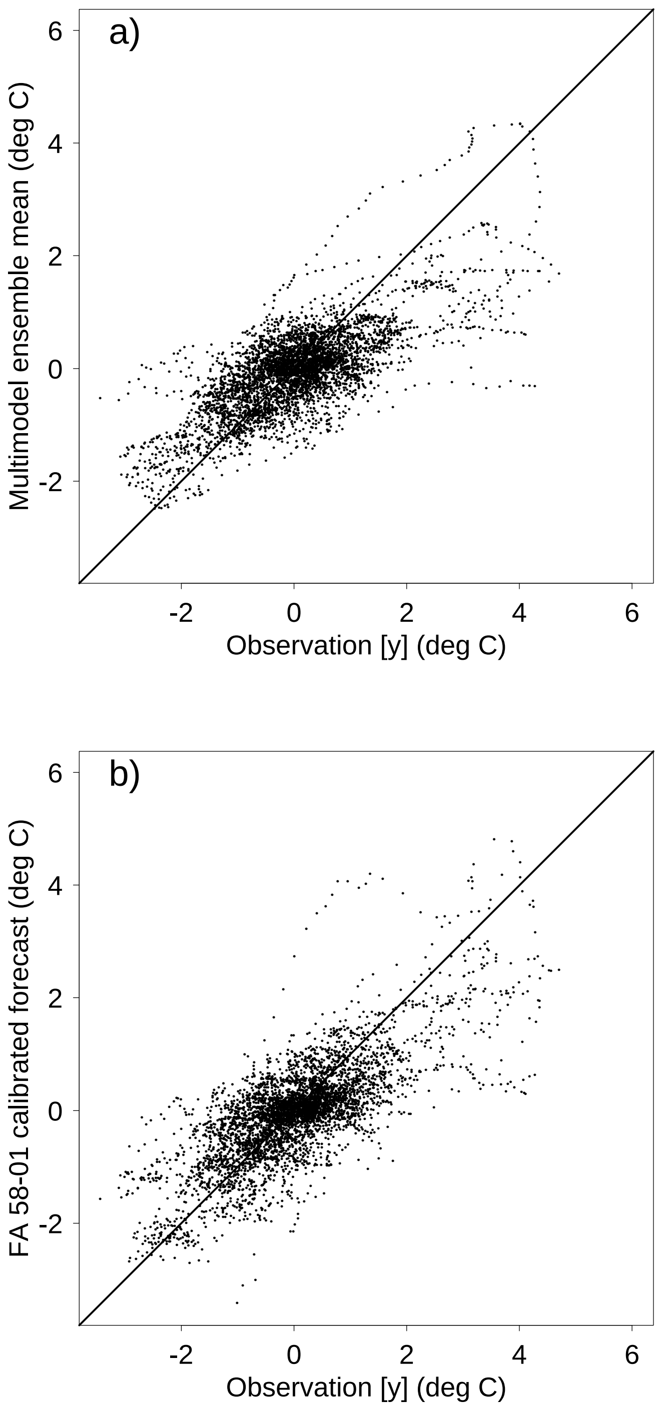


Figure 5.6: Scatter plots of combined forecasts versus observed sea surface temperature anomalies (^oC): a) multi-model ensemble mean forecast, b) forecast assimilation combined forecast with prior estimated using data from 1958-2001

of retained modes. For less than five modes, MCA outperforms CCA by producing forecasts with smaller RMSE and Brier scores. The CCA modes are generally noisier both spatially and temporally than the MCA modes. For more modes, MCA and CCA lead to similar forecast scores. The smallest RMSE and Brier scores are obtained using MCA with 3 modes and the scores are not highly sensitive to the addition of more modes. The three leading MCA modes consist of a basin-wide pattern, an east-west dipole over the central-east equatorial Pacific basin, and an east-west basin-wide dipole (not shown). A large fraction of the squared covariance between observations and model predictions (99.7%) is explained by the three leading MCA modes.

Figure 5.6 shows scatter plots of the combined forecast SST anomalies versus the observed SST anomalies for a) the ensemble mean forecasts, and b) the 1958-2001 forecast assimilation combined forecasts. The cloud of forecast assimilation forecasts lies much closer to the diagonal $y = x$ line than does the cloud of ensemble mean forecasts. Some positive skewness can also be discerned in the scatter plots (i.e. more positive than negative anomaly points) caused by the inherent positive skewness in eastern equatorial Pacific SSTs (Burgers and Stephenson 1999; Hannachi et al. 2003; Hannachi et al. 2004). A more sophisticated forecast assimilation model could be developed to take account of this deviation from normality (e.g. by using skewed probability models). The string of outlier points at warm temperatures is caused by warm ENSO events such as the 1997/98 event extending across the equatorial Pacific.

5.5 Summary

Forecast assimilation has been tested for the calibration and combination of DEMETER multi-model equatorial Pacific SST predictions. The resulting combined forecasts reproduce well the temporal and longitudinal variations observed in equatorial Pacific sea surface temperatures. In this example, the combined deterministic forecast obtained with forecast assimilation resembles the multi-model ensemble forecast of the mean. The ensemble mean works well in this case because the majority of the model predictions are for the most part rather similar to one another and closely resemble the observations. Applications with more disparate model predictions are likely to show enhanced skills for the combined forecasts compared to the simple approach based on equal-weight averaging the ensemble mean predictions.

Forecast assimilation has been shown to improve the skill of probabilistic forecasts. This is because forecast assimilation provides better estimates of prediction uncertainty than the simple multi-model prediction. The prediction uncertainty estimated by forecast assimilation is more realistic than the prediction uncertainty estimated by the standard deviation of the multi-model ensemble mean predictions. As a result of this the Brier score of predictions obtained with forecast assimilation is smaller (i.e. better) than of multi-model ensemble predictions. It is important for risk assessment purposes that climate forecasts are able to provide good estimates of forecast uncertainty in addition to providing forecasts of the mean. Forecast assimilation improved reliability of the predictions in the western Pacific and resolution in the eastern Pacific. The approach is easily applied to 2-dimensional gridded data. An example of this is given in chapter 6 for South American rainfall forecasts.

Chapter 6

A two-dimensional example: South American rainfall forecasts

6.1 Aim

The aim of this chapter is to produce improved (i.e. reliable and well-calibrated) Dec-Jan-Feb (DJF) South American rainfall probability forecasts. Section 6.2 provides an overview with a brief review of the literature on South American rainfall seasonal forecasting. Section 6.3 summarizes climatological features and some key elements of the climate system that influence South American rainfall. Particular attention is directed to understanding the relationship between SSTs and rainfall. This knowledge is used in section 6.4 that introduces an empirical model that uses Aug-Sep-Oct (ASO) Pacific and Atlantic SSTs as predictors for DJF South American rainfall. Section 6.5 tests the use of forecast assimilation for the calibration and combination of DJF South American forecasts. Forecast assimilation is performed using 1-month lead grid point forecasts of three DEMETER climate models (ECMWF, CNRM and UKMO) covering the period 1959-2001. Section 6.6 assesses the skill of empirical and combined and calibrated rainfall forecasts. The skill of combined and calibrated forecasts obtained with forecast assimilation is compared to the skill of both the simple multi-model forecast, produced by pooling/averaging together the forecasts of the three DEMETER climate models, and the empirical model that uses ASO SSTs as predictors. Section 6.7 presents an example of application of forecast assimilation for river flow forecasting. And finally, section 6.8 summarizes the main findings of the chapter.

6.2 Introduction

As in any other part of the world, South American rainfall seasonal forecasting is experimental. This is because our present knowledge about the climate system and its complex interactions is still far from comprehensive. This lack of knowledge is translated into climate models via simplifications/parameterisations of processes that are not yet fully understood. Although boundary conditions can provide some predictability of the atmosphere on seasonal time scales, the inherent variability of the atmosphere causes seasonal climate forecasts to be probabilistic.

South American rainfall seasonal forecasts are generally produced using two approaches: physically-derived dynamical modelling and empirically based (statistical) modelling. Several studies have used atmospheric GCMs forced with observed SSTs to simulate seasonal rainfall over South America (e.g. Folland *et al.* 2001; Cavalcanti *et al.* 2002; Marengo *et al.* 2003; Moura and Hastenrath 2004). These studies have demonstrated that atmospheric GCMs forced with observed SSTs have some predictive skill when forecasting rainfall in the tropical region of South America and over the south region of Brazil, Uruguay, Paraguay and northeast Argentina with all the other areas of South America presenting poor predictive skill. They all found that forecast skill is highly conditioned on the manifestation of ENSO events, with neutral years having less predictive skill. Both tropical South America and the south region of Brazil, Uruguay, Paraguay and northeast Argentina have strong ENSO signals (see Fig. 2.1).

Studies by Pezzi *et al.* (2000), Folland *et al.* (2001), Greischar and Hastenrath (2000) and Martis *et al.* (2002) have developed empirical models relating observed rainfall with SSTs over the Atlantic and Pacific oceans as well as the meridional surface wind component over the tropical Atlantic to predict seasonal rainfall over the south and northeast regions of Brazil and the Netherland's Antilles. Empirical models have been primarily developed for these regions because of the higher predictability of these regions compared to the other areas of South America. Empirically based rainfall predictions for the northeast region of Brazil are skillful during the period Mar-Apr-May (MAM), which is the rainy season for most parts of this region (Greischar and Hastenrath 2000; Folland *et al.* 2001; Moura and Hastenrath 2004). The empirical predictions of Pezzi *et al.* (2000) for the south of Brazil are generally less skillful than the predictions for the northeast region of Brazil, and El Niño years were found to be more predictable than neutral and La Niña years.

The comparative skill of physically-derived dynamical and empirically based seasonal forecasts of South American rainfall is not entirely known, and further systematic comparisons are desirable (Moura and Hastenrath 2004). Only a few comparison studies (Folland *et al.* 2001; van Oldenborgh *et al.* 2003; and Moura and Hastenrath 2004), focussing on rainfall forecasts for South America, have been carried out. The study by van Oldenborgh *et al.* (2003) concluded that physically-derived dynamical predictions outperform empirical predictions over tropical South America, northeast Brazil and Uruguay. Folland *et al.* (2001) and Moura and Hastenrath (2004), which focussed on rainfall forecasts for the northeast Brazil, both concluded that physically-derived dynamical predictions do not outperform empirically based predictions. Their conclusions are in accordance with other comparative skill assessment studies for other target regions outside South America (e.g. Barnston *et al.* 1999a; Anderson *et al.* 1999). This chapter (section 6.6) will contribute to this skill assessment exercise. It will compare the skill of an empirical model that uses observed Pacific and Atlantic SST anomalies of ASO to predict rainfall anomalies of the following DJF, with the skill of DEMETER coupled multi-model rainfall anomaly predictions for DJF produced with initial conditions of the 1st of November. The empirical model developed in this study differs from those of previous studies (Greischar and Hastenrath 2000; Pezzi *et al.* 2000; Folland *et al.* 2001; Moura and Hastenrath 2004) in the sense that it predicts rainfall anomalies for the entire South American continent, while the other studies focussed only on sub-regions.

Combining the predictions from these two approaches might yield better estimates of future climate. Section 6.6 of this chapter will examine this issue by combining physically-derived coupled model and empirically based predictions of DJF South American rainfall anomalies. The skill of the combined forecast can then be compared with the skill of each individual forecasting approach.

Good quality seasonal forecasts are fundamental for local governments to plan their actions in order to minimize human and economical losses that may be caused by anomalous climate events such as ENSO. In South America these forecasts are useful for civil defence, agricultural, fishery and water resources (reservoir management) planning. Brazil, the largest and most populated country of South America, produces more than 90% of its electricity with hydropower stations (http://www.ons.org.br), emphasising the need for good quality seasonal rainfall forecasts. The provision of improved seasonal rainfall forecasts will certainly help the Brazilian government to better plan its management actions in order to have a more efficient control of its national electricity production program.

Despite the recognised importance of good quality seasonal forecasts, no study has been published on statistical calibration of South America physically-derived climate model seasonal forecasts using past observations in order to improve the quality of the forecasts. Most studies (e.g. Cavalcanti *et al.* 2002; Marengo *et al.* 2003; Moura and Hastenrath 2004) investigated the ability of atmospheric GCMs forced with observed SSTs in simulating atmospheric climatological features such as the annual and seasonal cycles of rainfall for some regions of South America. These studies have identified systematic forecast errors, which have not been corrected to improve the forecasts. These errors arise from a combination of factors such as the chaotic evolution of the atmosphere, errors in the initial conditions of the model, and errors in model formulation/parameterisation. In order to produce well-calibrated and more reliable estimates of rainfall for South America this chapter uses the probabilistic Bayesian forecast assimilation procedure – introduced in Chapter 3 (section 3.4) – for the calibration/combination of the ensemble outputs of

DJF rainfall anomaly predictions produced by three DEMETER coupled models. The resulting calibrated/combined forecasts are summarized by the mean and the variance of a normal distribution at each grid point. In fact the resulting forecast is characterised by the mean and covariance of all grid points.

6.3 Climatology

6.3.1 Seasonal rainfall

South American seasonal rainfall is modulated by a few elements of the climate system. Among these elements are: *a)* the Intertropical Convergence Zone (ITCZ), a zonally oriented band of atmospheric convective activity that is observed in the tropics; *b)* the South Atlantic Convergence Zone (SACZ), a northwest-southeast oriented band of atmospheric convective activity that is observed over South America during DJF (Kodama 1992); *c)* Mesoscale Convective Systems (MCS), which are observed during DJF and MAM over subtropical South America (Velasco and Fritsch 1987); *d)* frontal systems originated in high and mid-latitudes of the Pacific ocean, which advect heat and humidity from the Pacific ocean to continental South America; and *e)* easterly trade winds resulting from the equatorial branches of the North and South Atlantic subtropical highs, which advect moisture from the Atlantic ocean to the continent.

Figure 6.1 shows the climatological mean rainfall for the four seasons of the year (Dec-Jan-Feb (DJF), Mar-Apr-May (MAM), Jun-Jul-Aug (JJA) and Sep-Oct-Nov (SON)). Fig. 6.1a shows a northwest-southeast oriented band of rainfall, with maximum over central South America, reflecting a rainfall pattern associated with the SACZ. Fig. 6.1b shows a zonally oriented band of rainfall in tropical South America, illustrating the rainfall pattern associated with the ITCZ. Fig. 6.1c shows that JJA is the dry season for most of the South America continent. Fig. 6.1d illustrates the pattern of rainfall primarily produced by South Pacific frontal systems.

Figure 6.2 shows the standard deviation of seasonal rainfall that is an indicator

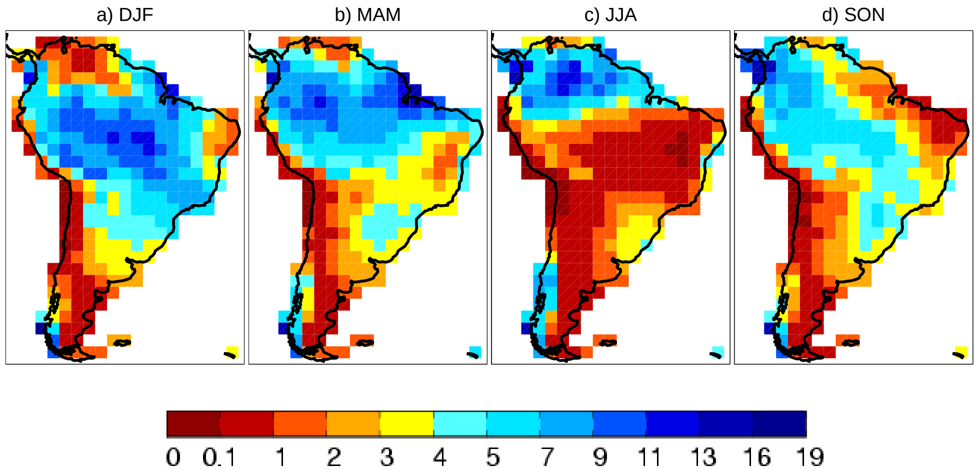


Figure 6.1: Seasonal mean rainfall (1948-2000) in $mm.day^{-1}$ from PREC/L V1.0 dataset (see Appendix C). a) DJF; b) MAM; c) JJA; and d) SON.

of rainfall variability. More variability is observed during DJF (Fig. 6.2a) and MAM (Fig. 6.2b) than in JJA (Fig. 6.2c) and SON (Fig. 6.2d). This is because DJF and MAM rainfall is strongly modulated by convective systems while JJA and SON rainfall is primarily produced by frontal systems. The dry season (JJA) has the smallest variability (Fig. 6.2c). Note that in DJF and MAM the southeast region of South America (formed by the south of Brazil, Uruguay, Paraguay and northeast Argentina) has a second maximum of rainfall variability, which is related to the frequent manifestation of MCS over this region during these periods.

Figure 6.3 shows the percentage of the annual rainfall observed during the four seasons of the year. More than 50% of the annual rainfall is observed during DJF (Fig. 6.3a), indicating that this is the rainy season for most of South America. Between 20 and 40% of the annual rainfall is observed during SON for most of South America (Fig. 6.3d). MAM is the season when most of South America observes between 10 and 20% of its annual rainfall (Fig. 6.3b). JJA is the dry season when most of South America (except Chile and Venezuela) experiences less than 10% of the annual rainfall (Fig. 6.3c).

From the exposition so far, DJF is the season when most of South America receives most of its rainfall. Therefore, good quality predictions of DJF seasonal

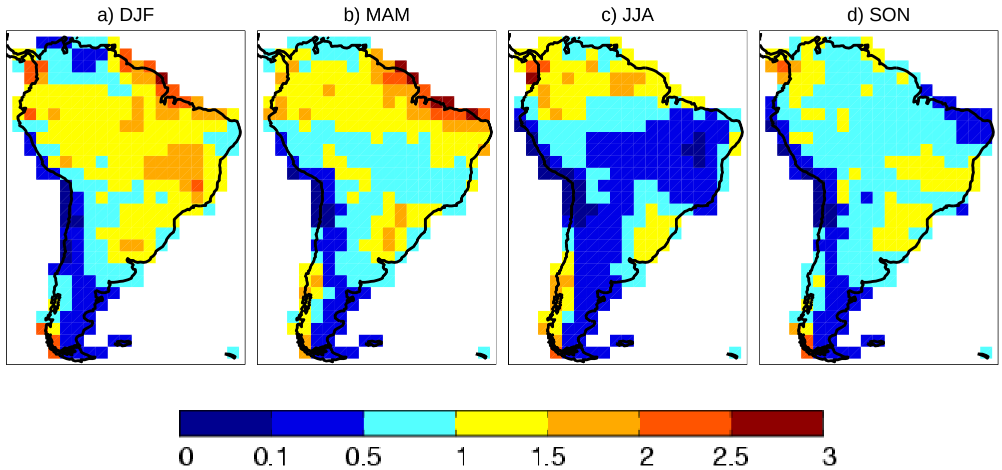


Figure 6.2: Interannual standard deviation of seasonal mean rainfall (1948-2000) in $mm.day^{-1}$ from PREC/L V1.0 dataset. a) DJF; b) MAM; c) JJA; and d) SON.

rainfall are crucial for all social and economical sectors that depend on seasonal rainfall for future planning and decision making. Hence, the next sections of this chapter will focus on analyses and predictions of DJF seasonal rainfall.

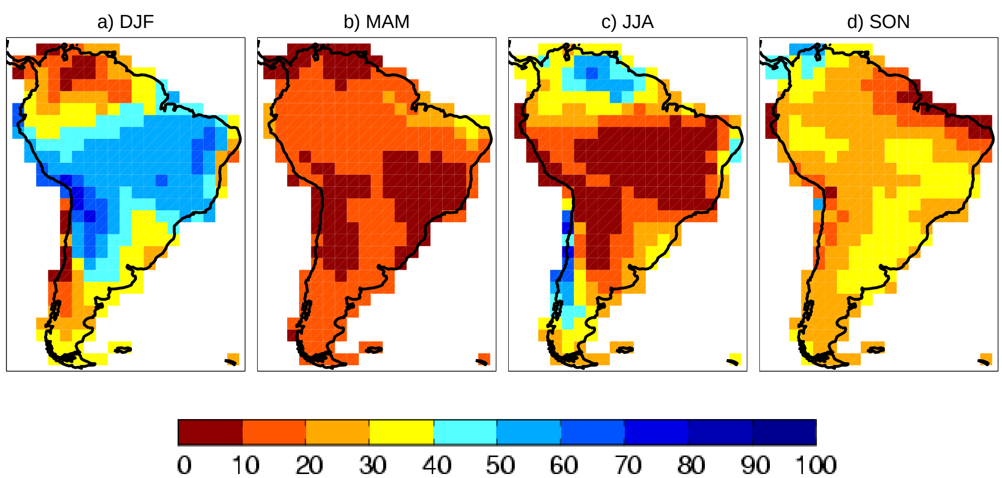


Figure 6.3: Percentage of the total annual rainfall (1948-2000) from PREC/L V1.0 dataset. a) DJF; b) MAM; c) JJA; and d) SON.

6.3.2 The relationship between SSTs and rainfall

South America is bordered by the Pacific and Atlantic oceans. Surface conditions of these two oceans are potential sources of predictability for South American climate (Moura and Shukla 1981; Mechoso *et al.* 1990; Marengo 1992; Nobre and Shukla 1996; Diaz *et al.* 1998; Uvo *et al.* 1998; Barros and Silvestri 2002; Coelho *et al.* 2002; Peagle and Mo 2002 among others). These studies identified regions of South America that are sensitive to SST anomalies in the Pacific and Atlantic oceans. This section uses MCA (section 3.4.2) to examine the relationship between Aug-Sep-Oct (ASO) Pacific and Atlantic SST anomalies (140^{o}E-10^{o}E; 15^{o}N-60^{o}S) and South American rainfall of the following DJF during 1959-2001. The study of lagged relationships like the one that is investigated here is useful for empirical predictions (e.g. the previous season ASO SST anomalies can be used as predictors for DJF rainfall anomalies). The previous season ASO is used here for consistency with the initial conditions of the 1st of November that are used by the three DEMETER coupled models to predict DJF rainfall.

Figure 6.4 shows the loadings (spatial patterns) and the expansion coefficients (time series) of the first mode of the MCA analysis between the observed ASO SST anomalies over the Pacific and Atlantic oceans and the observed DJF South American rainfall anomalies using data for the period 1959-2001. This mode explains a large amount (63.9$\%$) of the squared covariance between ASO SST and DJF rainfall. The SST pattern (Fig. 6.4a) shows basin-wide positive loadings in the equatorial Pacific, revealing that this mode is related to ENSO. Warm (El Niño) years are depicted by positive peaks in the time series of Fig. 6.4c and cold (La Niña) years are marked by minima in these time series. The rainfall pattern (Fig. 6.4b) has negative loadings over northern South America and south Chile and positive loadings over the east and south Brazil, Uruguay, Paraguay, northern Argentina and Ecuador. This figure reveals a dipole pattern that during El Niño years is marked by deficit of rainfall in northern South America and excess of rainfall in the east and south Brazil, Uruguay, Paraguay and northern Argentina. During La Niña years

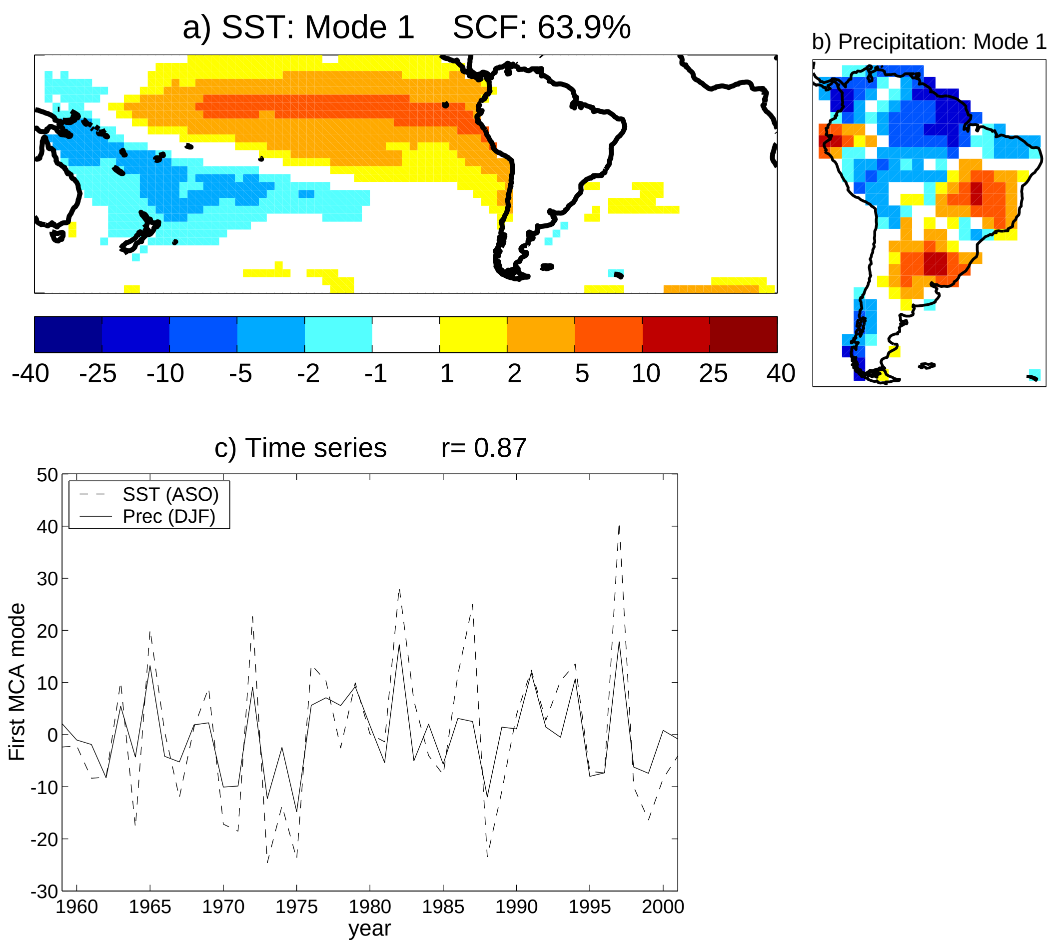


Figure 6.4: First MCA mode between ASO SST anomalies and DJF South America rainfall anomalies for the period 1959-2001. The squared covariance fraction (SCF), which indicates the percentage of the total squared covariance between ASO SST anomalies and DJF South America rainfall anomalies explained by this mode, is 63.9%. a) SST loadings. b) Rainfall loadings. c) Expansion coefficients (time series) of SST (dashed line) and rainfall (solid line). The correlation r between these two time series is indicated in panel c.

this pattern is reversed. A similar ENSO pattern to Fig. 6.4 has been identified by Ropelewski and Halpert (1987; 1989), Kiladis and Diaz (1989) and Peagle and Mo (2002).

Figure 6.5 shows composites of DJF equatorial Pacific SST anomalies and DJF South American rainfall anomalies for La Niña, neutral and El Niño events between 1959 and 2001 as defined by the Climate Prediction Center (http://www.cpc.noaa.gov/). These events are listed in Table 6.1. In total there are 13 La Niña, 14 neutral and 16 El Niño events.

	Years
La Niña	1964/65, 1970/71, 1971/72, 1973/74, 1974/75, 1975/76, 1983/84, 1984/85, 1988/89, 1995/96, 1998/99, 1999/00, 2000/01
Neutral	1959/60, 1960/61, 1961/62, 1962/63, 1966/67, 1967/68, 1978/79, 1980/81, 1981/82, 1985/86, 1989/90, 1993/94, 1996/97, 2001/02
El Niño	1963/64, 1965/66, 1968/69, 1969/70, 1972/73, 1976/77, 1977/78, 1979/80, 1982/83, 1986/87, 1987/88, 1990/91, 1991/92, 1992/93, 1994/95, 1997/98

Table 6.1: La Niña, neutral and El Niño years occurred during 1959-2001.

SST anomalies in the equatorial Pacific range between -0.5^oC and -1.5^oC for the La Niña composite (Fig. 6.5a), between -0.5^oC and 0.5^oC for the neutral composite (Fig. 6.5b) and between 0.5^oC and 1.5^oC for the El Niño composite (Fig. 6.5c). The dipole pattern of rainfall of Fig. 6.4b is reproduced by the El Niño composite of rainfall anomalies (Fig. 6.5f). Negative anomalies are observed in northern South America and positive anomalies in the east and south Brazil, Uruguay, Paraguay and northern Argentina. The La Niña composite (Fig. 6.5d) shows a pattern with positive rainfall anomalies in northern South America and negative anomalies in the east Brazil. The second centre of anomalies in south Brazil, Uruguay, Paraguay and northern Argentina is not observed in the composite of La Niña years. This is in accordance with the findings of Barros and Silvestri (2002) that suggest that during La Niña years this region is sensitive to SST

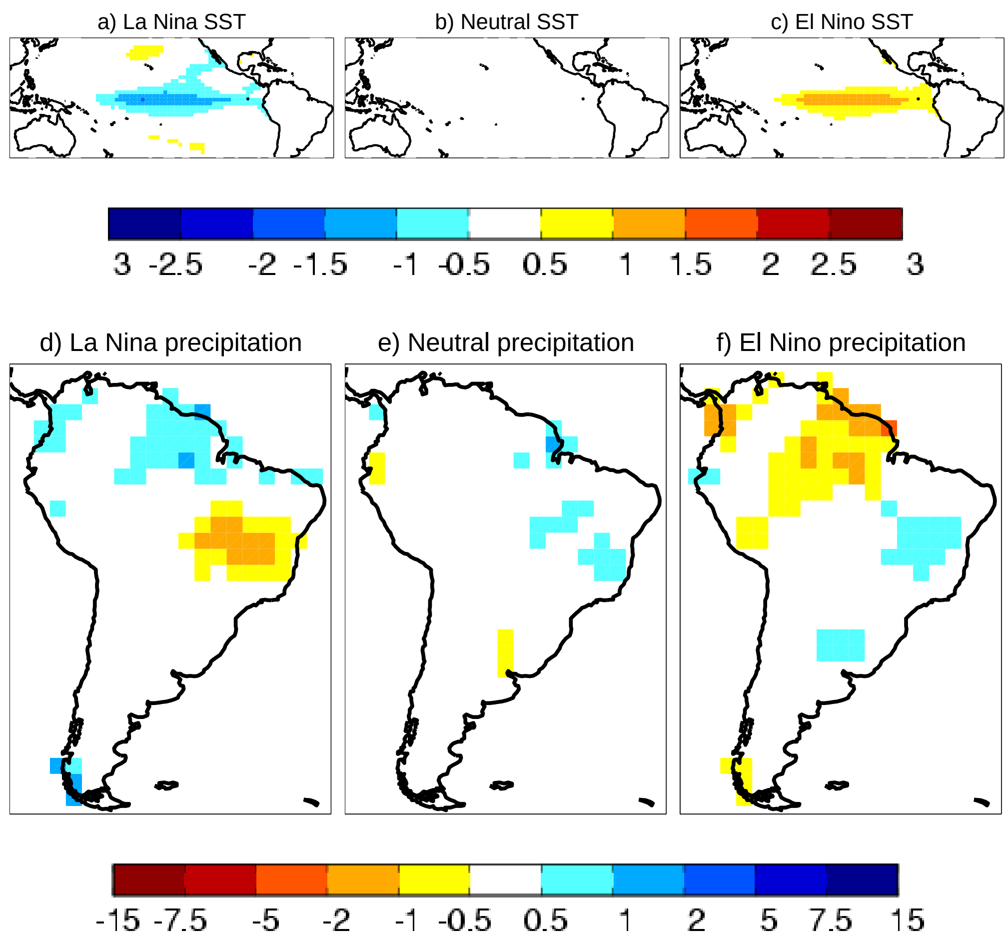


Figure 6.5: Composites of equatorial Pacific SST anomalies (^oC) and South American rainfall anomalies ($mm.day^{-1}$) for La Niña (panels a and d), neutral (panels b and e) and El Niño years (panels c and f) between 1959-2001. See Table 6.1 for the years used in each composite.

anomalies of the subtropical southwest Atlantic instead of equatorial Pacific SST anomalies. The link between rainfall in south Brazil, Uruguay, Paraguay and northern Argentina and SSTs in the subtropical Southwest Atlantic will be discussed in more detail later. The neutral years rainfall anomalies composite (Fig 6.5e) shows positive anomalies in northeast South America and negative anomalies in northwest South America and northeast Argentina.

The lack of rainfall in northern South America during El Niño years is a direct response to changes in the Walker circulation (Coelho *et al.* 2002; Peagle and Mo 2002). The increased convective activity observed over the anomalously warm equatorial waters near the west coast of South America during El Niño years has a direct response over the north of South America, where compensatory subsidence prevails and inhibits rainfall. The increased rainfall in south Brazil, Uruguay, Paraguay and northern Argentina is a teleconnection response to the anomalous convective activity observed in central equatorial Pacific during El Niño years. This teleconnection pattern receives the name Pacific South American (PSA) due to its analogy with the Pacific North American (PNA) teleconnection pattern (Wallace and Gutzler 1981; Karoly 1989). The PSA and the PNA teleconnection patterns are illustrated in Figs. 6.6a and 6.6b, respectively. The PSA consists of a high level wave train of alternated anti-cyclone and cyclone centres emanating from the anomalous convective heat source in the equatorial central Pacific, which curves towards subtropical South America. The wave train starts over the Pacific ocean with two anti-cyclonic circulations at low latitudes of both hemispheres centred east of the date line. It extends to the southeast down to the west of the Antarctic Peninsula, where it turns to the northeast towards the east South America. At subtropical latitudes, this implies an enhanced subtropical jet over most of the Pacific ocean. This wave train was documented by Karoly (1989) and attributed to barotropic Rossby wave propagation. At low levels in the Atlantic sector, there is an increased pressure gradient between the Chaco low (a cyclonic centre observed in subtropical South America) and the South Atlantic semipermanent subtropical high. This increased gradient results in a westward displacement and strengthening of the easterly flow from the northern branch of the Atlantic subtropical high. This contributes to an increased moisture supply over eastern Brazil and increased rainfall over this region. This easterly low-level flow is deflected southwards by the Andes and advects additional moisture from the Amazon rainforest to subtropical latitudes, contributing to the rainfall maximum over northern Argentina, Uruguay, Paraguay and south Brazil (Grimm *et al.* 2000).

Figure 6.7 shows the second MCA mode that explains 10.3% of the squared covariance between ASO SST and DJF rainfall. The SST pattern (Fig. 6.7a) shows

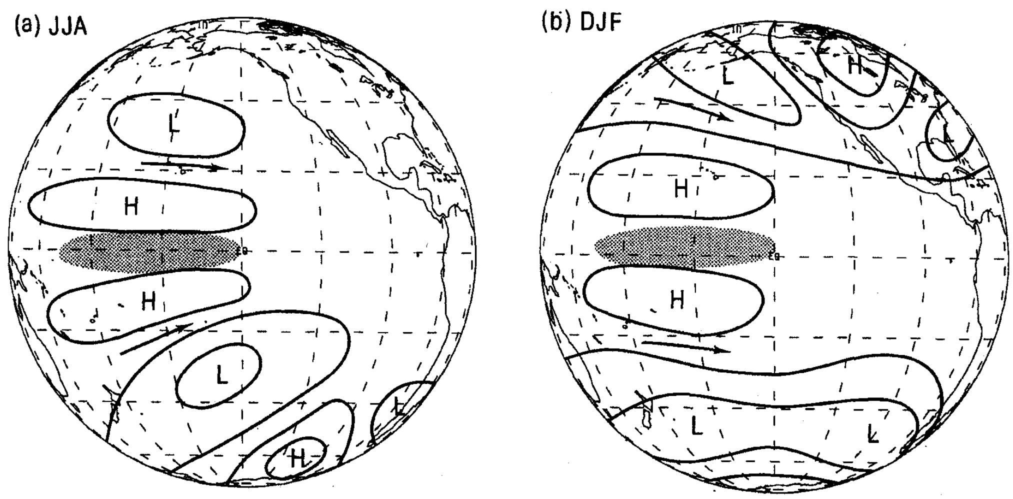


Figure 6.6: Schematic illustration of the upper troposphere circulation anomaly pattern over the Pacific ocean during a) the early stage of an El Niño event (JJA) and b) the mature stage of an El Niño event (DJF). The stippling shows the region of enhanced convection over the central equatorial Pacific and the arrows indicate the westerly wind anomalies in the jet streams. The letters H and L indicate anomalous centres of high and low pressure, respectively. Source: Karoly (1989).

three structures that are worth discussing: costal ENSO, equatorial Atlantic dipole and SSTs off the coast of south Brazil, Uruguay and Argentina. The first structure of Fig. 6.7a is found in the equatorial Pacific, with negative loadings near the west coast of South America resembling a costal ENSO signature (Fig. 6.7a). SST anomalies in this region produce rainfall anomalies with the same signs as those in neighbouring northwest South America, as illustrated in Fig 6.7b.

The second structure of Fig. 6.7a is the dipole in the equatorial Atlantic with positive loadings in the equatorial South Atlantic and negative loadings in the equatorial North Atlantic. This dipole has previously been reported by Moura and Shukla (1981), Nobre and Shukla (1996) and Uvo *et al.* (1998). The interhemispheric gradient of SST in the equatorial Atlantic modulates the meridional position of the ITCZ. Enhanced deep convection occurs over oceanic regions with positive SST anomalies. Warm SSTs in the equatorial South Atlantic and increased evaporation over the ocean, combined with the increased moisture flux over the northeast

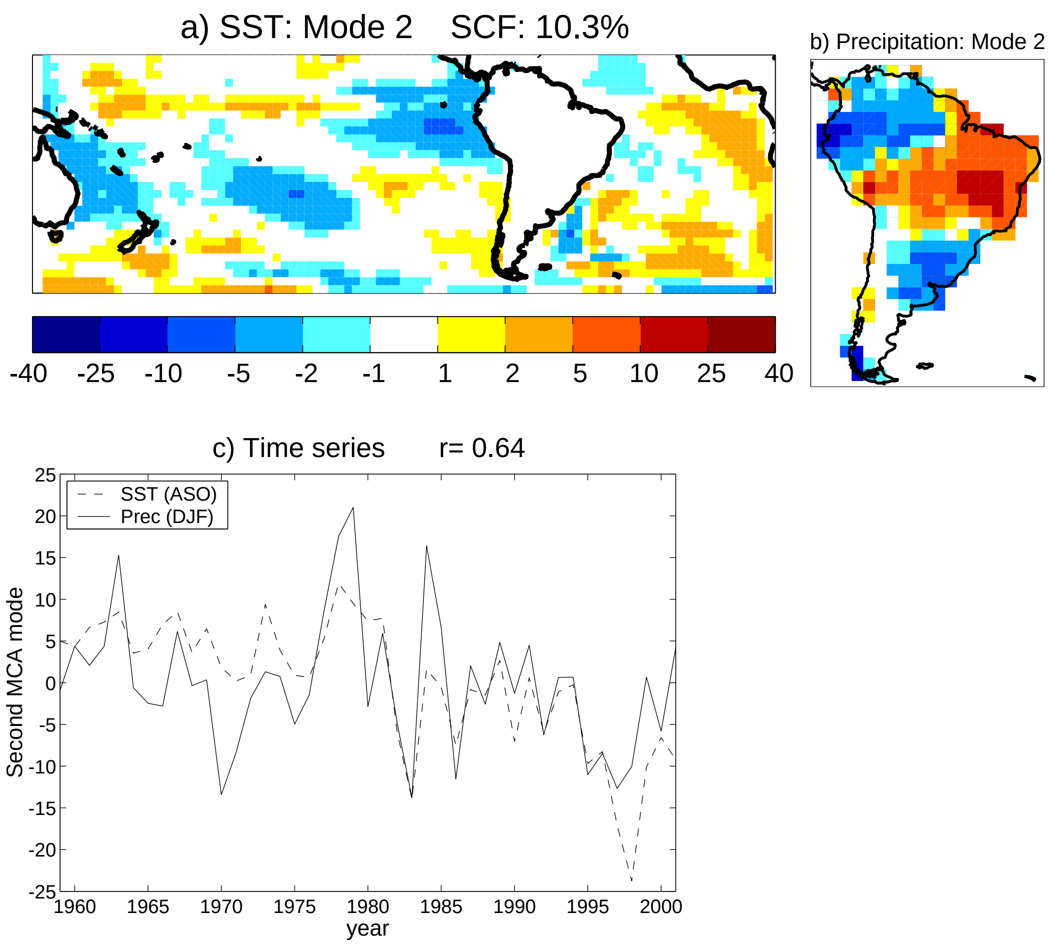


Figure 6.7: Second MCA mode between ASO SST anomalies and DJF South America rainfall anomalies for the period 1959-2001. The squared covariance fraction (SCF) is 10.3%. a) SST loadings. b) Rainfall loadings. c) Expansion coefficients (time series) of SST (dashed line) and rainfall (solid line). The correlation r between these two time series is indicated in panel c.

region of Brazil – produced by the equatorial southeasterly trade winds – result in increased rainfall over northeast Brazil. In contrast, when SSTs are warm in the equatorial North Atlantic and cold in the equatorial South Atlantic, evaporation is increased over the North Atlantic ocean resulting in enhanced convective activity in the ITCZ. Ascending vertical motion is observed in the North Atlantic over the region of warm SSTs and descending motion (subsidence) is observed in the northeast region of Brazil and the neighbouring oceanic areas. This link between SSTs in the equatorial Atlantic and rainfall in northeast Brazil is illustrated in Fig. 6.7a and 6.7b. Loadings in Northeast Brazil and in the equatorial South Atlantic have the same signs, but opposite signs to loadings in equatorial North Atlantic. This is in agreement with the physical explanations given above. A similar pattern to Fig. 6.7b has been identified by Peagle and Mo (2002) as the response to the equatorial Atlantic dipole.

The third structure of Fig. 6.7a is found in the subtropical southwest Atlantic, with negative loadings near the coast of southern Brazil, Uruguay and Argentina (Fig. 6.7a). SST anomalies in this region produce rainfall anomalies with the same signs as those in neighbouring south Brazil, Uruguay, Paraguay and northeast Argentina as illustrated in Fig 6.7b. Frontal systems that reach the subtropical southwest Atlantic are intensified when SSTs over this region are above normal, leading to above normal rainfall over south Brazil, Uruguay, Paraguay and northeast Argentina. This link has previously been documented by Diaz *et al.* (1998) and Barros and Silvestri (2002).

Figure 6.8 shows the third MCA mode that explains only 5.5% of the squared covariance between ASO SST and DJF rainfall. The SST pattern (Fig. 6.8a) shows negative loadings in a large area of the Atlantic. The pattern is linked to the pattern of negative loadings in the northeast Brazil (Fig. 6.8b). This indicates that positive rainfall anomalies are observed in the northeast Brazil when the equatorial Atlantic is dominated by positive SST anomalies, Conversely, negative rainfall anomalies are observed in the northeast Brazil when the equatorial Atlantic is dominated by negative SST anomalies. South Brazil, Uruguay, Paraguay and northern

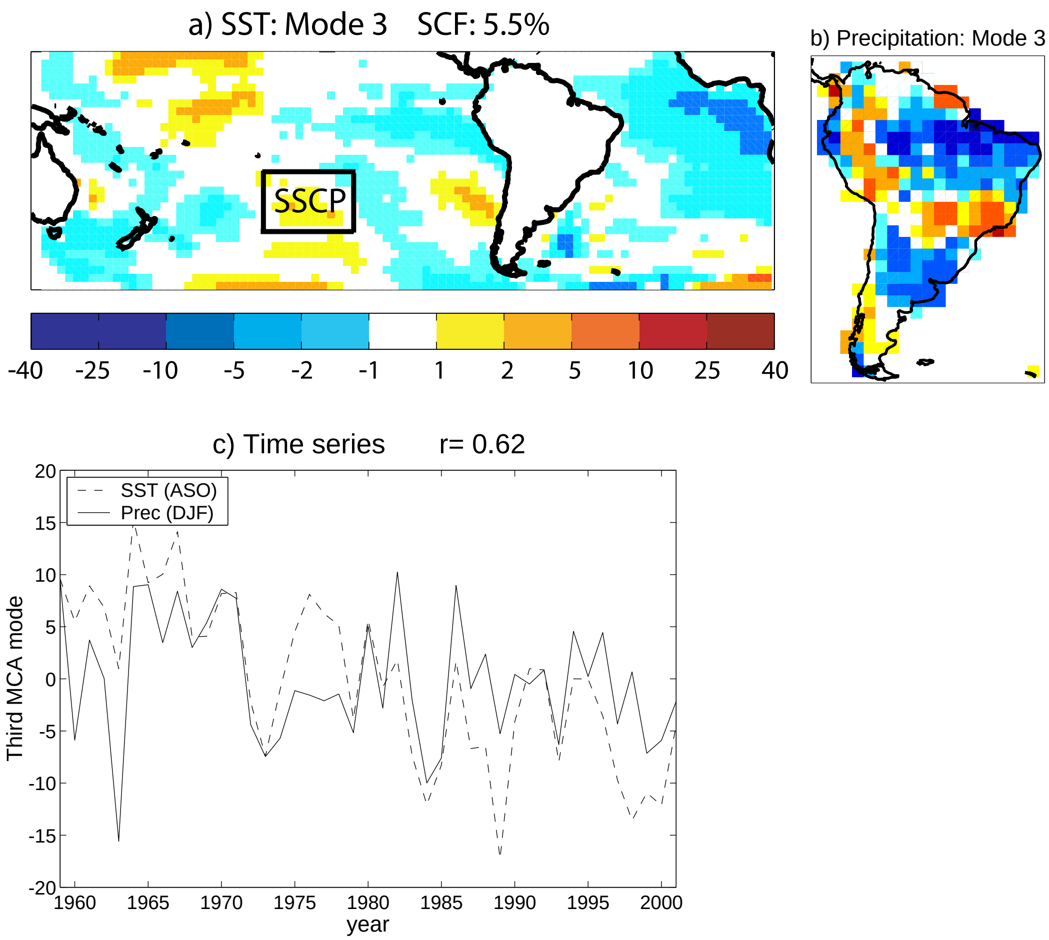


Figure 6.8: Third MCA mode between ASO SST anomalies and DJF South America rainfall anomalies for the period 1959-2001. The squared covariance fraction (SCF) is 5.5%. a) SST loadings. b) Rainfall loadings. c) Expansion coefficients (time series) of SST (dashed line) and rainfall (solid line). The correlation r between these two time series is indicated in panel c.

Argentina also have negative loadings (Fig. 6.8b). Barros and Silvestri (2002) showed that rainfall variability in this region is not only modulated by ENSO as illustrated by Fig. 6.4. They found that the SSTs in the subtropical south-central Pacific (SSCP), which is marked in Fig. 6.8b, also influence rainfall in south Brazil, Uruguay, Paraguay and northern Argentina. Their findings suggest that the ENSO PSA pattern is not the result of equatorial forcing alone, but is actually caused by the forcing of both the equatorial and the subtropical SST forcing. They found a negative correlation between rainfall in south Brazil, Uruguay, Paraguay and northern Argentina and SST in the SSCP, which is also evident from Figs 6.8a and 6.8b (although not very intense). Positive loadings are found in the SSCP and negative loadings in south Brazil, Uruguay, Paraguay and northern Argentina. Barros and Silvestri (2002) still suggested that the SST in the SSCP has considerable low-frequency variability. Visual inspection of Fig. 6.8c also suggests that the third MCA mode has low-frequency variability. Additionally, expansion coefficients of Fig. 6.8c reveal a downward trend in SST. The subtropical Southwest Atlantic, near the coast of south Brazil and Uruguay also appears with the same sign of rainfall loadings in the south Brazil, Uruguay and northern Argentina (Figs 6.8a and 6.8b), as previously noted in Fig. 6.7.

6.4 Empirical predictions

6.4.1 The empirical model

The existence of relationships between ASO Pacific and Atlantic SST anomalies and South American rainfall anomalies in the following DJF as those found in the previous sections suggests that ASO SST observations could be used as predictors of DJF rainfall. The simplest model for forecasting DJF South American rainfall anomalies uses multivariate linear regression with the preceding Pacific and Atlantic ASO SST anomaly time series at each grid point v as the linear predictors for rainfall at each grid point q over South America. One assumes that the observable rainfall anomalies y given the observable SST anomalies z are (multivariate) normally distributed:

$$y = M(z - z_0) + \epsilon_T \tag{6.1}$$

where ϵ_T is a (multivariate) normally distributed error with zero mean and *empirical prediction error covariance* T. For generality, a bias term z_0 has also been included. The $(q \times v)$ empirical prediction operator M can be estimated by regression of observed SST anomalies on the observed rainfall anomalies. The equation can be rewritten more succinctly as the following probability model:

$$y|z \sim N(M(z - z_0), T). \tag{6.2}$$

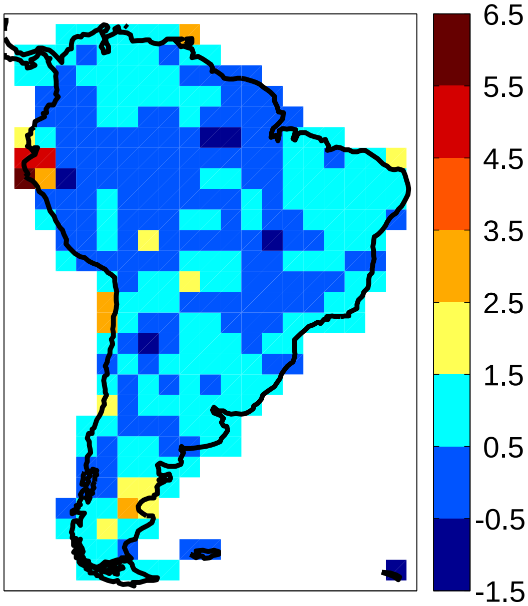


Figure 6.9: Map of skewness of mean DJF rainfall anomalies (1959-2001).

The normality assumption is generally acceptable for seasonal rainfall anomalies. Figure 6.9 shows a moment measure of skewness $b_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{s_y}\right)^3$ of the series of $n = 43$ years (1959-2001) of mean DJF rainfall anomalies. The skewness

b_1 measures the asymmetry of the distribution. Values above zero indicate that the distribution is positively skewed. Values below zero indicate that the distribution is negatively skewed. Values close to zero indicate that the distribution is symmetric (close to normal). Fig. 6.9 shows that most South America has values between -0.5 and +0.5, indicating that DJF seasonal rainfall anomalies are close to following a normal distribution. The normality assumption substantially simplifies both modelling and parameter estimation.

6.4.2 Parameters estimation

For this multivariate normal model one needs estimates of the vector z_0 and matrices T and M. These parameters are obtained by performing a multivariate regression of observed SST anomalies on the observed rainfall anomalies for a period when both SST and rainfall anomaly observations are available. The common period for the example shown in this chapter is 1959-2001 (43 years). The slope M, bias vector z_0, and the empirical prediction error covariance T can be obtained using ordinary least squares estimation:

$$M = S_{yz}S_{zz}^{-1} \tag{6.3}$$

$$z_0 = -(\overline{y} - \overline{z}M^T)M(M^TM)^{-1} \tag{6.4}$$

$$T = S_{yy} - S_{yz}S_{zz}^{-1}S_{yz}^T \tag{6.5}$$

where S_{zz} is the $(v \times v)$ covariance matrix of the observed SST anomalies, S_{yy} is the $(q \times q)$ covariance matrix of the observed rainfall anomalies, and S_{yz} is the $(q \times v)$ cross-covariance matrix.

Reliable parameter estimation is difficult because of the large dimensionality of gridded data sets (e.g. v=2761 grid points of SST anomalies over the Pacific and Atlantic and q=312 grid points of rainfall anomalies over South America) and the strong dependency between values at neighbouring grid points. As previously discussed in chapter 3 (section 3.4.2) poor conditioning of matrices such as S_{zz} makes parameter estimation problematic (or impossible). This problem can be avoided using multivariate dimension reduction techniques to reduce the dimensionality of the data sets. Instead of considering grid point variables, one can project the data onto a small set of spatial patterns to obtain a small number of indices. In this chapter Maximum Covariance Analysis (MCA) has been used to extract leading co-varying modes from the observed ASO Pacific and Atlantic SST anomalies and the observed South American rainfall of the following DJF. MCA with up to 8 retained modes has been tested. It was found that MCA with 6 modes gave the best cross-validated forecast results, which are shown in this chapter. Figure 6.10 shows the squared covariance fraction (SCF) as a function of the number of modes for the MCA between observed ASO SST anomalies in the Pacific and Atlantic and observed rainfall over South America using data for the period 1959-2001. Note that the SCF substantially drops up until 6 modes. These first 6 modes explain 89.3$\%$ of the squared covariance between SST and rainfall anomalies. After 6 modes a very small amount of the squared covariance is explained by each MCA mode. The first 3 modes have been shown and discussed in section 6.3.2 of this chapter (Figs. 6.4, 6.7 and 6.8) and explain a total of 79.7$\%$ of the squared covariance between SST and rainfall anomalies.

The empirical prediction that is presented later in section 6.6 is performed as follows:

1. In order to produce cross-validated forecasts on data not used in the estimation, the year to be forecast is removed from the data set.

2. The time mean is subtracted from the remaining observations to make anomalies stored in a $(n \times q)$ data matrix Y of observed DJF rainfall and a $(n \times v)$ data matrix Z of observed ASO Atlantic and Pacific SSTs, where n is the length of the data time series. In the example of this chapter $n = 42$, $q = 312$ and $v = 2761$.

3. An SVD analysis is performed of the matrix $Y^T Z = U\Sigma^* V^T$ to determine the leading MCA modes.

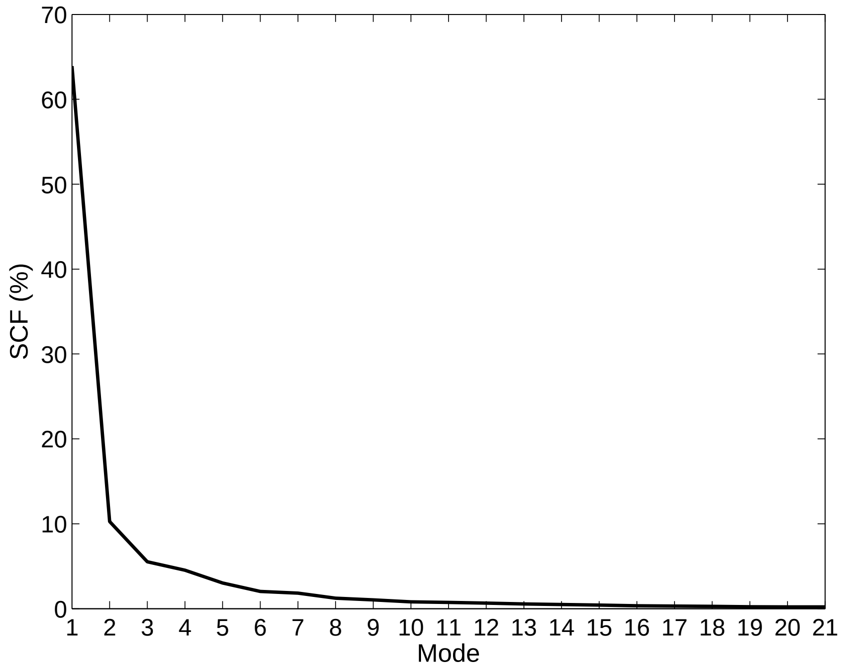


Figure 6.10: SCF as a function of the number of modes for the MCA between observed ASO SST anomalies and observed DJF rainfall anomalies over South America (1959-2001). The first 6 modes explain 89.3$\%$ of the covariance between SST and rainfall anomalies

4. A multivariate regression of the k-leading MCA DJF rainfall anomaly modes on the k-leading MCA ASO SST anomaly modes is performed in order to estimate M, z_0, and T.

5. The estimated quantities M, z_0, and T are then used to forecast DJF rainfall anomalies for the removed year using the observed ASO SST anomalies from that year.

6.5 Forecast assimilation of coupled model predictions

The Bayesian forecast assimilation procedure introduced in chapter 3 (section 3.4) has been used for the calibration and combination of DJF South American rainfall anomaly forecasts produced by three DEMETER coupled models (ECMWF, CNRM and UKMO). This procedure uses forecasts for DJF covering the period

1959-2001 that were produced using initial conditions of the 1st November (1-month lead) and is referred here to as forecast assimilation (FA). A particular usefulness of the forecast assimilation procedure is that it allows predicted patterns to be shifted around in order to correct coupled model predictions. The prior distribution was estimated using observations over the calibration period 1959-2001. As it will be shown later in section 6.6.1, forecast assimilation has also been used to combine forecasts produced by DEMETER coupled models with empirical forecasts produced by the model described in the previous section. In one of the two combination procedures empirical predictions have been used to estimate the prior distribution. The first three leading modes of the MCA between observed and coupled model predictions of DJF South American rainfall anomalies were used in the forecast assimilation procedure. Forecast assimilation was tested with up to 8 retained modes. It was found that forecast assimilation with 3 modes gave the best cross-validated forecast results. These three modes explain 86.5% of the squared covariance between observed and forecast rainfall. Figure 6.11 shows the SCF as a function of the number of modes for the MCA between observed and coupled model predictions of DJF South American rainfall anomalies. Note that the SCF substantially drops up until 3 modes. After 3 modes a very small amount of the squared covariance is explained by each MCA mode. The first three modes are shown in Figs. 6.12, 6.13 and 6.14 and discussed below.

Figure 6.12 shows the loadings (spatial patterns) and the expansion coefficients (time series) of the first mode of the MCA between observed and coupled model predictions of DJF South American rainfall anomalies. This mode explains 65% of the squared covariance between observed and predicted rainfall. The pattern of observed rainfall (Fig. 6.12a) shows a similar pattern to the rainfall pattern of the first MCA mode of Fig. 6.4b, which is related to ENSO. The correlation between the expansion coefficients (time series) of observed rainfall of the first MCA mode (solid line in Fig. 6.12e) and the expansion coefficients (time series) of observed rainfall of the first MCA mode of Fig. 6.4c (solid line) is 0.97. This correlation is statistically significant at the 1% level. Figures 6.12b, 6.12c and 6.12d show the

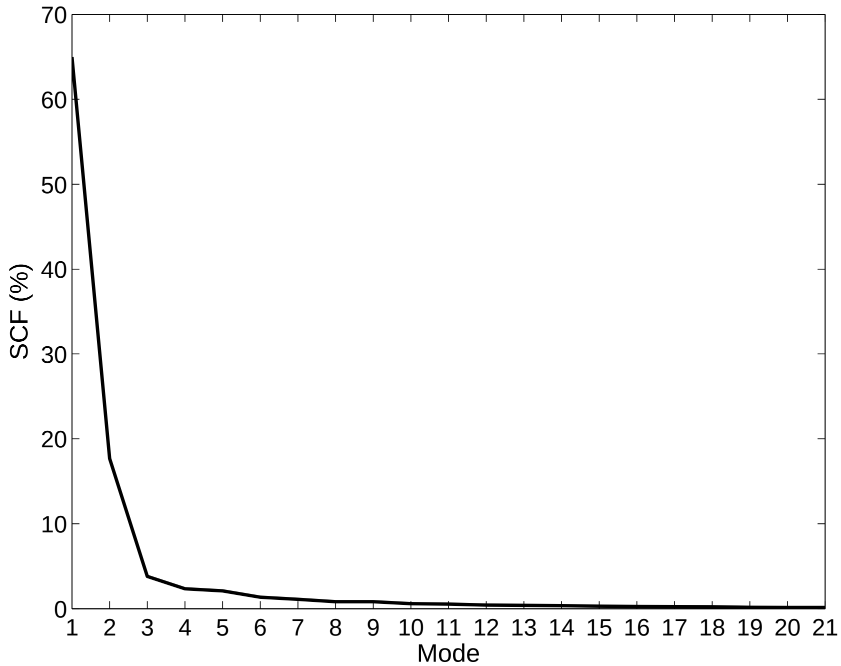


Figure 6.11: SCF as a function of the number of modes for the MCA between observed and coupled model predictions of DJF South American rainfall anomalies (1959-2001). The first 3 modes explain 86.5% of the covariance between observed and predicted DJF rainfall anomalies

spatial patterns of the predictions produced by the three coupled models here investigated (CNRM, ECMWF and UKMO). The spatial structure of these patterns (Figs 6.12b, 6.12c and 6.12d) when compared to the observed pattern (Fig. 6.12a) provides an indication of the ability of these models to reproduce the observed rainfall. The magnitude of the loadings of Figs 6.12b, 6.12c and 6.12d gives an indication of the weights attributed to each model in the forecast assimilation combination procedure. The three models are able to reproduce the rainfall pattern over the central region of northern South America. The pattern of negative loadings in the northwest South America, near Ecuador, is partially captured by the ECMWF and UKMO models, while CNRM fails to reproduce this feature. Over east Brazil CNRM and UKMO partially capture the signal of positive loadings observed in this region but ECMWF does not reproduce this feature. Both ECMWF and UKMO are able to capture the sign of positive loadings in the south Brazil, Uruguay, Paraguay

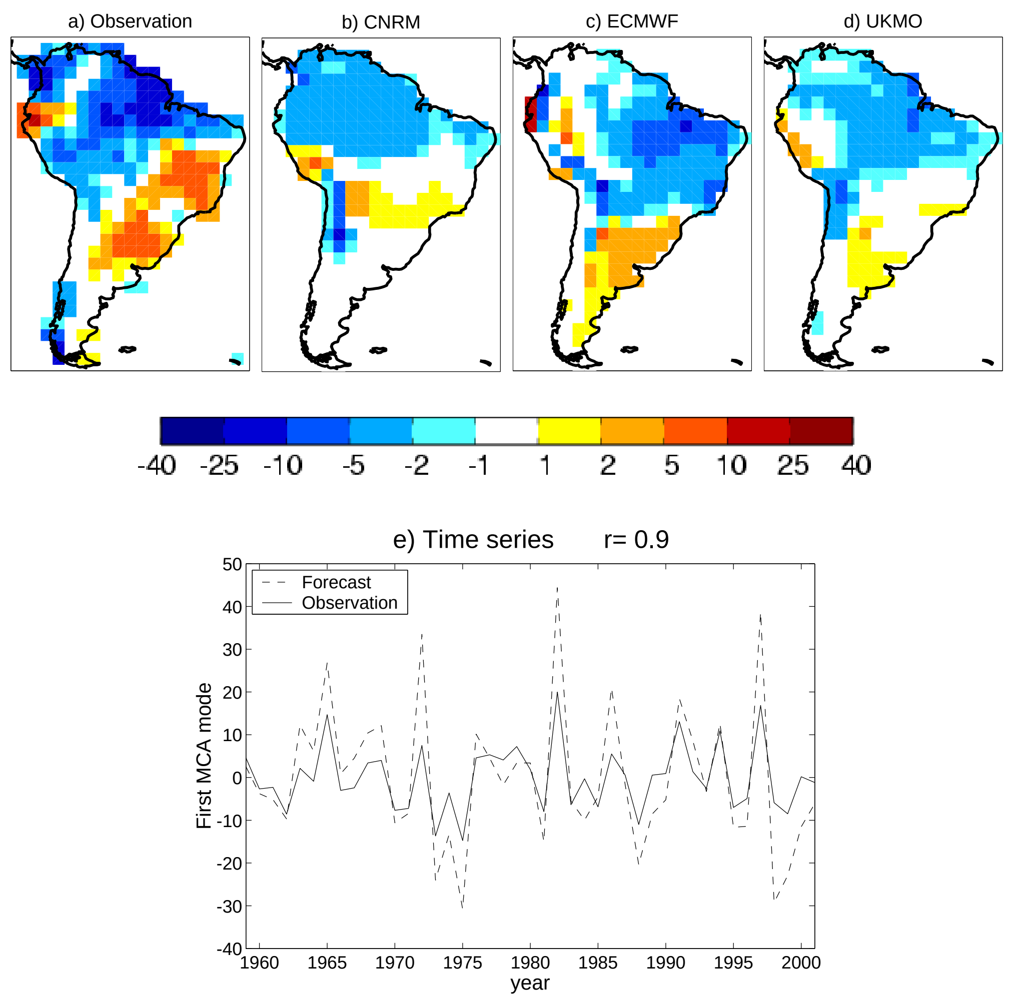


Figure 6.12: First MCA mode between observed and predicted DJF South America rainfall anomalies for the period 1959-2001. The SCF is 65.0%. Spatial patterns (loadings): a) Observation. b) CNRM. c) ECMWF. d) UKMO. e) Expansion coefficients (time series) of observed rainfall (solid line) and predicted rainfall of these three coupled models (dashed line). The correlation r between these two time series is indicated in panel e.

and northern Argentina. This feature is not reproduced by CNRM.

Figure 6.13 shows the second MCA mode that explains 17.7% of the squared covariance between observed and predicted rainfall. The pattern of observed rainfall (Fig. 6.13a) shows a similar pattern to the rainfall pattern of the third MCA of Fig. 6.8b. This rainfall pattern is related to the SST patterns of the equatorial Atlantic, the southwestern subtropical Atlantic and the SSCP region in the Pacific (Fig. 6.8a). The correlation between the expansion coefficients (time series) of observed rainfall of the second MCA mode (solid line in Fig. 6.13e) and the expansion coefficients (time series) of observed rainfall of the third MCA mode of Fig. 6.8c (solid line) is 0.85 and is statistically significant at the 1% level. The three models are able to reproduce the rainfall pattern over northeast Brazil. They all have negative loadings over this region (Figs. 6.13b, 6.13c and 6.13d) in accordance with the observed pattern (Fig. 6.13a). The negative loadings observed in central South America (in the Amazonian region) and in south Brazil, Uruguay, Paraguay and northern Argentina (Fig. 6.13a) are not reproduced by any model. Only ECMWF (Fig. 6.13c) has some negative loadings in northern Argentina. The positive loadings observed in the southeast region of Brazil (Fig. 6.13a) are well reproduced by all models. The three models show positive loadings over this region (Figs. 6.13b, 6.13c and 6.13d). The positive loadings observed in northwest South America (Fig. 6.13a) are well reproduced by UKMO (Fig. 6.13d), while both CNRM and ECMWF do not capture this feature appropriately (Figs. 6.13b and 6.13c).

Figure 6.14 shows the third MCA mode that explains 3.8% of the squared covariance between observed and predicted rainfall. The observed rainfall pattern (Fig. 6.14a) is not similar to any pattern described in section 6.3.2. The time series of Fig 6.14e suggest that this mode has a very low frequency (decadal) variability. The three models (Figs 6.14b, 6.14c and 6.14d) have difficulty in reproducing the observed pattern (Fig 6.14a). CNRM (Fig 6.14b) nearly completely fails to reproduce the observed pattern (Fig 6.14a). ECMWF is able to capture the observed pattern of positive loadings in the south of Brazil and central-west South America

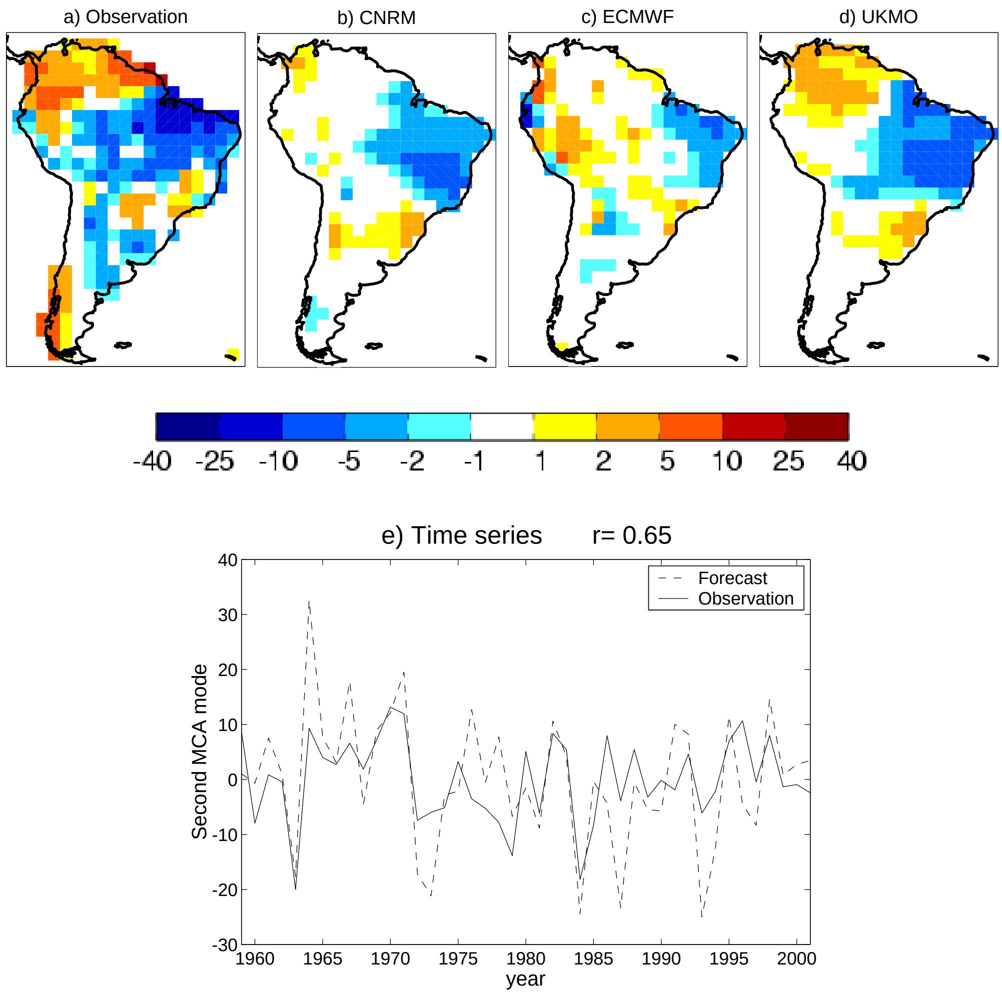


Figure 6.13: Second MCA mode between observed and predicted DJF South America rainfall anomalies for the period 1959-2001. The SCF is 17.7%. Spatial patterns (loadings): a) Observation. b) CNRM. c) ECMWF. d) UKMO. e) Expansion coefficients (time series) of observed rainfall (solid line) and predicted rainfall of these three coupled models (dashed line). The correlation r between these two time series is indicated in panel e.

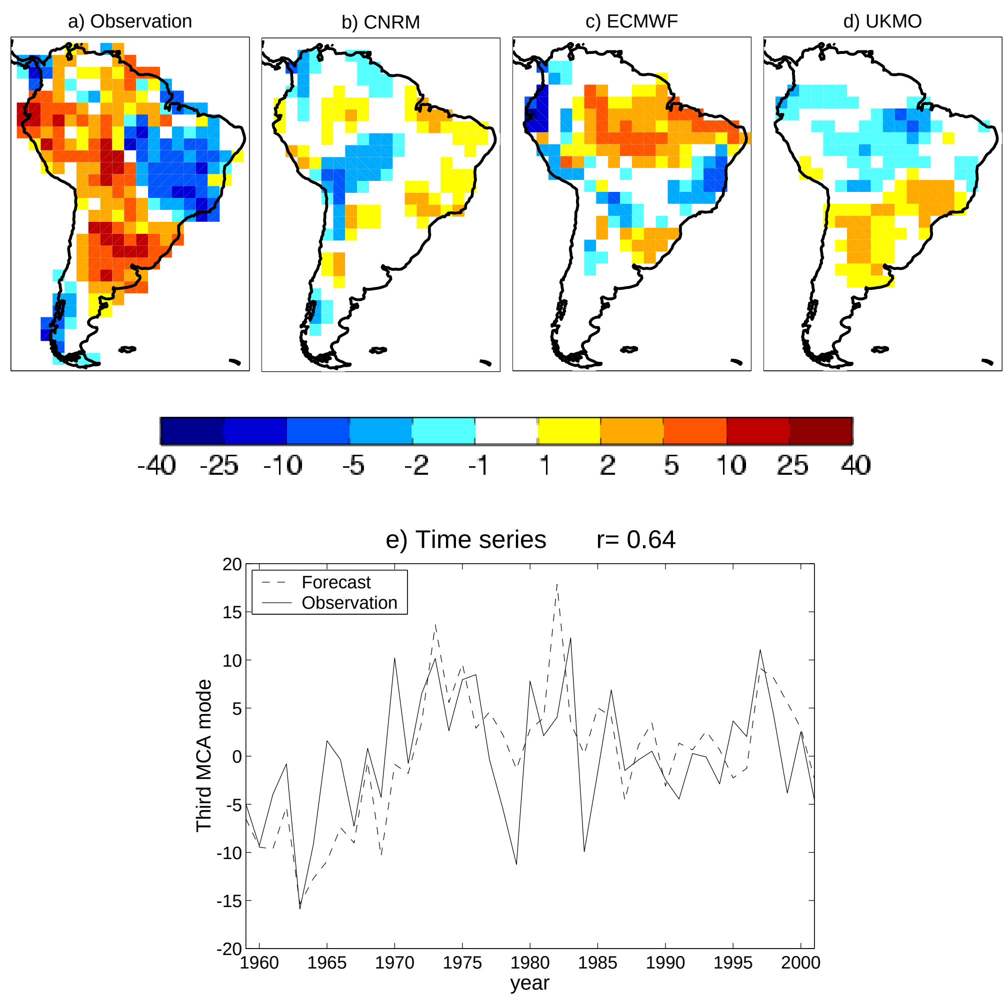


Figure 6.14: Third MCA mode between observed and predicted DJF South America rainfall anomalies for the period 1959-2001. The SCF is 3.8%. Spatial patterns (loadings): a) Observation. b) CNRM. c) ECMWF. d) UKMO. e) Expansion coefficients (time series) of observed rainfall (solid line) and predicted rainfall of these three coupled models (dashed line). The correlation r between these two time series is indicated in panel e.

and also negative loadings in the east coast of Brazil (Fig 6.14c). UKMO reproduces the observed pattern of positive loadings in northern Argentina.

6.6 Rainfall forecasts

6.6.1 Skill assessment

The Bayesian forecast assimilation procedure of chapter 3 (section 3.4) has also been used for the combination of forecasts produced by the three DEMETER coupled models with empirical forecasts produced by the model described in section 6.4. Two additional combined and calibrated forecasts have been produced. One uses empirical forecasts to estimate the prior distribution in the forecast assimilation procedure and is referred here to as Forecast Assimilation with Empirical Prior (FAEP). The other uses the empirical forecasts as an additional model in the forecast assimilation procedure. This means that the empirical forecasts are used as the fourth model in the forecast assimilation. In other words, the matrix X used in FA contains not only forecasts of the three DEMETER coupled models, but also contains the empirical model forecasts. These combined and calibrated forecasts are referred here to as Forecast Assimilation of Coupled model and Empirical forecasts (FACE).

This section assesses the skill of DJF rainfall forecasts produced by empirical, multi-model and the three different Bayesian combined and calibrated predictions (FA, FAEP and FACE). Deterministic and probabilistic skill measures are examined and discussed. The skill of combined and calibrated forecasts is compared to the skill of both the simple multi-model forecast, produced by pooling/averaging together the forecasts of the three DEMETER climate models, and the empirical model that uses ASO SSTs as predictors. The comparison of performance of different climate prediction methods requires temporal consistency in both predictor and predictand and a long common reference period. These two requirements are satisfied in this assessment exercise. Both empirical and coupled models produced predictions for the common period (1959-2001). Consistently, the empirical model uses ASO SST to predict rainfall for the following DJF and coupled models use initial conditions of the 1st of November to predict rainfall for DJF.

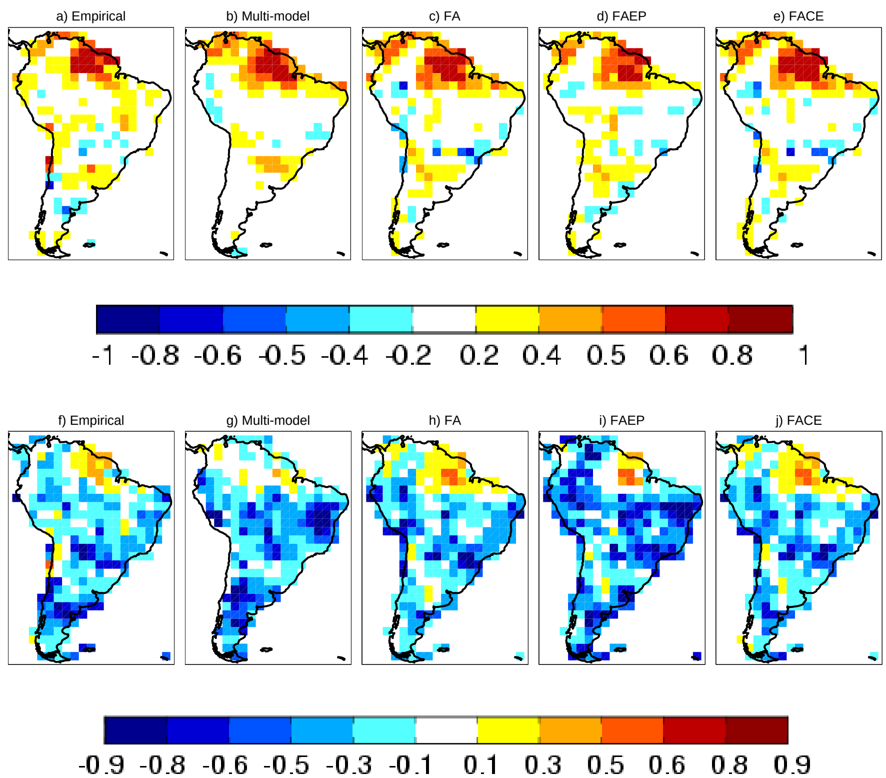


Figure 6.15: Correlation and Brier skill score maps of DJF rainfall anomaly predictions for the period 1959-2001. The Brier skill score is for the event ’rainfall anomalies less than or equal to zero’.

Figure 6.15 shows correlation maps (Figs. 6.15a-e) and Brier Skill Score maps (Figs. 6.15f-j) of rainfall anomaly predictions for empirical, multi-model, FA, FAEP and FACE forecasts for the period 1959-2001. Correlation maps show the correlation between observed and predicted anomalies at each grid point. The BSS is for the event ’rainfall anomaly less than or equal to zero’. The BSS represents the level of improvement of the Brier score (Brier 1950) compared to that of a reference forecast (in this case, the climatological probability of the event). The BSS is designed to range from one for perfect predictions, through zero for predictions that provide no improvement over the reference forecast, to negative values for predictions that are worse than the reference forecast. The tropical region, in northern South America, is the most skilful region with correlations between 0.6 and 0.8 and BSS between 0.1 and 0.6. The subtropics (south Brazil, Uruguay, Paraguay and northern Argentina) also show some skill. Correlations between 0.2 and 0.5 are found in this region. These two regions are well known to be influenced by ENSO. This suggests that most of the skill of South American rainfall predictions is ENSO derived. This is in accordance with Figs. 6.4 and 6.12, which show that most of the variability of South American DJF rainfall is related to ENSO. Empirical, multi-model, FA, FAEP and FACE predictions have similar correlation maps (Figs. 6.15a-e), indicating that all these approaches have comparable level of deterministic skill. The probabilistic measure of skill (Figs. 6.15f and 6.15g) shows that empirical predictions are more skilful than multi-model predictions, particularly in the tropical region where empirical predictions have higher BSS. Bayesian combined and calibrated predictions obtained with forecast assimilation (Figs. 6.15h and 6.15j) have higher BSS than uncalibrated multi-model predictions (Fig. 6.15g). This indicates that the calibration provided by forecast assimilation improves the skill of the multi-model predictions. Combined and calibrated predictions obtained with forecast assimilation have now comparable level of probabilistic skill as empirical predictions. This increase in BSS is mainly due to improvements in the reliability of the predictions (Fig. 6.16a-e), with the tropical regions also showing improvements in resolution (Fig. 6.16f-j). The predominance of negative BSS in Figs. 6.15f-j is due to some properties of this score. Mason (2004) has shown that the expected value of the BSS is less than zero if nonclimatological forecast probabilities are issued. As a result, negative skill scores can often hide useful information content in the forecasts. Therefore, negative skill scores need to be interpreted with caution.

Figure 6.17 shows the mean anomaly correlation coefficient (ACC) for La Niña, neutral and El Niño years occurred during 1959-2001 (Table 6.1) and all (1959-2001) years. The ACC of each year is given by the correlation between the observed and predicted spatial anomaly pattern (Jolliffe and Stephenson 2003, their section 6.3.1). La Niña and El Niño years have higher mean ACC than neutral

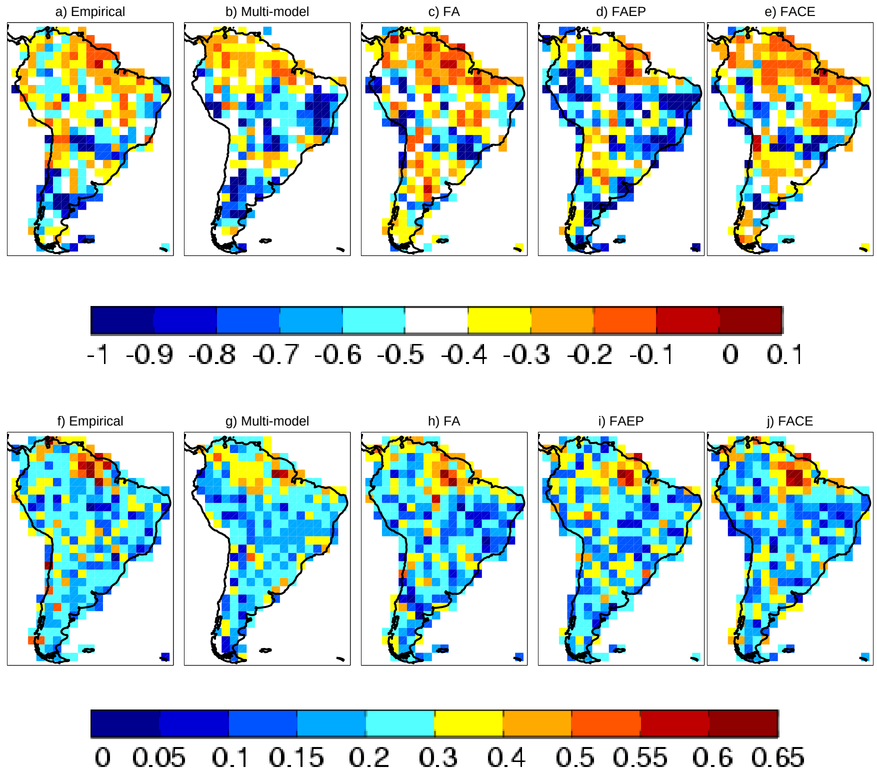


Figure 6.16: Reliability component of the BSS given by $-Rel/BSc$ (panels a-e) and resolution component of the BSS given by Res/BSc (panels f-j) for empirical, multi-model, FA, FAEP and FACE, where Rel and Res are the reliability and resolution components of the Brier Score (see Appendix D), respectively, and BSc is the Brier score of the climatological forecast. Large values of both $-Rel/BSc$ and Res/BSc contribute for increase in the BSS and therefore indicate improved forecast skill.

years, indicating that predictions for ENSO years are more skillful than predictions for neutral years. El Niño and La Niña predictions obtained with forecast assimilation (FA, FAEP and FACE) show an increase in the mean ACC compared to the uncalibrated multi-model. Neutral years have nearly null mean ACC, indicating that rainfall anomalies of these years are hardly predicted. The higher predictability of ENSO years compared to neutral years supports the idea that most of the skill of DJF South American rainfall forecasts is ENSO derived. The mean ACC for all years show that empirical and combined and calibrated forecasts obtained with forecast assimilation have similar levels of skill.

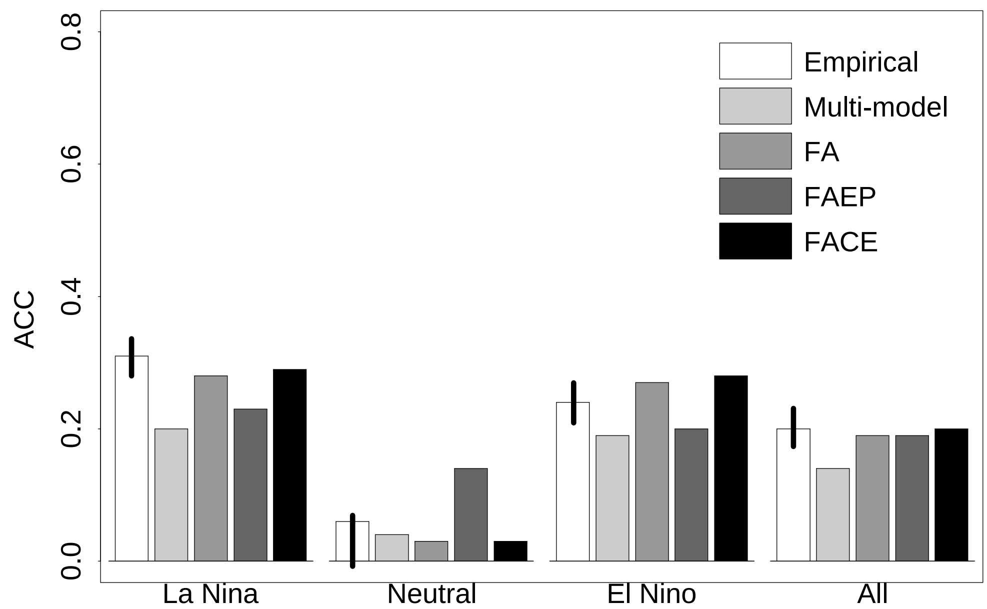


Figure 6.17: DJF mean anomaly correlation coefficient (ACC) for empirical, multi-model, FA, FAEP and FACE forecasts of La Niña, neutral and El Niño years (listed in section 6.3.2 of this chapter) used to produce the composites of Fig. 6.5. The vertical solid lines on the top of the white bars indicate the 95% confidence interval for the mean ACC of empirical forecasts, which were obtained using a bootstrap resampling procedure

The vertical solid lines on the top of the white bars of Fig. 6.17 indicate the 95% confidence interval for the mean ACC of empirical predictions. These intervals were obtained using a bootstrap resampling procedure as described in section 5.3.2 of Wilks (1995). Empirical predictions have the highest mean ACC among all the predictions for La Niña years. FA and FACE predictions have the second and third highest mean ACC for La Niña years, respectively, but they are within the range of values of the 95% confidence interval of the mean ACC of empirical predictions. This indicates that the mean ACC of these three forecasts cannot be considered different from each other from the statistical point of view at the 5% significance level. This suggests that empirical, FA and FACE have similar level of skill when forecasting rainfall of La Niña years. Multi-model and FAEP have smaller mean ACC than empirical, FA and FACE predictions.

The 95% confidence interval of the mean ACC of empirical predictions for neutral years in Fig 6.17 indicates that empirical, multi-model, FA and FACE predictions have similar levels of skill. FAEP has the largest mean ACC indicating that the calibration and combination obtained with FAEP can improve forecast skill when forecasting neutral years. For El Niño years the mean ACC of empirical predictions cannot be considered different from the mean ACC of FA and FACE forecasts at the 5% significance level. This indicates that empirical, FA and FACE predictions have similar levels of skill. As for La Niña years, multi-model and FAEP have smaller mean ACC than empirical, FA and FACE predictions.

6.6.2 ENSO Composites

This section examines the ability of empirical, multi-model, FA, FAEP and FACE predictions in reproducing observed El Niño and La Niña composite patterns during 1959-2001. Figure 6.18 shows observed and predicted DJF South American rainfall anomaly composites for those La Niña and El Niño years listed in Table 6.1. Figures 6.18b and 6.18h show that the empirical model that uses the previous season ASO Pacific and Atlantic SST anomalies as predictor for DJF South American rainfall reproduces remarkably well both El Niño and La Niña observed composite patterns (Figs. 6.18a and 6.18g). The correlation between the empirical forecast and the observed pattern is 0.95 for the La Niña composite and 0.97 for the El Niño composite. Figures 6.18c and 6.18i show that the multi-model composed by three DEMETER coupled models (ECMWF, CNRM and UKMO) partially reproduces the observed pattern in equatorial South America and fails to reproduce the observed pattern in the other regions of the continent. The correlation between the multi-model forecast composite and the observed composite is 0.28 for La Niña and 0.51 for El Niño.

Figures 6.18d and 6.18j show that the use of the Bayesian forecast assimilation (FA) procedure for the calibration and combination of the predictions produced by the three coupled models resulted in forecast composites in much better agreement

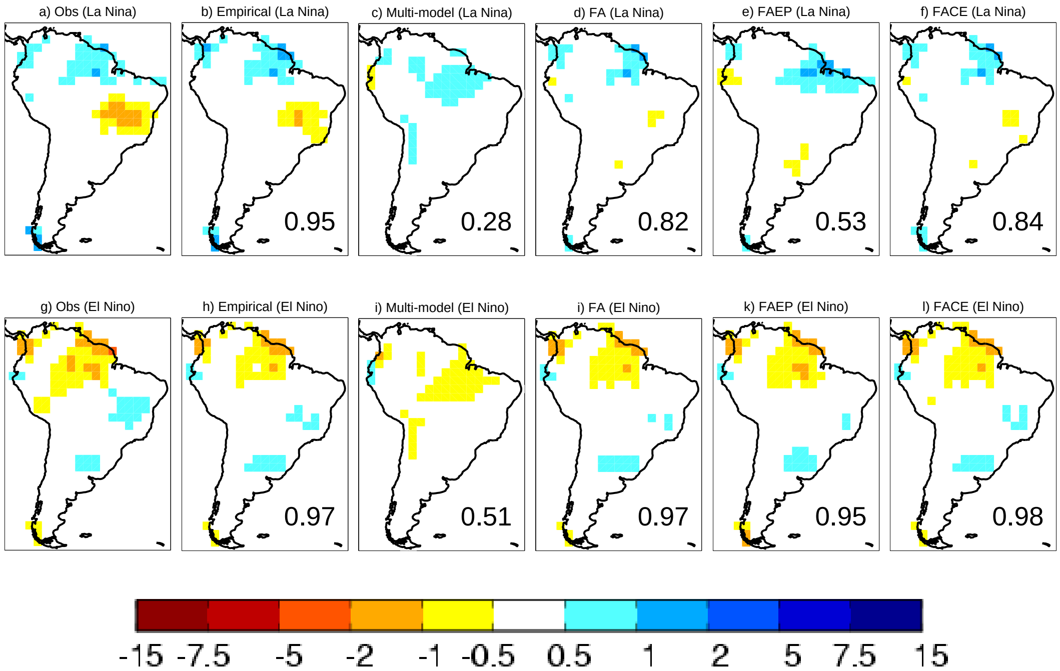


Figure 6.18: DJF South American rainfall anomaly composites ($mm.day^{-1}$) for those La Niña and El Niño years listed in Table 6.1. a) La Niña composite of observed rainfall anomalies . Panels b-f) Empirical, multi-model, FA, FAEP and FACE La Niña forecast composites. g) El Niño composite of observed rainfall anomalies. Panels h-l) Empirical, multi-model, FA, FAEP and FACE El Niño forecast composites. The number in the bottom right hand corner of panels b-f) and h-l) is the correlation between the observed (panels a and g) and the forecast composite.

with the observations (Figs. 6.18a and 6.18g). The correlation between FA composites and observed composites is 0.82 for La Niña and 0.97 for El Niño, being now comparable to the correlation values of empirical forecast composites. These results show that more realistic patterns can be obtained with the calibration and combination of coupled model forecasts using the multimodel Bayesian forecast assimilation approach. The La Niña composite of Fig. 6.18e shows that the use of empirical model predictions to estimate the prior distribution in the forecast assimilation procedure (FAEP) improves over the multi-model (Fig. 6.18c) but does not outperform FA (Fig. 6.18d). The correlation between the FAEP La Niña composite and the observed composite is 0.53 compared to 0.82 for FA. Figure 6.18k shows that for El Niño FAEP provides a similar composite to FA (Fig. 6.18j), and a correlation with the observed pattern of 0.95 compared to 0.97 for FA. Figures 6.18f and 6.18l show that the use of empirical model forecasts as an additional model in the forecast assimilation procedure (FACE) produced similar composites as FA (Figs. 6.18d and 6.18j). The correlation between FACE composites and observed composites is 0.84 for La Niña and 0.98 for El Niño, being these values slightly larger than those of FA composites (Figs. 6.18d and 6.18j)

These results and those presented in the previous section of this chapter (section 6.6.1) indicate that empirical and forecast assimilated combined and calibrated forecasts have similar level of skill for 1-month lead DJF South American rainfall predictions. However, one might argue that at longer leads coupled models may outperform empirical predictions. In order to examine this issue, an empirical model that uses the previous season MJJ (May-Jun-Jul) Pacific and Atlantic SST anomalies as predictors for NDJ South American rainfall has been developed and its forecasts have been compared to DEMETER coupled model predictions produced by ECMWF, CNRM and UKMO for NDJ with initial conditions of the 1st of August (3-month lead). Results shown in Appendix F confirm that empirical and coupled model predictions of South America rainfall have comparable level of skill, even at this extended lead time. It should be noted, however, that the difference between empirical and combined and calibrated predictions is larger for 3-month lead (Figs. F.2) than for 1-month lead (Figs. 6.17) predictions and that La Niña years are less predictable at the extended 3-month lead time.

6.6.3 Rainfall indices

The Bayesian forecast assimilation framework can be used to produce probability forecasts at specific locations/regions (down-scaling). This section demonstrates its utility for forecasting the mean rainfall anomaly at specific regions of South America. The resulting forecast is specified by the mean and the variance of a normal distribution. These two estimates are used to construct interval forecasts. The skill of the forecasts is assessed both deterministically and probabilistically.

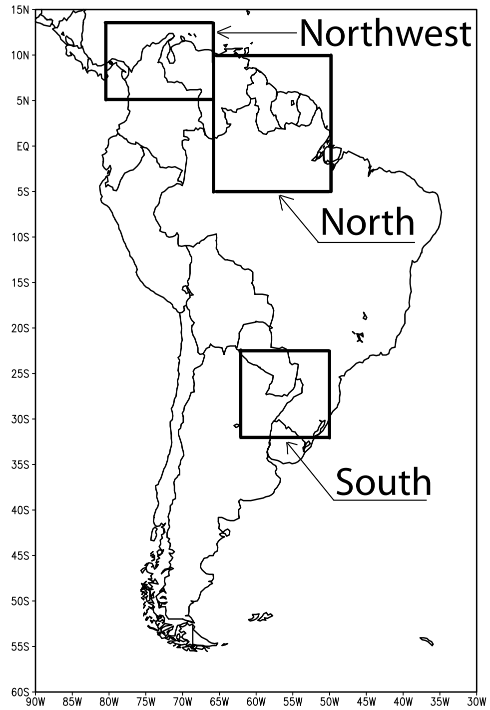


Figure 6.19: Rainfall index boxes

The correlation and Brier skill score maps of Fig. 6.15 reveal three potentially predictable areas over South America. These three regions are in the northwest (5^oN-13^oN; 65^oW-80^oW), north (5^oS-10^oN; 50^oW-65^oW) and south (22.5^oS-32.5^oS; 50^oW-62.5^oW) of South America, as illustrated by the boxes in Fig. 6.19. The observed DJF rainfall of all grid points inside each box of Fig. 6.19 was averaged to produce an index for each region. The index was computed for all the years of the period 1959-2001 in order to produce a $n = 43$ year long time series. This index was used to compose the $n \times q$ matrix Y needed for both empirical predictions and forecast assimilation as described in section 6.4.2 and section 3.4.2, respectively. Note that here Y is a one column matrix (i.e. q=1). All results shown here were obtained in cross-validation mode.

Figure 6.20 shows empirical, multi-model, FA and FAEP 1959-2001 DJF rainfall anomaly predictions for the north box illustrated in Fig. 6.19. Predictions are given by the mean forecast anomaly value (solid line) and the 95$\%$ prediction in-

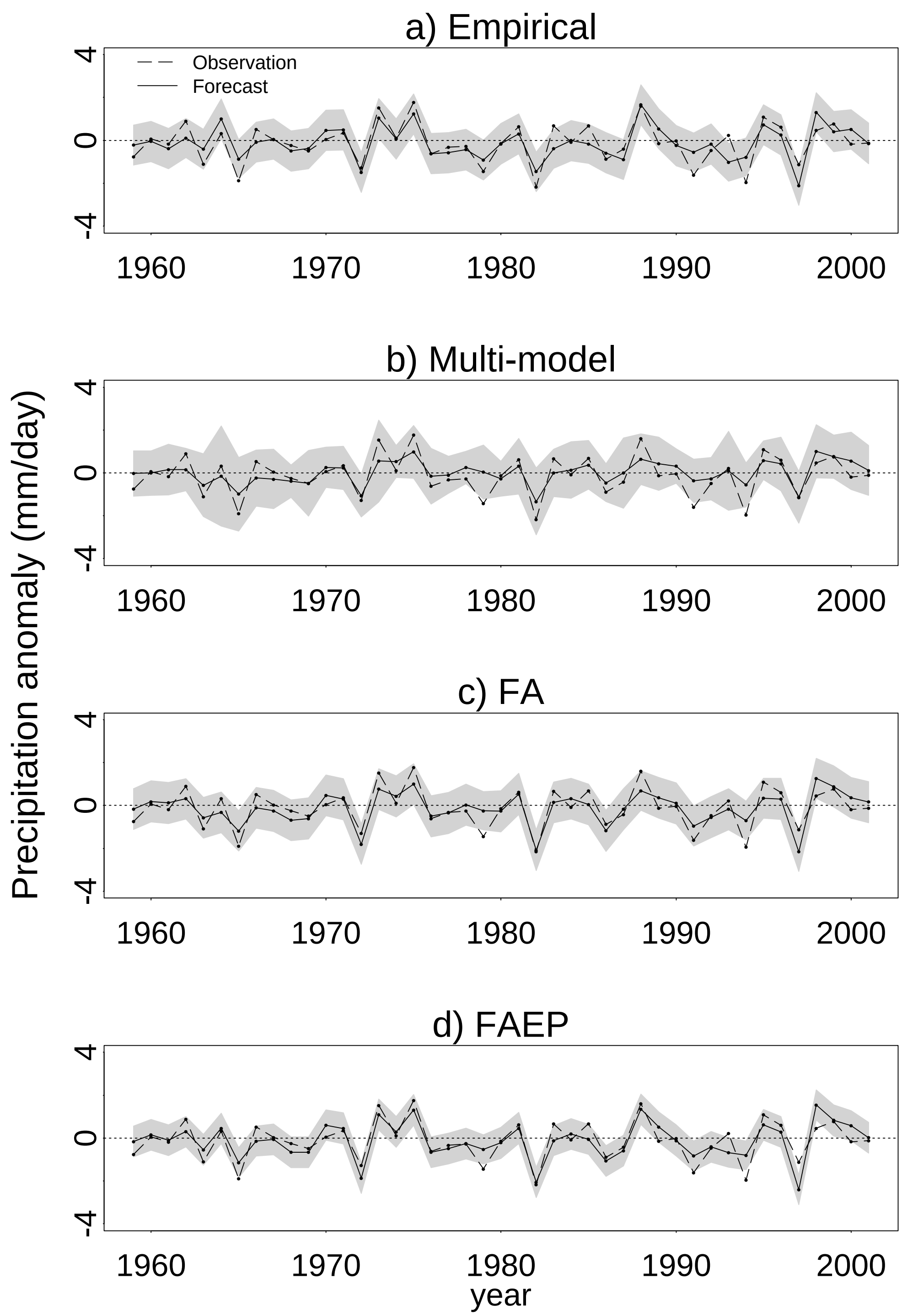


Figure 6.20: Empirical, multi-model, FA and FAEP 1959-2001 DJF rainfall anomaly predictions for the north box illustrated in Fig. 6.16. Mean predicted anomaly (solid line), observed anomaly (dashed line) and the 95% prediction interval (grey shading)

Forecast	MSE $[mm^2]$	Correlation	Uncertainty $[mm]$	BS
Climatology	0.84	-	0.93	0.25
Empirical	0.34	0.77	0.47	0.14
Multi-model	0.36	0.79	0.61	0.17
FA	0.30	0.80	0.48	0.15
FAEP	0.29	0.81	0.36	0.13
FACE	0.29	0.81	0.48	0.14

Table 6.2: Skill and uncertainty measures of DJF rainfall anomaly predictions for the north box. MSE in mm^2, correlation, mean predicted uncertainty in mm, Brier score for the event 'rainfall anomaly less than or equal to zero'. Values refer to the forecast period 1959-2001.

terval (grey shading). This interval is defined by the mean forecast anomaly value plus or minus 1.96 times the prediction standard deviation. Empirical predictions (Fig. 6.20a) use Pacific and Atlantic ASO SSTs as predictors for DJF rainfall, as described in section 6.4.2. Multi-model predictions (Fig. 6.20b) are obtained from the 27 member ensemble produced by ECMWF, CNRM and UKMO. FA predictions (Fig. 6.20c) are produced with the Bayesian forecast assimilation procedure of section 3.4.2 (chapter 3). FAEP predictions (Fig. 6.20d) use empirical predictions (Fig. 6.20a) as estimates of the prior distribution in the forecast assimilation procedure. The four predictions of Fig. 6.20 reproduce remarkably well the observed anomalies (dashed line) during the period 1959-2001.

Table 6.2 shows skill and uncertainty measures for the four prediction methods of Fig. 6.20 that are used to forecast DJF rainfall anomalies in the north box, in addition to climatological and FACE predictions. Climatological predictions have the highest MSE and Brier score of all methods, indicating that they provide the least skillful prediction for the north box. Climatological predictions also provide the largest mean forecast uncertainty of all prediction methods. Empirical and multi-model predictions have comparable MSE and correlation but empirical predictions have less uncertainty than multi-model predictions. Empirical predictions have a smaller Brier score than multi-model predictions. This indicates that

Forecast	MSE $[mm^2]$	Correlation	Uncertainty $[mm]$	BS
Climatology	0.32	-	0.58	0.24
Empirical	0.25	0.52	0.40	0.20
Multi-model	0.21	0.61	0.47	0.20
FA	0.19	0.64	0.40	0.19
FAEP	0.25	0.58	0.32	0.21
FACE	0.19	0.64	0.40	0.19

Table 6.3: Skill and uncertainty measures of DJF rainfall anomaly predictions for the northwest box. MSE in mm^2, correlation, mean predicted uncertainty in mm, Brier score for the event 'rainfall anomaly less than or equal to zero'. Values refer to the forecast period 1959-2001.

empirical predictions have better estimates of forecast uncertainty than multi-model predictions. FA, FAEP and FACE predictions have comparable MSE, correlation and Brier score. FAEP has a smaller mean forecast uncertainty estimate than FA and FACE, as noted in Fig. 6.20 by the smaller 95$\%$ prediction interval of FAEP compared to FA. This results in a slightly smaller Brier score for FAEP when compared to the Brier score of FA and FACE. FA, FACE and FAEP have slightly smaller MSE and slightly higher correlations than empirical and multi-model predictions. This indicates that Bayesian forecast assimilated predictions provide slightly better estimates of the mean rainfall anomaly than empirical and multi-model predictions. In probabilistic terms, the Brier score reveals that Bayesian forecast assimilated predictions have comparable level of skill to empirical predictions and slightly better skill than multi-model predictions.

Figure 6.21 shows 1959-2001 DJF rainfall anomaly predictions for the northwest box illustrated in Fig. 6.19. The four prediction methods are able to reproduce most of the interannual variability of the observed anomalies. Table 6.3 shows that climatological predictions have the highest MSE and Brier score of all methods, indicating that they provide the least skillful prediction for the northwest box. Climatological predictions also provide the largest mean forecast uncertain of all prediction methods. Multi-model predictions have smaller MSE and higher corre-

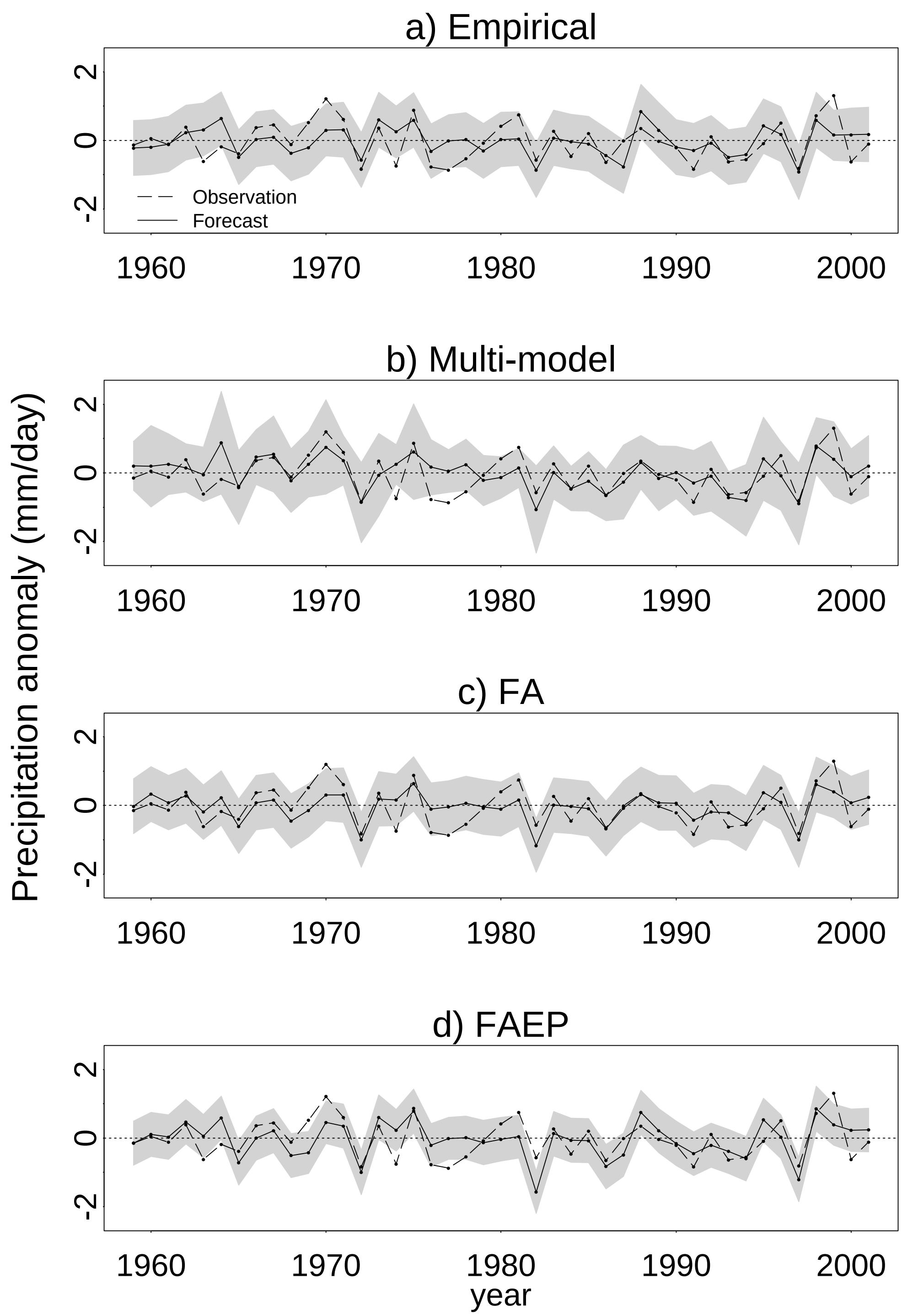


Figure 6.21: Empirical, multi-model, FA and FAEP 1959-2001 DJF rainfall anomaly predictions for the northwest box illustrated in Fig. 6.16. Mean predicted anomaly (solid line), observed anomaly (dashed line) and the 95% prediction interval (grey shading)

lation than empirical predictions, indicating that multi-model predictions provide better estimates of the mean rainfall anomaly than empirical predictions. Empirical predictions have less uncertainty than multi-model predictions and coincidently these two prediction methods have the same Brier score. FA and FACE have the same MSE, correlation, mean prediction uncertainty and Brier score. The MSE of FA is slightly smaller than the MSE of multi-model predictions and the correlation of FA is slightly larger than the correlation of multi-model predictions. This indicates that the calibration produced by forecast assimilation improves the estimate of the mean rainfall anomaly. FA predictions have less uncertainty than the multi-model and a slightly smaller Brier score than the multi-model, indicating that forecast uncertainty is slightly better estimated by FA. As for the north box, FAEP has the smallest predicted uncertainty as illustrated by the smaller 95$\%$ prediction interval in Fig. 6.21d compared to Figs. 6.21a, 6.21b and 6.21c. Note, however, that a large number of observations lie outside the 95$\%$ prediction interval, what makes FAEP the least reliable forecast among all here investigated.

Forecast	MSE $[mm^2]$	Correlation	Uncertainty $[mm]$	BS
Climatology	0.56	-	0.75	0.25
Empirical	0.56	0.28	0.60	0.27
Multimodel	0.47	0.43	0.46	0.24
FA	0.47	0.42	0.62	0.21
FAEP	0.58	0.38	0.52	0.27
FACE	0.47	0.42	0.62	0.21

Table 6.4: Skill and uncertainty measures of DJF rainfall anomaly predictions for the south box. MSE in mm^2, correlation, mean predicted uncertainty in mm, Brier score for the event 'rainfall anomaly less than or equal to zero'. Values refer to the forecast period 1959-2001.

Figure 6.22 shows 1959-2001 DJF rainfall anomaly predictions for the South box of Fig. 6.19. The four prediction methods have difficulties in reproducing the observed anomalies. Table 6.4 shows that FAEP predictions have the highest MSE and Brier score of all methods, indicating that they provide the least skillful predic-

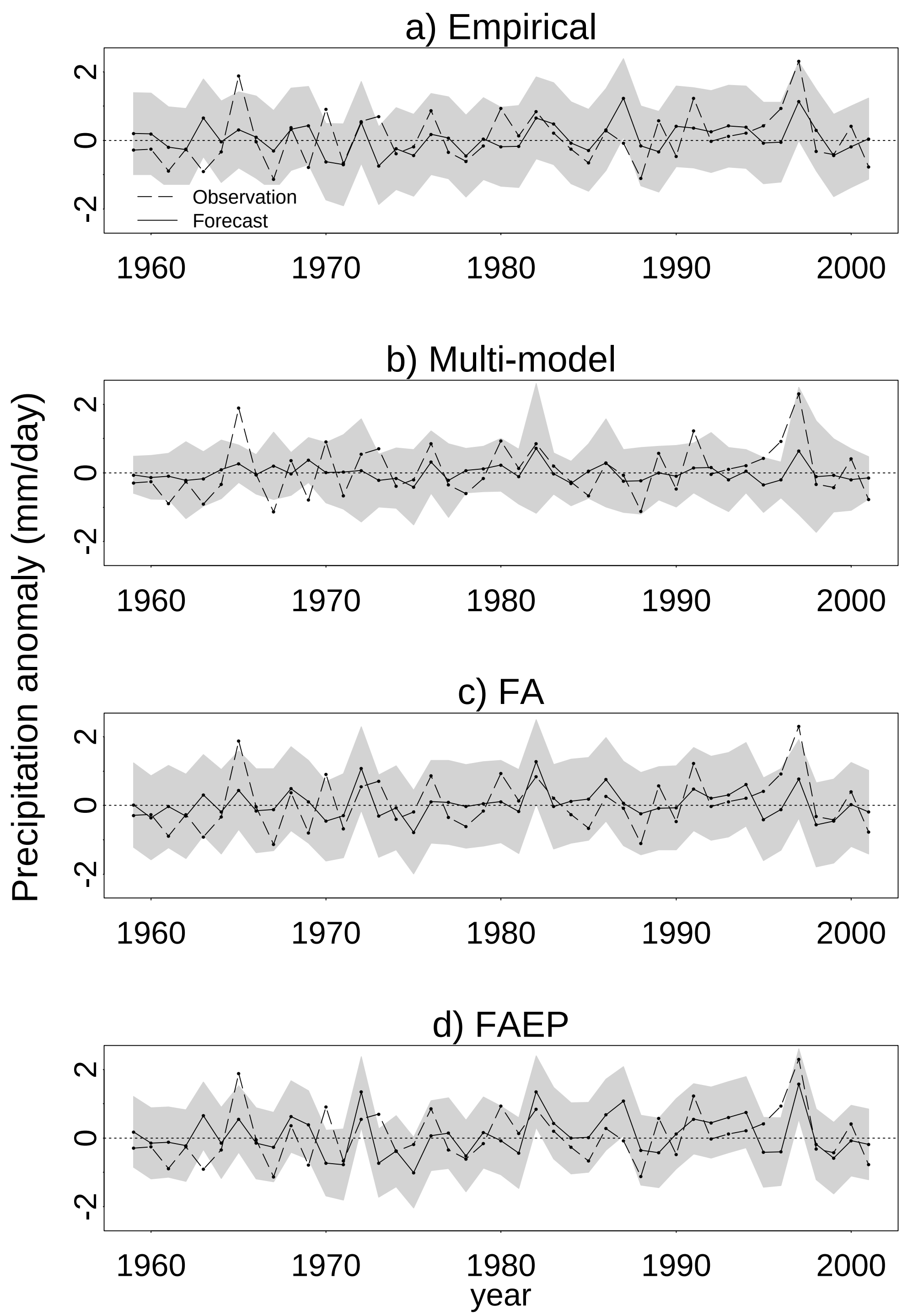


Figure 6.22: Empirical, multi-model, FA and FAEP 1959-2001 DJF rainfall anomaly predictions for the south box illustrated in Fig. 6.16. Mean predicted anomaly (solid line), observed anomaly (dashed line) and the 95% prediction interval (grey shading)

tion for the south box. Climatological predictions provide the largest mean forecast uncertainty of all prediction methods. Multi-model predictions have smaller MSE and higher correlation than empirical predictions, indicating that multi-model predictions provide better estimates of the mean rainfall anomaly than empirical predictions. Empirical predictions have more uncertainty than multi-model predictions and a higher Brier score than multi-model predictions, indicating that multi-model predictions are more skillful than empirical predictions. FA and FACE have the same MSE, correlation, mean prediction uncertainty and Brier score. FA and multi-model predictions have the same MSE and FA has a slightly smaller correlation than multi-model predictions. Note, however, that FA has a larger mean prediction uncertainty than the multi-model, resulting in a smaller Brier score than the multi-model. This indicates that the calibration produced by forecast assimilation improved probabilistic forecast skill.

The comparison of the magnitude of the MSE, Brier score and correlation of Tables 6.2, 6.3 and 6.4 reveals that the north and northwest boxes are more predictable than the south box. Note, however, that the south box shows the highest difference between empirical and FA (or multi-model), indicating that coupled models provide useful forecast information for this region.

6.7 River flow forecasts

This section shows an example of application of forecast assimilation for river flow forecasting at Tucuruí (3.75^{o}S, 49.68^{o}W) in the north of Brazil. Tucuruí is the second largest hydropower station in Brazil, capable of producing peak power of 4240 MW. Forecast assimilation has been performed using the observed 1959-2001 DJF Tocantins river flow at Tucuruí to compose the one column matrix Y as described in section 3.4.2. The matrix X was composed by 1-month lead DJF rainfall predictions produced by ECMWF, CNRM and UKMO coupled models.

Figure 6.23 shows cross-validated DJF river flow predictions for Tucuruí obtained with forecast assimilation (solid line). Forecast assimilation reproduces well

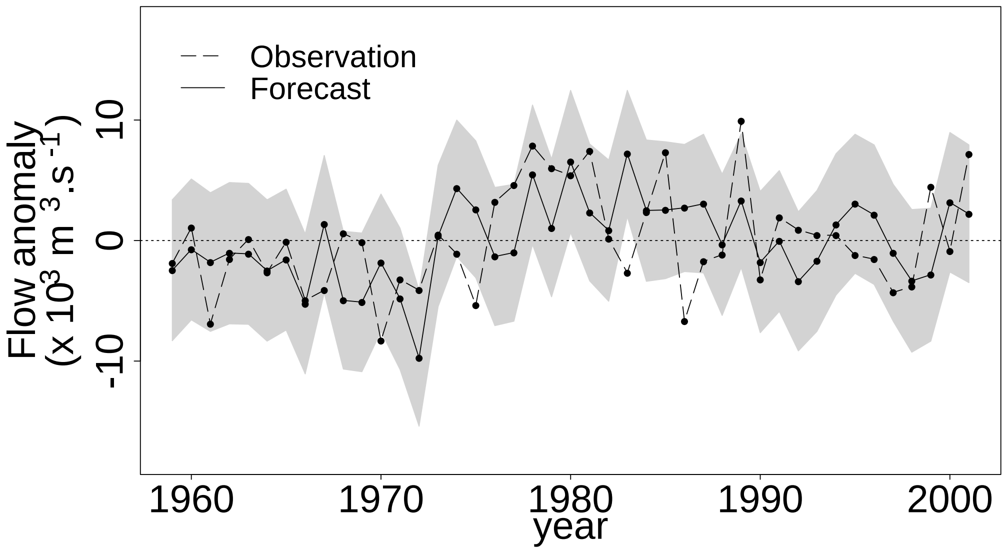


Figure 6.23: FA 1959-2001 DJF flow anomaly predictions for Tucuruí. Mean predicted anomaly (solid line), observed anomaly (dashed line) and the 95% prediction interval (grey shading)

the inter-decadal variability of the observed anomalies (dashed line). The correlation between observed and predicted anomalies is 0.35. Most observations are within the 95% prediction interval (grey shading), indicating that forecast assimilation provides reliable interval forecasts of river flow anomalies for Tucuruí.

6.8 Summary

This study addressed seasonal predictability of South American rainfall. The skill of empirical, DEMETER coupled multi-model and combined and calibrated predictions obtained with forecast assimilation has been assessed and compared. This comparison revealed that when seasonally forecasting Dec-Jan-Feb South American rainfall at 1-month lead-time the current generation of coupled models have comparable level of skill to those obtained using a simplified empirical approach. The same conclusion still holds for longer (e.g. 3-month) lead times. This result is in agreement with findings of previous comparison studies (e.g. Folland *et al.* 2001;

Moura and Hastenrath 2004). This implies that both empirical and coupled model predictions are sufficient for each other. In other words, in the example presented here, empirical predictions do not provide additional skill to coupled model predictions and vice-versa. The tropics and the area of south Brazil, Uruguay, Paraguay and northern Argentina have been found to be the two most predictable regions of South America. South American rainfall is generally only predictable in ENSO years rather than in neutral years, which exhibit very little skill.

Bayesian forecast assimilation has been shown to be a powerful tool for the calibration of multi-model predictions. The resulting forecasts have been shown to have improved Brier scores compared to the simple multi-model prediction. This is because forecast assimilation provides reliable estimates of forecast uncertainty. Forecast assimilation predictions also helped to produce probability forecasts with skill in southeastern South America – an important region for South American hydroelectricity production. Additionally, forecast assimilation rainfall composites for El Niño and La Niña years have been shown to be in much better agreement with observed composites than multi-model composites. Finally, forecast assimilation has been shown to be useful for local down-scaling of rainfall and river flow anomalies.

Chapter 7

Conclusion

7.1 Summary of thesis results

This thesis has reviewed the literature of forecast calibration and combination in atmospheric sciences and economics. Additionally, it has also introduced a new framework for the calibration and combination of forecasts. This framework has been proposed for the production of calibrated probability forecasts of observable variables from multi-model ensemble predictions. In analogy with data assimilation, the concept of forecast assimilation has been introduced. Statistical modelling methods have previously been developed for the production of empirical predictions, which are entirely based on observations, for down-scaling the outputs of physically-derived dynamical models to specific locations/regions, and also for the calibration of deterministic predictions produced by physically-derived dynamical models (MOS). This thesis proposes forecast assimilation as a statistical modelling method for the treatment of multi-model ensemble predictions. Forecast assimilation is an inherently Bayesian procedure for making improved forecasts of observable variables based on information provided by ensemble predictions produced by distinct climate models. It incorporates many previous techniques such as MOS and statistical downscaling as special cases.

The methodology used here has been developed and tested progressively in three stages. First, a univariate Bayesian approach for calibrating and combining empirical and raw (uncorrected) coupled model ensemble forecasts of the Niño-3.4 index has been developed. Combined 5-month lead Niño-3.4 forecasts for December have been shown to have greater skill than either coupled model or empirical forecasts separately. Combined forecasts provided more reliable estimates of forecast uncertainty than raw and bias-corrected coupled model forecast. Combined forecasts have shown a deterministic skill score 5% larger than bias-corrected forecasts and 13% larger than empirical forecasts. Combined forecasts have also shown a probabilistic skill score 13% larger than bias-corrected forecasts and 3% larger than empirical forecasts. The Bayesian approach for calibrating and combining Niño-3.4 forecasts improves both accuracy and reliability of the predictions.

Forecast assimilation has then been used for the calibration and combination of DEMETER coupled model predictions of equatorial Pacific SSTs. Therefore, in this second stage the method acquired its first spatial (longitudinal) component. Combined and calibrated forecasts obtained with forecast assimilation have been shown to have improved reliability in the western Pacific and improved resolution in the eastern Pacific. The improvement in reliability in the western Pacific has been obtained without degrading the resolution of the predictions. Reliability is a highly desirable characteristic of a prediction system. Forecast assimilation improved by around 75% the reliability of the predictions in the western Pacific and around 25% the resolution of the predictions in the eastern Pacific. Such an improvement in resolution could not be achieved by calibration of predictions of a single model.

In the third stage, the Bayesian multi-model forecast assimilation procedure has been applied for the calibration and combination of spatial field forecasts of rainfall over South America in DJF. An empirical model based on MCA of rainfall with SSTs has also been developed to predict DJF South American rainfall. Empirical predictions have been combined with 1-month lead DEMETER coupled model predictions using the forecast assimilation procedure. Bayesian forecast assimilation has been shown to improve forecast skill over some regions of South America. Forecast assimilation improved both reliability and resolution of the predictions in tropical South America. Forecast assimilation improved the reliability of the predictions in tropical South America by 65% and the resolution of the predictions by 20%. Eastern Brazil and southeastern South America also had the reliability of the predictions improved by forecast assimilation. Eastern Brazil had an improvement of about 50% in reliability while southeasten South America had an improvement of about 70% in reliability. Combined and calibrated forecasts produced with forecast assimilation have been compared with both empirical and DEMETER multi-model forecasts. This comparison revealed that empirical and coupled model forecast have similar level of skill. For both forecasting systems ENSO years have been found to be more predictable than neutral years, which have hardly any skill. The tropics and the area of south Brazil, Uruguay, Paraguay and northern Argentina are the two most predictable regions of South America. Finally, forecast assimilation has successfully been used for statistical down-scaling of DJF precipitation indices for three regions of South America and also for local river flow predictions for Tucuruí (north of Brazil). Forecast assimilation is a powerful new tool for the calibration and combination of predictions, capable of improving the skill of probabilistic forecasts.

7.2 Future areas of research

The statistical method proposed in this thesis (forecast assimilation) has some advantages:

- produces well-calibrated probability forecasts
- able to deal with ensemble predictions
- able to deal with multi-model predictions
- preserves spatial structure present in the datasets
- allows spatial patterns to be shifted/corrected

Forecast assimilation also has some potential disadvantages:

- normality assumption
- need for data reduction to be able to estimate regression parameters
- relationships can change with time (stability)

• need to re-compute calibration equations (regression) each time the forecasting system changes

In summary, the examples of forecast assimilation presented here assume that the data are close to normally distributed and treat gridded fields as multivariate vectors instead of taking account of the true spatial nature of the datasets. It would be interesting in future studies to extend the methods to non-normally distributed data, to deal with weather and climate extremes, and to develop flexible functional methods that can exploit the smoothness of spatial fields as an additional constraint. There is also scope for improving forecast assimilation by including state-dependent ensemble spread information in the estimation of the prediction error covariance S. Another interesting avenue for future research would be to develop forecast assimilation approaches that use knowledge of the data assimilation operator H to help estimate G. The use of calibrated predictions obtained with forecast assimilation as the first guess analysis for data assimilation is also potentially feasible. In this way forecast assimilation and data assimilation would share information with each other, making the prediction process cyclic. Non-stationarity of climate can also affect forecast calibration. The development of more generalised methods capable of dealing with non-stationary time series might also improve forecast assimilation. Forecast assimilation is an essential yet often poorly acknowledged aspect of the forecasting process and hopefully this thesis will stimulate more coordinated activity in this area for weather and climate predictions on all lead times.

Appendix A

Classical and inverse regression

Rather than regress the forecasts x on the observations y as illustrated in Fig. 3.3, it might at first appear more natural to regress the observations y on the forecasts x (as is done in MOS). In other words, one can use the physically-derived coupled model forecasts as predictors in a regression model to obtain predictions of the observations as illustrated in Fig. A.1. However, the (explanatory) forecast values are not deterministic control variables but instead contain large amounts of uncertainty. Furthermore, it can be assumed that climate forecasts are generally more uncertain than the observed values. This is clearly illustrated in Fig. 3.3, where the forecasts spread along the vertical axis for a fixed (more certain) observed value. Besides, the least-squares estimation minimises the (vertical) error of the response variable for a fixed value of the explanatory variable. However, by fitting a regression model of observations on forecasts as illustrated in Fig. A.1 one assumes that forecasts are more certain than are observations – which from Figs. 3.3 and A.1 is clearly not the case – and minimises (vertical) errors of observations for fixed forecast values. For these reasons and what follows, it is better to develop a regression model of the forecasts as a function of the observed values rather than regress observations on forecasts as in Fig. A.1. Least-squares estimation then corresponds to minimising forecast error for fixed values of the observed variable.

The calibration of the forecast $\bar{x}_t$ to the observed predictand y_t can be considered as a classical calibration problem for an instrumental device. This is a

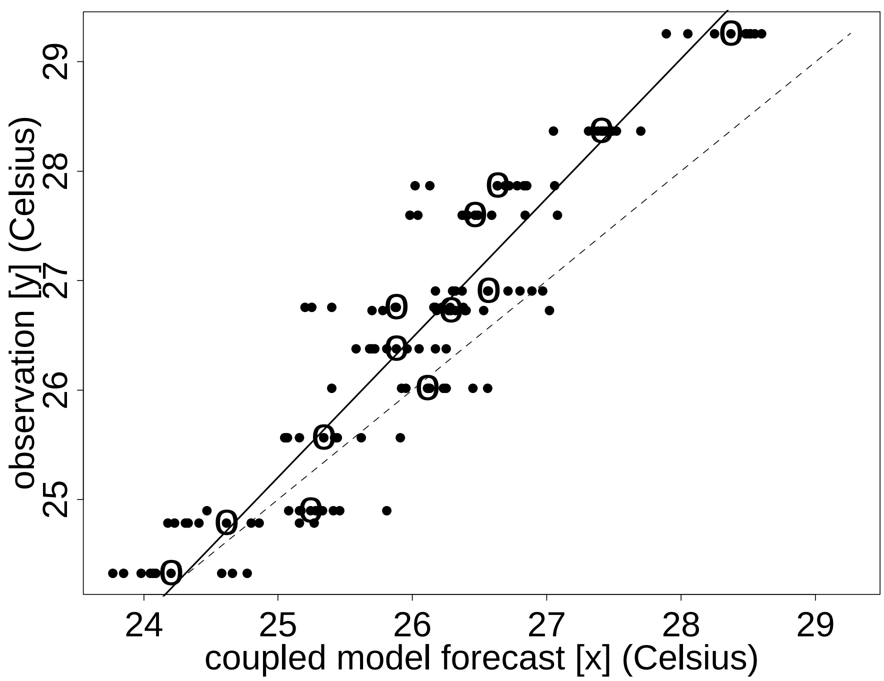


Figure A.1: Regression of observed values y on ECMWF coupled model ensemble mean forecasts $\bar{x}$ of December Niño-3.4 index for the period 1987-1999 (solid line). Parameter estimates of Eqn. A.8 are $\hat{a} = -6.68^oC$, $\hat{b} = 1.28$ ($R^2 = 0.93$). Each black dot is one of the $m = 9$ ensemble members. Big open circles are ensemble means $\bar{x} = x_m$. The dashed line shows what would be obtained for perfect forecasts.

long standing issue in statistical literature, often referred to as the *inverse regression problem* (Brown 1994). It is relevant to probability forecasting and so will be briefly reviewed here.

In the simplest classical calibration setting, a precise instrument gives a measurement y_t, while a less precise instrument, to be calibrated, produces $\bar{x}_t$ for the same quantity. The calibration database consists of a time series of paired values $\{(y_t, \bar{x}_t), t = 1, 2, ..., n\}$. Some classical examples for y_t and $\bar{x}_t$ are respectively pressure and gauge readings (Seber 1977), tree-ring counts and (the less precise) carbon dating measure (Draper and Smith 1998), or a long and costly laboratory method for determining the concentration of a certain enzyme in blood plasma samples and a quick and cheap autoanalyser device (Aitchison and Dunsmore 1975).

In chapter 4, y_t is the (more precise) best estimate of the observed variable (e.g. Niño-3.4 index), while $\bar{x}_t$ is the (less precise) physically-derived raw coupled model ensemble-mean forecast of the same variable for the same year t. The coupled model forecast can be considered to be an instrument for diagnosing the predictand, and calibrating the forecasts then becomes a standard issue of instrumental calibration (Swets 1988). The problem of estimating y_t when a new reading $\bar{x}_t$ becomes available is then an inverse regression problem. This is precisely the problem of calibrating some new forecast $\bar{x}_t$ when an historical database is available.

The established protocol stems at least from Eisenhart (1939) (see also Seber 1977; Aitchison and Dunsmore 1975; Draper and Smith 1998; and Brown 1982). Since the errors in y-values are negligible when compared with the device (forecast) errors, y_t can be treated as the fixed control values and then one obtains the *classical regression model* of $\bar{x}_t$ versus y_t:

$$\bar{x}_t = \alpha + \beta y_t + \epsilon_t \tag{A.1}$$

where ϵ_t are independent normally distributed random variables with zero mean and variance δ. This regression model (illustrated in Fig. 3.3) has been defined in probabilistic notation by Eqn. (3.5).

The least squares solutions of the calibration equation (A.1) are given by

$$\hat{\alpha} = \overline{\bar{x}} - \hat{\beta}\,\overline{y} \tag{A.2}$$

$$\hat{\beta} = \frac{s_{\bar{x}}}{s_y}\, r \tag{A.3}$$

where the hat symbol denotes an estimated parameter, the overbar indicate time mean over the calibration period *n*, r is the mutual correlation between $\bar{x}_t$ and y_t and $s_{\bar{x}}$ and s_y are the sample standard deviations of $\bar{x}_t$ and y_t, respectively.

Then the classical maximum likelihood (ML) estimate of y_t is

$$\hat{y}_t = (\bar{x}_t - \hat{\alpha})/\hat{\beta}. \tag{A.4}$$

To avoid explosive estimates when $\hat{\beta} \approx 0$, truncated forms of Eqn. (A.4) can be defined. In summary, the classical calibration model (A.1) considers the conditional distribution of $\bar{x}$ given y (i.e., $\bar{x}|y$), since the calibrating equation (A.1) describes the stochastic measures conditionally to the true quantities.

Whereas Williams (1969) and others advocated the use of Eqn. (A.1) to derive the ML estimate of y_t (Eqn. A.4), one can also think of defining the *inverse regression model* for $y|\bar{x}$ and then use it directly for estimating y_t. Following this idea, Krutchkoff (1967, 1969), suggested the so-called inverse estimate

$$\hat{y}_t^K = \hat{a} + \hat{b}\bar{x}_t \tag{A.5}$$

based on the least squares estimates

$$\hat{a} \;=\; \overline{y} - \hat{b}\,\overline{\overline{x}} \tag{A.6}$$

$$\hat{b} \;=\; \frac{s_y}{s_{\bar{x}}}\,r \tag{A.7}$$

obtained from the *inverse regression model*

$$y_t = a + b\bar{x}_t + e_t. \tag{A.8}$$

Classical (Eqn. A.4) and inverse (Eqn. A.5) estimates coincide only when $\bar{x}_t$ is perfectly correlated with y_t in the calibration database, i.e., when $r = 1$, $\overline{\overline{x}} = \bar{y}$ and $s_{\bar{x}} = s_y$. The inverse regression approach is currently the prevalent method for correcting forecast biases in atmospheric sciences (see Kharin and Zwiers 2002; Pavan and Doblas-Reyes 2000)

Krutchkoff (1967) used simulations to show that the inverse method can have smaller MSE than the classical calibration approach (even in the truncated form). This led to a controversy in which the MSE criterion was criticised for this particular case. An alternative criteria was proposed and the conditions of relative superiority of one method over the other were investigated in depth by Williams (1969), Berkson (1969), Halperin (1970) and Hoadley (1970) among others, and later on by Chow and Shao (1990).

The Bayesian approach was useful in clarifying the controversy (Hoadley 1970; Aitchison and Dunsmore 1975). Ideally, one would like the conditional distribution of $y|\bar{x}$ but of course this cannot be obtained from the conditional distribution of $\bar{x}|y$ without also having an estimate of the marginal prior distribution $p(y)$. By means of the prior $p(y)$ and the likelihood $p(\bar{x}|y)$ the distribution $p(y|\bar{x})$ can be obtained using Bayes' theorem (Eqn. 3.3) and the inverse regression problem can be solved. In order to understand the relative merits of classical and inverse estimators, note that both are special cases of the Bayesian estimator with exactly the same normal likelihood distribution given by Eqn. (3.5), but with two different prior distributions $p(y)$ (Hoadley 1970). The classical maximum likelihood estimator (Eqn. A.4) is obtained using a uniform prior distribution $p(y) \propto 1$. A normal likelihood distribution (Eqn. 3.5) when combined with a uniform prior distribution produces a posterior distribution $p(y_t|\bar{x}_t)$ that is normal with mean given by $\hat{y}_t$ as in Eqn. (A.4). As demonstrated by Hoadley (1970), the inverse estimator $\hat{y}_t^K$ of Eqn. (A.5) is obtained using the normal likelihood distribution of Eqn. (3.5) and a more informative prior, which is given by a normal distribution with mean and variance estimated from the same n historical values of y used to build the likelihood.

These correspondences are valid for any continuous variables but are easily demonstrated when $\bar{x}$ and y are standardized variables, i.e. variables with zero mean and unit variance ($\overline{\overline{x}} = \bar{y} = 0$ and $s_{\bar{x}} = s_y = 1$). For such variables, Eqns.

(3.8), (3.9), (A.2) and (A.3) give

$$\frac{1}{\sigma_t^2} = \frac{1}{{\sigma_{ot}}^2} + \frac{r^2}{\delta} \tag{A.9}$$

$$\frac{\mu_t}{\sigma_t^2} = \frac{\mu_{ot}}{{\sigma_{ot}}^2} + \frac{r^2}{\delta}\left(\frac{\bar{x}_t}{r}\right) \tag{A.10}$$

For a uniform prior distribution the term $1/\sigma_{ot}^2$ is zero in Eqns. (A.9) and (A.10) leading to the Bayesian estimator

$$\hat{\mu}_t = \bar{x}_t/r \tag{A.11}$$

that is exactly the same as the classical estimator $\hat{y}_t = \bar{x}_t/r$ obtained from Eqns (A.2), (A.3) and (A.4) for standardized variables.

For a normal prior distribution with mean $\mu_{ot} = \bar{y} = 0$ and variance $\sigma_{ot}^2 = s_y = 1$ (since y is a standardized variable) estimated from the same n historical values of y used to build the likelihood, it follows from Eqns. (A.9) and (A.10) that

$$\hat{\mu}_t = \frac{r}{\delta_* + r^2}\,\bar{x}_t \tag{A.12}$$

with $\delta_* = E[(\bar{x} - \alpha - \beta y)^2] = 1 - r^2$, where the symbol E denotes expectation (population mean). Then the Bayesian estimator is

$$\hat{\mu}_t = r\,\bar{x}_t \tag{A.13}$$

that is exactly the same as the inverse estimator $\hat{y}_t^K = r\,\bar{x}_t$ obtained from Eqns (A.5), (A.6) and (A.7) for standardized variables.

In the current comparison between classical and inverse estimators, the inverse will do well if y_t lies centrally in the set of previous y-values used in fitting the inverse calibration (Eqn. A.8). On the other hand, the classical estimator, obtained using a uniform prior distribution, will be more efficient for more extreme y_t-values (Brown 1982). Since the prior of the inverse regression model is centred on the calibration mean $\bar{y}$, the comparison of inverse and classical estimates will be unfair to the latter if the calibration database coincides with the verification database.

Note, however, that rather than using different estimators, the best method is to choose the best prior for any particular application (the Bayesian approach). To do this, one needs extra information about y alone. In forecast calibration this is the most common situation, where a short bivariate time series $\{(y_t, \bar{x}_t), t = 1, 2, ..., n\}$ is used for the calibration and a longer historical climatology can be used to estimate the prior. The utility and flexibility of the Bayesian approach in combining the two sources of information is apparent. The use of more complex prior distributions including other predictors can further help in adapting the prior to the particular forecasting conditions.

Appendix B

Derivation of the univariate posterior distribution

From Eqns. (3.4) and (3.5) the prior and the likelihood p.d.f.'s are respectively:

$$p(y_t) = N(\mu_{ot}, \sigma_{ot}^2) = \frac{1}{(2\pi)^{\frac{1}{2}}\,\sigma_{ot}}\ exp\left[-\frac{(y_t-\mu_{ot})^2}{2\ \sigma_{ot}{}^2}\right]$$

$$p(\bar{x}_t \mid y_t) = N(\alpha+\beta\, y_t, \delta) = \frac{1}{(2\pi)^{\frac{1}{2}}\,(\delta)^{\frac{1}{2}}}\ exp\left[-\frac{(\bar{x}_t-\alpha-\beta\, y_t)^2}{2\ \delta}\right]$$

Changing the variable to $g_t = \frac{\bar{x}_t-\alpha}{\beta}$ in the likelihood function then gives

$$p(g_t \mid y_t) = \frac{\beta}{(2\pi)^{\frac{1}{2}}\,(\delta)^{\frac{1}{2}}}\ exp\left[-\frac{\beta^2(g_t-y_t)^2}{2\ \delta}\right]$$

which is a normal distribution for the random variable g_t with mean y_t and variance $\delta\,/\,\beta^2$:

$$p(y_t) = N(\mu_{ot}, \sigma_{ot}^2)$$

$$p(g_t \mid y_t) = N(y_t, \frac{\delta}{\beta^2})$$

This is the normal-normal Bayesian model in standard form. Using Bayes' theorem (Eqn. 3.3), this can be shown to have a posterior p.d.f., which is normal, with posterior precision (reciprocal variance) given by the sum of prior precision and likelihood precision (Lee 1997):

$$\frac{1}{\sigma_t^2} = \frac{1}{{\sigma_{ot}}^2} + \frac{\beta^2}{\delta}$$

while the posterior mean is the weighted average of prior mean and the rescaled forecast g_t, with weights given by the respective precisions. Substituting g_t by $\frac{\bar{x}_t - \alpha}{\beta}$ then gives

$$\frac{\mu_t}{\sigma_t^2} = \frac{\mu_{ot}}{{\sigma_{ot}}^2} + \frac{\beta^2}{\delta}\left(\frac{\bar{x}_t - \alpha}{\beta}\right)$$

Appendix C

Observational datasets

Historical (1950-2001) Niño-3.4 index data were obtained from Reynolds optimum interpolation version 2 SST dataset[1] (Reynolds *et al.*, 2002). This dataset was used for the verification of the Niño-3.4 index forecasts produced in chapter 4.

ERA-40 reanalysis[2] sea surface temperatures (referred to as observations) obtained from ECMWF were used for the verification of equatorial Pacific SST anomaly forecast presented in chapter 5.

A dataset of global monthly precipitation over land (PREC/L)[3] on a 2.5 degree latitude/longitude grid for a 50-year period from 1950 to 2001 (Chen *et al.* 2002) was used for the verification of DJF South American precipitation anomaly forecasts in chapter 6. This gridded field of monthly precipitation was produced by Chen *et al.* (2002) by interpolating gauge observations from the version 2 dataset of the Global Historical Climatology Network (GHCN) of NOAA/NCDC and the Climate Anomaly Monitoring System (CAMS) of NOAA/CPC using an optimum interpolation algorithm.

[1]Available at http://www.cpc.ncep.noaa.gov/data/indices/index.html

[2]ERA-40 provides global analysis of variables for the atmosphere, land and ocean surface for the period 1958-2001. More information is available at http://www.ecmwf.int/research/era/

[3]Available at ftp://ftp.ncep.noaa.gov/pub/precip/50yr/gauge/2.5deg

Appendix D

Brier score decomposition

Following Murphy (1973), the Brier score (Eqn. 4.6) can be expressed as the sum of three components:

$$BS = \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i(p_i - \bar{o}_i)^2}_{(reliability)} - \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i(\bar{o}_i - \bar{o})^2}_{(resolution)} + \underbrace{\bar{o}(1-\bar{o})}_{(uncertainty)}$$

where N_i is the number of times each probability forecast p_i is used in the set of forecasts being verified. The total number of forecast/event pairs n is simply the sum of these counts: $n = \sum_{i=1}^{I} N_i$, where I is the number of discrete forecast values p_i. For each probability p_i, depicted by the I allowable forecast values, there is a relative frequency $\bar{o}_i$ of the observed event. Since the observed event is dichotomous, a single conditional relative frequency defines the conditional distribution of observations given each forecast p_i. The subsample relative frequency, or conditional average observation, is given by

$$\bar{o}_i = p(o_1|p_i) = \frac{1}{N_i}\sum_{k \in N_i} o_k,$$

where $o_k = 1$ if the event occurs for the kth forecast/event pair and $o_k = 0$ if it does not occur, and the summation is over only those values of k corresponding to occasions when the forecast p_i was issued. Similarly, the overall (unconditional) relative frequency, or sample climatology, of the observations is given by

$$\bar{o} = \frac{1}{n}\sum_{k=1}^{n} o_k.$$

More accurate forecasts are characterized by small values of Brier score. Therefore, the forecaster would aim for the reliability component of the Brier score to be as small as possible, and the resolution component to be as large (in the absolute sense) as possible. The uncertainty component depends only on the sample climatological relative frequency $\bar{o}$, and is not affected by the forecasts.

The reliability component summarizes the calibration, or conditional bias, of the forecasts. It consists of a weighted average of the squared differences between the forecast probabilities p_i and the relative frequencies of the forecast event in each subsample i. For perfectly reliable forecasts the subsample relative frequency $\bar{o}_i$ is exactly equal to the forecast probability p_i in each subsample. The relative frequency of the forecast event $\bar{o}_i$ should be small when $p_i = 0$ is forecast, and should be large when $p_i = 1$ is forecast. When $p_i = 0.5$, $\bar{o}_i$ should be near 0.5. For reliable, or well-calibrated forecasts, all the squared differences in the reliability component of the Brier score will be near zero, and their weighted average will be small.

The resolution component summarizes the ability of the forecasts to discern subsample relative frequencies forecasts $\bar{o}_i$ from the observed overall sample climatology relative frequency $\bar{o}$. The forecast probabilities p_i do not appear explicitly in this term, yet it still depends on the forecasts through the sorting of the events making up the subsample relative frequency $\bar{o}_i$. The resolution component is a weighted average of the squared differences between $\bar{o}_i$ and $\bar{o}$. Thus, if the forecasts sort the observations into subsamples having substantially different relative frequencies $\bar{o}_i$ than the overall sample climatology $\bar{o}$, the resolution term will be large. This is a desirable situation, since the resolution component is subtracted in the Brier score decomposition equation. Conversely, if the forecasts sort the events into subsamples with very similar event relative frequencies $\bar{o}_i$, the squared differences in the summation of the resolution term will be small. In such a situation the forecasts resolve the event only weakly, and the resolution component will be small.

The uncertainty component depends only on the variability of the observations, and is not influenced by the forecasts. It has minima at zero when the climatological probability $\bar{o}$ is either zero or one, and a maximum of 0.25 when $\bar{o} = 0.5$. The uncertainty in the forecasting situation is small (close to zero) when the event being forecast almost always happens ($\bar{o}$ close to 1) or when the event being forecast almost never happens ($\bar{o}$ close to 0). In such situation, always forecasting the climatological probability $\bar{o}$ will give generally good results. When the climatological probability is close to 0.5, there is substantially more uncertainty inherent in the forecasting situation, and the uncertainty component of the Brier score is commensurately larger.

Appendix E

Reliability diagram

The three components of the Brier score can be interpreted geometrically using a device known as reliability diagram. Figure 5.3 in chapter 5 shows examples of this diagram. The dashed line in this diagram is obtained by joining the points defined by the subsample relative frequency $\bar{o}_i$ and the correspondent forecast probability p_i. Forecast probabilities have been ordered and grouped in 10 equally spaced probability bins from 0 to 1 and the points are plotted in the centre of each bin. For perfectly reliable forecasts the paired quantity ($\bar{o}_i$,p_i) should have exactly the same value, yielding all points falling on the solid diagonal line of the diagram. In other words, perfectly reliable probability forecasts p_i are those that are able to equate the observed relative frequency $\bar{o}_i$. For example, if one looks at a forecast probability of lets say $p_i = 0.7$ that was issued a sufficiently large number of times, and observes that in 70$\%$ of the forecast cases the event was observed, then the forecast is statistically consistent. If this consistency is noted for all possible forecast probabilities p_i (from 0 to 1) then the forecast is reliable. The histogram plot in the bottom right corner of the diagram shows a summary of the forecast probability frequency for the 10 equally spaced probability bins.

The reliability component of the Brier score is the weighted average of the squared vertical distances between the points of the dashed line and the points of the solid diagonal line. The closer the dashed line is to the diagonal line the better is the reliability of the forecasts. Therefore, the reliability can be geometrically measured by the area between the dashed and the diagonal lines. The smaller this area the better is the reliability of the forecasting system.

The geometric interpretation of the resolution component of the Brier score is achieved by the identification of subsamples with event relative frequencies $\bar{o}_i$ different from the climatological probability $\bar{o}$, which is represented in the reliability diagram by the horizontal dotted line. This corresponds to points in the reliability diagram being well removed (vertically) from the level of the overall sample climatology (horizontal dotted line). Forecasts for which the points fall on this horizontal line are unable to resolve occasions where the event is more or less likely than the overall sample climatology and therefore have no resolution. The weighted average constituting the resolution component is the squares of the vertical distances between the points of the dashed line and the points of the horizontal (no resolution) dotted line. These distances will be large for forecasts exhibiting good resolution. In this case the resolution component will contribute to a small (i.e., good) Brier score. Forecasts that are most different from the sample climatology probability make the largest contribution to the resolution component. The resolution can be geometrically measured by the area between the dashed and the horizontal dotted lines. The larger this area the better is the resolution of the forecasting system.

The uncertainty can be interpreted imagining the reliability diagram for climatological forecasts. In such situation the diagram is composed by a single dot, since only a single forecast value (the climatological probability) is ever used ($I = 1$). The horizontal position of this dot is at the climatological probability for the event, and if the sample size is sufficiently large, the long-term climatological probability will be close to the sample climatological relative frequency. The vertical position of the single dot will be at the sample climatological relative frequency, locating it at the intersection of the perfect reliability (diagonal solid) and the no resolution (horizontal dotted) lines. Therefore, climatological forecasts have perfect (zero) reliability, since the forecast p_i and the conditional relative frequency $\bar{o}_i$ are both equal to the climatological probability $\bar{o}$. Similarly, climatological forecasts have zero resolution since the existence of only $I = 1$ forecast category prevent discerning different subsets of forecasting occasions with differing relative frequencies of the outcomes. Since the reliability and resolution components are both zero, the Brier score for climatological forecasts is exactly the uncertainty.

Appendix F

3-month lead NDJ rainfall predictions

This appendix assesses the skill of 1959-2001 NDJ South American rainfall forecasts produced by empirical, multi-model and three different Bayesian combined and calibrated predictions obtained with forecast assimilation (FA, FAEP and FACE) as described in section 6.5. Empirical forecasts have been produced with the empirical model described in section 6.4.1 but here using the previous season MJJ (May-Jun-Jul) Pacific and Atlantic SST anomalies as predictor for NDJ South American rainfall. DEMETER coupled model predictions produced by ECMWF, CNRM and UKMO for NDF with initial conditions of the 1st of August (3-month lead) are used to produce the multi-model forecast. This ensures that empirical and coupled model predictions have temporal consistency of both predictors and predictands, and a fair comparison is performed.

Figure F.1 shows correlation and BSS maps of NDJ rainfall anomaly predictions produced by empirical, multi-model, FA, FAEP and FACE for the period 1959-2001. Figure F.2 shows the mean ACC of empirical, multi-model, FA, FAEP and FACE predictions for NDJ. The means ACC is computed for those La Niña, neutral and El Niño years listed in Table 6.1 and for all (1959-2001) years. Figure F.3 shows observed and predicted NDJ South American rainfall anomaly composites for La Niña and El Niño years produced by the empirical, multi-model, FA, FAEP and FACE systems. These figures reveal that: a) the tropics and the area of South Brazil, Uruguay, Paraguay and Northern Argentina are the two most predictable regions of South America; b) ENSO years are more predictable than neutral years, the latter having nearly null skill; c) empirical and coupled models have similar level of skill when seasonally forecasting NDJ South American rainfall anomalies at 3-month lead time; and d) Bayesian forecast assimilation improves forecast skill in terms of both ACC and BSS.

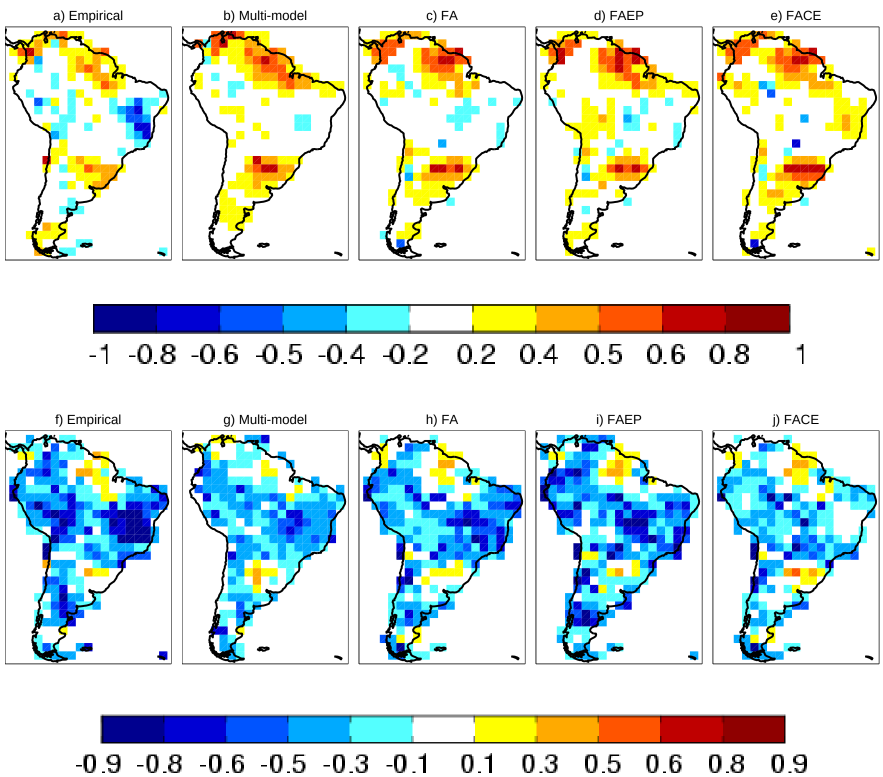


Figure F.1: Correlation and BSS maps of NDJ rainfall anomaly predictions for the period 1959-2001. The BSS is for the event 'rainfall anomalies less than or equal to zero'.

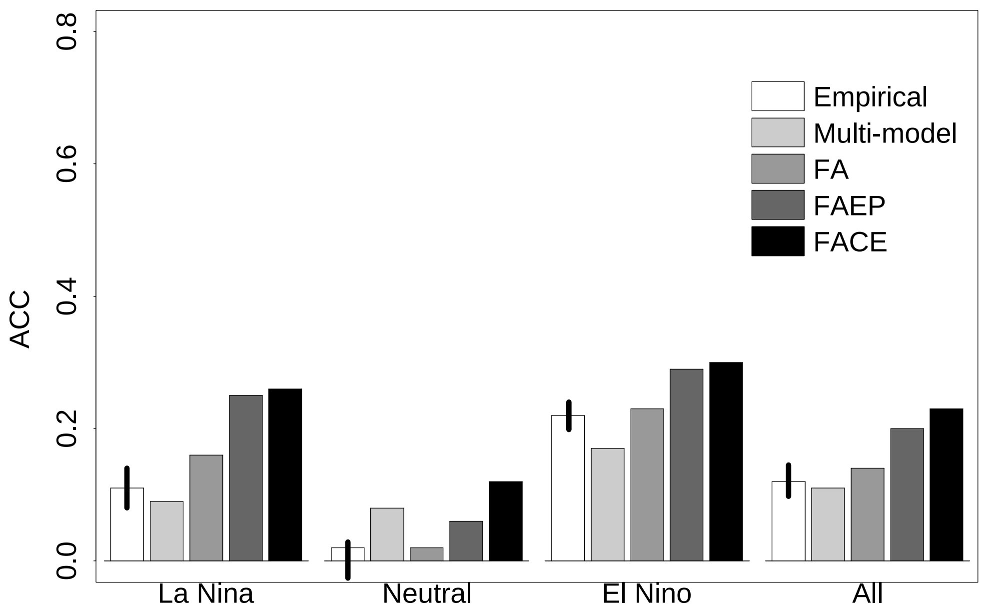


Figure F.2: NDJ mean anomaly correlation coefficient (ACC) for empirical, multi-model, FA, FAEP and FACE forecasts of La Niña, neutral, El Niño years (listed in Table 6.1) and all (1959-2001) years. The vertical solid lines on the top of the white bars indicate the 95% confidence interval for the mean ACC of empirical forecasts, which were obtained using a bootstrap resampling procedure.

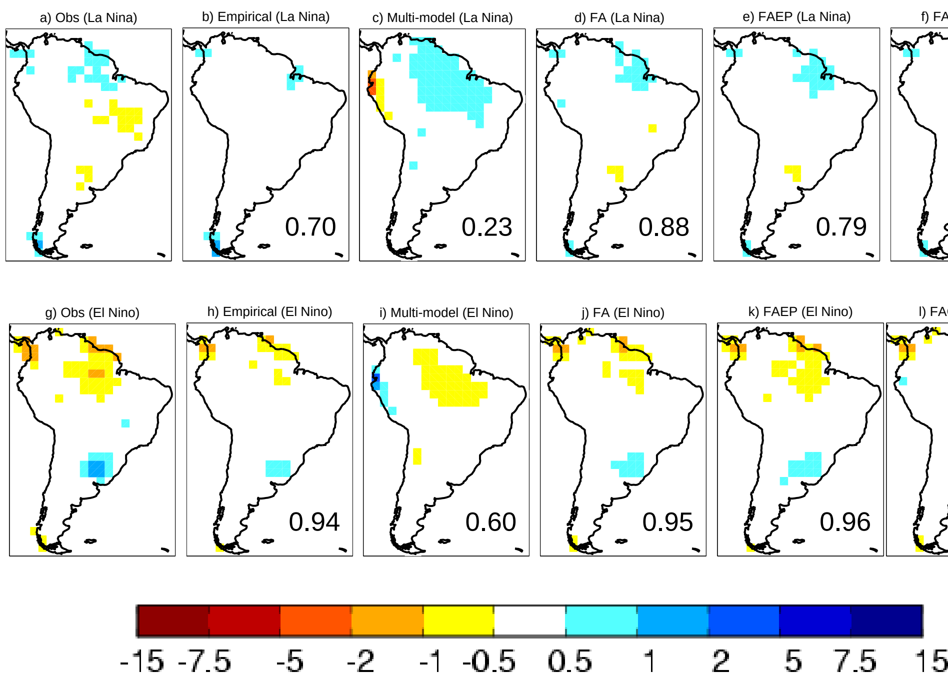


Figure F.3: NDJ South American rainfall anomaly composites ($mm.day^{-1}$) for those La Niña and El Niño years listed in Table 6.1. a) La Niña composite of observed rainfall anomalies . Panels b-f) Empirical, multi-model, FA, FAEP and FACE La Niña forecast composites. g) El Niño composite of observed rainfall anomalies. Panels h-l) Empirical, multi-model, FA, FAEP and FACE El Niño forecast composites. The number in the bottom right hand corner of panels b-f) and h-l) is the correlation between the observed (panels a and g) and the forecast composite.

Glossary of Acronyms

ACC	Anomaly Correlation Coefficient
AGCM(s)	Atmospheric General Circulation Model(s)
ASO	August-September-October
BS	Brier Score
BSS	Brier Skill Score
BMA	Bayesian Model Averaging
CAMS	Climate Anomaly Monitoring System
CCA	Canonical Correlation Analysis
CERFACS	European Centre for Research and Advanced Training in Scientific Computation
CGCM(s)	Coupled ocean-atmosphere General Circulation Model(s)
CNRM	Centre National de Recherches Météorologiques
CPC	Climate Prediction Center
DEMETER	Development of a European Multimodel Ensemble system for seasonal to inTERannual prediction
DJF	December-January-February
DSP	Dynamical Seasonal Prediction
ECMWF	European Centre for Medium-range Weather Forecasts
ENSO	El Niño-Southern Oscillation
ERA-40	ECMWF 40 years Re-Analysis
EU	European Union
FA	Forecast Assimilation
FACE	Forecast Assimilation of Coupled model and Empirical forecasts
FAEP	Forecast Assimilation with Empirical Prior
GCM(s)	General Circulation Model(s)

GHCM	Global Historical Climatology Network
GMT	Greenwich Mean Time
hPa	Hecto Pascal
INGV	Istituto Nazionale de Geofisica e Vulcanologia
IRI	International Research Institute for Climate Prediction
ITCZ	Intertropical Convergence Zone
JJA	June-July-August
LODYC	Laboratoire d’ Océanographie Dynamique et de Climatologie
MAE	Mean Absolute Error
MAM	March-April-May
MAP	Maximum a Posteriori
MAX	Maximum
MCA	Maximum Covariance Analysis
MCS	Mesoscale Convective Systems
MIN	Minimum
MJJ	May-June-July
ML	Maximum Likelihood
MOS	Model Output Statistics
MPI	Max-Plank Institut für Meteorologie
MSE	Mean Square Error
NCEP	National Centers for Environmental Prediction
NDJ	November-December-January
NOAA	National Oceanic and Atmospheric Administration
p.d.f.	Probability density function
OLS	Ordinary Least squares
P.I.	Prediction Interval
PNA	Pacific North American
PREC/L	Precipitation Reconstruction over land
PROVOST	Prediction of Climate Variations on Seasonal to Interannual Time-scales
PSA	Pacific South American

RMSE	Root Mean Square Error
SACZ	South Atlantic Convergence Zone
SCF	Squared Covariance Fraction
SLP	Sea Level Pressure
SON	September-October-November
SSCP	Subtropical South-Central Pacific
SST(s)	Sea Surface Temperature(s)
SVD	Singular Value Decomposition
UKMO	United Kingdom Meteorological Office

Glossary of Notation

μ mean

σ^2 variance

σ standard deviation

r correlation

R^2 coefficient of determination

b_1 moment measure of skewness

$|$ standard statistical symbol denoting *"given"* (conditional upon)

$\sim$ standard statistical symbol denoting *"is distributed as"*

N standard statistical symbol denoting a *"Normal"* (Gaussian) *p.d.f.*

$\hat{}$ standard statistical symbol denoting an estimated quantity or parameter

E standard statistical symbol denoting expectation (population mean)

$p(.)$ probability density function (*p.d.f.*)

$Pr(E)$ probability of an event E

$\Phi(y_*)$ area under the standard normal curve to the left of y_*

m total number of members of an ensemble forecast

n number of sample objects in the calibration period

t time for which a forecast is issued

x model state or model prediction data

$\bar{x}$ ensemble mean forecast or historical mean of x

$\overline{\overline{x}}$ sample mean of $\bar{x}$

$\bar{x}_t$ ensemble mean forecast for time t

y observational data or future observable

$\bar{y}$	sample mean of y
y_t	observable variable at time t
z	observational SST data
$\bar{z}$	sample mean of z
Z_t	standardized forecast errors
μ_{ot}	mean of the prior distribution of the univariate Normal model
${\sigma_{ot}}^2$	variance of the prior distribution of the univariate Normal model
μ_t	mean of the posterior distribution of the univariate Normal model
${\sigma_t}^2$	variance of the posterior distribution of the univariate Normal model
α	intercept parameter of the linear regression of $\bar{x}_t$ on y_t
β	slope parameter of the linear regression of $\bar{x}_t$ on y_t
δ	constant variance parameter of the linear regression of $\bar{x}_t$ on y_t
ϵ_t	residuals of the linear regression of $\bar{x}_t$ on y_t
$s_{\bar{x}}$	sample standard deviation of $\bar{x}_t$
s_y	sample standard deviation of y_t
s_x	sample standard deviation of an ensemble forecast composed by m members
a	intercept parameter of the linear regression of y_t on $\bar{x}_t$
b	slope parameter of the linear regression of y_t on $\bar{x}_t$
λ	constant variance parameter of the linear regression of y_t on $\bar{x}_t$
e_t	residuals of the linear regression of y_t on $\bar{x}_t$
ψ_t	July Niño-3.4 monthly values
$\bar{\psi}_t$	sample mean of ψ_t
S_t^2	standard deviation of ψ_t

β_0	intercept of the linear regression of December on July (ψ_t) Niño-3.4
β_1	slope of the linear regression of December on July (ψ_t) Niño-3.4
ϵ'_t	residuals of the linear regression of December on July (ψ_t) Niño-3.4
p	model space dimension
q	observation space dimension
v	observed SST space dimension
F^T	transpose of matrix F
F^{-1}	inverse of matrix F
X	matrix of model prediction anomalies
Y	matrix of observable anomalies
Z	matrix of observed SST anomalies
S_{xx}	covariance matrix of model predictions
S_{yy}	covariance matrix of observables
S_{zz}	covariance matrix of observed SST
S_{xy}	cross-covariance matrix
S_{yx}	cross-covariance matrix
S_{yz}	cross-covariance matrix
S_{zy}	cross-covariance matrix
Σ	covariance of the multivariate Normal distribution
U	matrix of Empirical Orthogonal Functions
V	matrix of Principal Components
Σ^*	matrix of singular values

	Data assimilation		Forecast Assimilation
x_a	analysis model state	y_a	forecast observable state
x_b	background model state	y_b	background observable state
A	analysis error covariance	D	forecast error covariance
B	background forecast error covariance	C	background observable covariance
H	observation operator	G	forecast operator
K	gain/weight matrix	L	gain/weight matrix
R	observation error covariance	S	forecast error covariance
$J_{x\|y}$	cost function	$J_{y\|x}$	cost function

ϵ_B normally distributed errors with zero mean and covariance B

ϵ_R normally distributed errors with zero mean and covariance R

ϵ_C normally distributed errors with zero mean and covariance C

ϵ_S normally distributed errors with zero mean and covariance S

M empirical prediction operator

T empirical prediction covariance

ϵ_T normally distributed errors with zero mean and covariance T

y_o bias vector

z_o bias vector

k numbering of the n forecast/observation pairs

p_k forecast probability

o_k binary observation

$\bar{o}$ sample climatological mean of binary observation

References

Aitchison, J. and I. R. Dunsmore, 1975. Statistical prediction analysis. *Cambridge University Press*. 273 pp.

Alves, O., G. Wang, A. Zhong, N. Smith, G. Warren, A. Marshall, F. Tzeitkin, and A. Schiller, 2002. POAMA: Bureau of Meteorology operational coupled model seasonal forecast system. *Proceeding of the ECMWF Workshop on the Role of the Upper Ocean in Medium and Extended Range Forecasting*, pages 22–32.

Anderson, J. L., 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**:1518–1530.

Anderson, J. L., H. van den Dool, A. Barnston, W. Chen, W. Stern, and J. Ploshay, 1999. Present-day capabilities of numerical and statistical models for atmospheric extratropical seasonal simulation and prediction. *Bull. Am. Meteorol. Soc.*, **80**:1349–1361.

Atger, F., 2003. Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.*, **131**: 1509–1523.

Balmaseda, M. A., D. L. T. Anderson, and M. K. Davey, 1994. ENSO prediction using a dynamical ocean model coupled to statistical atmospheres. *Tellus*, **46A**: 497–511.

Balmaseda, M. A., M. K. Davey, and D. L. T. Anderson, 1995. Decadal and seasonal dependence of ENSO prediction skill. *J. Climate*, **8**:2705–2715.

Barnard, G. A., 1963. New methods of quality control. *Journal of the Royal Statistical Society, Series A*, **126**:255–259.

Barnston, A. G., M. H. Glantz, and Y. He, 1999a. Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997-98 El Niño episode and the 1998 La Niña onset. *Bull. Am. Meteorol. Soc.*, **80**:217–243.

Barnston, A. G., A. Leetmaa, V. E. Kousky, R. E. Livezey, E. A. O'Lenic, H. van den Dool, A. J. Wagner, and D. A. Unger, 1999b. NCEP forecasts of the El Niño of 1997-1998 and its US impacts. *Bull. Am. Meteorol. Soc.*, **80**: 1829–1852.

Barros, V. and G. E. Silvestri, 2002. The relation between sea surface temperature at the subtropical south-central Pacific and precipitation in southeastern South America. *J. Climate*, **15**:251–267.

Bates, J. M. and C. W. J. Granger, 1969. The combination of forecasts. *Operational Research Quarterly*, **20**:451–468.

Battisti, D. S., 1998. Dynamics and thermodynamics of a warming event in a coupled atmosphere-ocean model. *J. Atmos. Sci.*, **45**:2889–2919.

Berkson, J., 1969. Estimation of a linear function for a calibration line: Consideration of a recent proposal. *Technometrics*, **11**:649–660.

Berliner, L. M., R. A. Levine, and D. J. Shea, 2000a. Bayesian climate change assessment. *J. Climate*, **13**:3805–3820.

Berliner, L. M., C. K. Wikle, and N. Cressie, 2000b. Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *J. Climate*, **13**:3953–3968.

Bjerknes, J., 1966. A possible response of the atmospheric Hadley circulation to equatorial anomalies of temperature. *Tellus*, **18**:820–829.

Bouttier, F. and P. Courtier, 1999. Data assimilation concepts and methods. *ECMWF Meteorological Training Course Lecture Series*. 59 pp. Available from http://www.ecmwf.int.

Brankovic, C., T. N. Palmer, and L. Ferranti, 1994. Predictability of seasonal atmospheric variations. *J. Climate*, **7**:217–237.

Brier, G. W., 1950. Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**:1–3.

Brown, P. J., 1982. Multivariate calibration. *Journal of the Royal Statistical Society, Series B*, **44**(3):287–321.

Brown, P. J., 1994. Measurement, regression and calibration. *Oxford Science Publications*. Oxford Statistical Science Series, **12**. 210 pp.

Brunet, N., R. Verret, and N. Yacowar, 1988. An objective comparison of model output statistics and ”perfec prog” systems in producing numerical weather element forecasts. *Weather and Forecasting*, **3**:273–283.

Bunn, D. W., 1975. A Bayesian approach to the linear combination of forecasts. *Operational Research Quarterly*, **26**:325–329.

Bunn, D. W., 1985. Statistical efficiency on the linear combination of forecasts. *Int. Journal of Forecasting*, **1**:151–163.

Burgers, G. and D. B. Stephenson, 1999. The ”Normality” of El Niño. *Geophysical Research Letters*, **26**(8):1027–1030.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989. Statistical forecasts based on the National Meteorological Center’s numerical weather prediction system. *Weather and Forecasting*, **4**:401–412.

Cavalcanti, I. F. A., J. A. Marengo, P. Satyamurty, C. A. Nobre, I. Trosnikov, J. P. Bonatti, A. O. Manzi, T. Tarasova, L. P. Pezzi, C. D’Almeida, G. Sampaio, C. C.

Castro, M. B. Sanches, and H. Camargo, 2002. Global climatological features in a simulation using the CPTEC-COLA AGCM. *J. Climate*, **15**(21):2965–2988.

Challinor, A. J., J. M. Slingo, T. R. Wheeler, and F. J. Doblas-Reyes, 2005. Probabilistic simulations of crop yield over western India using the DEMETER seasonal hindcast ensembles. *Tellus*. In press.

Charney, J. G. and J. Shukla, 1981. Predictability of monsoons. *Monsoon Dynamics*, pages 99–109. Cambridge University Press. Editors: Lighthill, J. and Pearce, R.

Chatfield, C., 2001. Time-series forecasting. *Chapman and Hall/CRC*. 267 pp.

Chen, M., P. Xie, J. E. Janowiak, and P. A. Arkin, 2002. Global land precipitation: A 50-yr monthly analysis based on gauge observations. *J. Hydrometeorol.*, **3**: 249–266.

Chow, S. and J. Shao, 1990. On the difference between the classical and inverse methods of calibration. *Appl. Statist.*, **39**(2):219–228.

Clemen, R. T., 1985. Extraneous expert information. *Journal of Forecasting*, **4**: 329–348.

Clemen, R. T., 1989. Combining forecasts: A review and annotated bibliography. *Int. Journal of Forecasting*, **5**:559–583.

Clemen, R. T. and A. H. Murphy, 1986a. Objective and subjective precipitation probability forecasts: statistical analysis of some interrelationships. *Weather and Forecasting*, **1**:56–65.

Clemen, R. T. and A. H. Murphy, 1986b. Objective and subjective precipitation probability forecasts: some methods for improving forecast quality. *Weather and Forecasting*, **1**:213–218.

Coelho, C. A. S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes, and D. B. Stephenson, 2003. Skill of coupled model seasonal forecasts: A Bayesian assessment of

ECMWF ENSO forecasts. *European Centre for Medium-Range Weather Forecasts Technical Memorandum number 426*. 16 pp.

Coelho, C. A. S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes, and D. B. Stephenson, 2004. Forecast calibration and combination: A simple Bayesian approach for ENSO. *J. Climate*, **17**:1504–1516.

Coelho, C. A. S., C. B. Uvo, and T. Ambrizzi, 2002. Exploring the impacts of the tropical Pacific SST on the precipitation patterns over South America during ENSO periods. *Theor. Appl. Climatol.*, **71**:185–197.

Colman, A. W. and M. K. Davey, 2003. Statistical prediction of global sea-surface temperature anomalies. *Int. J. Climatol.*, **23**:1677–1697.

Conte, M., C. DeSimone, and C. Finizio, 1980. Post-processing of numerical models: Forecasting the maximum temperature at Milano Linate. *Rev. Meteorol. Aeronaut.*, **40**:247–265.

Courtier, P., 1997. Dual formulation of four-dimensional variational assimilation. *Q. J. R. Meteorol. Soc.*, **123**:2449–2461.

Daley, R., 1991. Atmospheric data analysis. *Cambridge Atmospheric and Space Science Series*. Cambridge University Press. ISBN 0-521-38215-7, 457 pp.

Davey, M. K., M. Huddleston, K. R. Sperber, P. Braconnot, F. Bryan, D. Chen, R. A. Colman, C. Cooper, U. Cubasch, P. Delecluse, D. DeWitt, L. Fairhead, G. Flato, C. Gordon, T. Hogan, M. Ji, M. Kimoto, A. Kitoh, T. R. Knutson, M. Latif, H. Le Treut, T. Li, S. Manabe, C. R. Mechoso, G. A. Meehl, S. B. Power, E. Roeckner, L. Terray, A. Vintzileos, R. Voss, B. Wang, W. M. Washington, I. Yoshikawa, J.-Y. Yu, S. Yukimoto, and S. E. Zebiak, 2002. STOIC: a study of coupled model climatology and variability in tropical ocean regions. *Clim. Dynamics*, **18**:403–420.

Davey, M. K., S. Ineson, and M. A. Balmaseda, 1994. Simulation and hindcasts of tropical Pacific Ocean interannual variability. *Tellus*, **46A**:433–447.

Davis, R. E., 1976. Predictability of sea surface temperature and sea level pressure anomalies over the north Pacific ocean. *J. Phys. Oceanography*, **6**:249–266.

de Menezes, L. M., D. W. Bunn, and J. W. Taylor, 2000. Review guidelines for the use of combined forecasts. *European Journal of Operational Research*, **120**: 190–204.

Délécluse, P., M. K. Davey, Y. Kitamura, S. G. H. Philander, M. Suarez, and L. Bengtsson, 1998. Coupled general circulation modeling of the tropical Pacific. *J. Geophys. Res.*, **103**:14357–14373.

Derome, J., G. Brunet, A. Plante, N. Gagnon, G. J. Boer, F. W. Zwiers, S. J. Lambert, J. Sheng, and H. Ritchie, 2001. Seasonal predictions based on two dynamical models. *Atmosphere-Ocean*, **39**:485–501.

Diaz, A. F., C. D. Studzinski, and C. R. Mechoso, 1998. Relationships between precipitation anomalies in Uruguay and Southern Brazil and sea surface temperature in the Pacific and Atlantic oceans. *J. Climate*, **11**:251–271.

Diebold, F. X. *Elements of forecasting*. 1998, Cincinnati: South-Western College Publishing, 1998.

Diebold, F. X. and J. A. Lopez. *Forecast evaluation and combination*, volume 14. 1996, Amsterdam: North-Holland, 1996. Handbook of Statistics, (Eds.) Maddala, G. S. and Rao, C. R. pp. 241-248.

Dirmeyer, P. A., 2003. The role of the land surface background state in climate predictability. *J. Hydrology*, **4**:599–610.

Doblas-Reyes, F. J., M. Déqué, and J.-P. Piedelièvre, 2000. Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Q. J. R. Meteorol. Soc.*, **126**: 2069–2088.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting. Part II: Calibration and combination. *Tellus*. In press.

Doblas-Reyes, F. J., V. Pavan, and D. B. Stephenson, 2003. The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation. *Clim. Dynamics*, **21**:501–514.

Douville, H., 2003. Assessing the influence of soil moisture on seasonal climate variability with AGCMs. *J. Hydrology*, **4**:1044–1066.

Douville, H. and J.-F. Royer, 1996. Sensitivity of the Asian summer monsoon to anomalous Eurasian snow cover within the Météo-France GCM. *Clim. Dynamics*, **12**:449–466.

Draper, N. R. and H. Smith, 1998. Applied regression analysis. *John Wiley and Sons, Inc*. Third edition. 706 pp.

Drought Monitoring Centre - Nairobi, 1998. Climate Outlook Forum for Eastern Africa Sub-region for the period ending December 1998. *Available from the Drought Monitoring Centre - Nairobi, Kenya*. 58 pp.

Eisenhart, C., 1939. The interpretation of certain regression methods and their use in biological and industrial research. *Ann. Math. Statist.*, **10**:162–186.

Epstein, E. S., 1962. A Bayesian approach to decision making in applied meteorology. *J. Appl. Meteorol.*, **1**:169–177.

Epstein, E. S., 1985. Statistical inference and prediction in climatology: A Bayesian approach. *Meteorological Monographs*, **20**(42). Amer. Meteor. Soc., 199 pp.

Feddersen, H., 2003. Predictability of seasonal precipitation in the Nordic region. *Tellus*, **55A**:385–400.

Feddersen, H. and U. Andersen, 2005. A method for statistical downscaling of seasonal ensemble predictions. *Tellus*. In press.

Feddersen, H., A. Navarra, and M. N. Ward, 1999. Reduction of model systematic error by statistical correction for dynamical seasonal predictions. *J. Climate*, **12**: 1974–1989.

Folland, C. K., A. W. Colman, D. P. Rowell, and M. K. Davey, 2001. Predictability of Northeast Brazil rainfall and real-time forecast skill, 1987-98. *J. Climate*, **14** (9):1937–1958.

Fraedrich, K. and L. M. Leslie, 1987. Combining predictive schemes in short-term forecasting. *Mon. Wea. Rev.*, **115**:1640–1644.

Fraedrich, K. and N. R. Smith, 1989. Combining predictive schmes in long-range forecasting. *J. Climate*, **2**:291–294.

Francis, P. E., A. P. Day, and G. P. Davis, 1982. Automated temperature forecasting, and application of Model Output Statistics to the Meteorological Office numerical weather prediction model. *Meteorol. Mag.*, **111**:73–87.

Francis, R. I. C. C. and J. A. Renwick, 1998. A regression-based assessment of the predictability of New Zealand climate anomalies. *Theor. Appl. Climatol.*, **60**: 21–36.

Geisser, S., 1965. A Bayes approach for combining correlated estimates. *Journal of the American Statistical Association, Series A*, **60**:602–607.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 1995. Bayesian Data Analysis. *Chapman and Hall/CRC.* Second Edition. 526 pp.

Glahn, B., 2004. Discussion of verification concepts in Forecast Verification: A Practitioner's Guide in Atmospheric Science. *Weather and Forecasting*, **19**:769–775.

Glahn, H. R. and D. A. Lowry, 1972. The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.*, **11**:1203–1211.

Goddard, L. and N. E. Graham, 1999. The importance of the Indian ocean for simulating rainfall anomalies over eastern and southern Africa. *J. Geophys. Res.*, **104**:19099–19116.

Goddard, L., S. J. Mason, and A. W. Robertson, 2003. Multi-model ensembling: Combining and refining. *Geophysical research abstracts*, **5**(11216). European Geophysical Society Conference. Nice - France.

Goddard, L., S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane, 2001. Current approaches to seasonal-to-interannual climate predictions. *Int. J. Climatol.*, **21**:1111–1152.

Graham, N. E. and T. P. Barnett, 1987. Sea surface temperature, surface wind divergence, and convection over tropical oceans. *Science*, **238**:657–659.

Graham, R. J., A. Evans, K. Mylne, M. Harrison, and K. B. Robertson, 2000. An assessment of seasonal predictability using atmospheric general circulation models. *Q. J. R. Meteorol. Soc.*, **126**:2211–2240.

Granger, C. W. J., 1989. Combining forecasts - Twenty years later. *Journal of Forecasting*, **8**:167–173.

Granger, C. W. J. and R. Ramanathan, 1984. Improved methods of forecasting. *Journal of Forecasting*, **3**:197–204.

Greischar, L. and S. Hastenrath, 2000. The rainy seasons of the 1990s in Northeast Brazil: Real-time forecasts and verification. *J. Climate*, **13**(21):3821–3826.

Grimm, A. M., V. R. Barros, and M. E. Doyle, 2000. Climate variability in southern South America associated with El Niño and La Niña events. *J. Climate*, **13**:35–38.

Halperin, M., 1970. On inverse estimation in linear regression. *Technometrics*, **12**: 727–736.

Hamill, T. and S. J. Colucci, 1998. Verification of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**:711–724.

Hannachi, A., D. B. Stephenson, and K. R. Sperber, 2003. Probability-based methods for quantifying nonlinearity in the ENSO. *Clim. Dynamics*, **20**:241–256.

Hannachi, A., D. B. Stephenson, and K. R. Sperber, 2004. Corrigendum: Probability-based methods for quantifying nonlinearity in the ENSO. *Clim. Dynamics*, **22**:69–70.

Hastenrath, S., L. Greischar, and J. van Heerden, 1995. Prediction of the summer rainfall over South Africa. *J. Climate*, **8**:1511–1518.

Hoadley, B., 1970. A Bayesian look at inverse linear regression. *Journal of the American Statistical Association, Series A*, **65**:356–369.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999. Bayesian model averaging: a tutorial. *Statistical Science*, **14**(4):382–417.

Holden, K., D. A. Pell, and J. L. Thomson. *Economic forecasting: An introduction*. 1990, Cambridge University Press, Cambridge, 1990.

Horel, J. D. and J. M. Wallace, 1981. Planetary-scale atmospheric phenomena associated with the Southern Oscillation. *Mon. Wea. Rev.*, **109**:813–829.

Huang, J., H. M. van den Dool, and A. G. Barnston, 1996. Long-lead seasonal temperature prediction using optimal climate normals. *J. Climate*, **9**:809–817.

Ineson, S. and M. K. Davey, 1997. Interannual climate simulation and predictability in coupled TOGA GCM. *Mon. Wea. Rev.*, **125**:721–741.

Jolliffe, I. N., 2002. Principal Component Analysis. *Springer Verlag*. Second edition. 487pp.

Jolliffe, I. N. and D. B. Stephenson, 2003. Forecast Verification: A practitioner's guide in atmospheric science. *Wiley and Sons*. First edition. 240 pp.

Kanamistu, M., A. Kumar, H.-M. H. Juang, K. Schemm, J, W. Wang, F. Yang, S.-Y. Hong, P. Peng, W. Chen, S. Moorthi, and M. Ji, 2002. NCEP dynamical seasonal forecast system 2000. *Bull. Am. Meteorol. Soc.*, **83**:1019–1037.

Karoly, D. J., 1989. Southern Hemisphere circulation features associated with El Niño-Southern Oscillation events. *J. Climate*, **2**:1239–1252.

Kharin, V. V. and F. W. Zwiers, 2002. Climate predictions with multimodel ensembles. *J. Climate*, **15**:793–799.

Kiladis, G. N. and H. F. Diaz, 1989. Global climate anomalies associated with extremes in the Southern Oscillation. *J. Climate*, **2**:1069–1090.

Kodama, Y., 1992. Large-scale common features of subtropical precipitation zones (the Baiu frontal zone, the SPCZ, and the SACZ) Part I: Characteristics of subtropical frontal zones. *J. Meteorol. Soc. Japan*, **70**(4):813–836.

Krishnamurti, T. N., C. M. Kishtawal, T. LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999. Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**:1548–1550.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000a. Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**:4196–4216.

Krishnamurti, T. N., D. W. Shin, and C. E. Williford, 2000b. Improving tropical precipitation forecasts from a multianalysis superensemble. *J. Climate*, **13**:4217–4227.

Krishnamurti, T. N., S. Surendran, D. W. Shin, R. J. Correa-Torres, T. S. V. V. Kumar, E. Williford, C. Kummeow, R. F. Adler, J. Simpson, R. Kakar, W. S. Olson, and F. J. Turk, 2001. Real-time multianalysis-multimodel superensemble forecasts of precipitation using TRMM and SSM/I products. *Mon. Wea. Rev.*, **129**:2861–2883.

Krutchkoff, R. G., 1967. Classical and inverse methods of calibration. *Technometrics*, **9**:525–539.

Krutchkoff, R. G., 1969. Classical and inverse methods of calibration in extrapolation. *Technometrics*, **11**:605–608.

Krzysztofowicz, R., 1983. Why should a forecaster and a decision maker use Bayes theorem. *Mon. Wea. Rev.*, **19**:327–336.

Krzysztofowicz, R. and H. D. Herr, 2001. Hydrologic uncertainty processor for probabilistic river stage forecasting: precipitation-dependent model. *J. Hydrology*, **249**:46–68.

Kumar, A., A. Barnston, P. Peng, and M. P. a. G. L. Hoerling, 2000. Changes in the spread of the variability of the seasonal mean atmospheric states associated with enso. *J. Climate*, **13**:3139–3151.

Landman, W. A. and S. J. Mason, 1999. Operational prediction of south african rainfall using canonical correlation analysis. *Int. J. Climatol.*, **19**:1073–1090.

Landsea, C. and A. Knaff, 2000. How much skill was there in forecasting the very strong 1997-1998 El Niño? *Bull. Am. Meteorol. Soc.*, **8**:2107–2119.

Laplace, P. S., 1818. Deuxième supplément a la théorie analytique des probabilités. Paris: Courcier. Reprinted (1847) in *Oeuvres de Laplace* **7**, pp. 569-623. Paris: Imprimerie Royale; (1886) in *Oeuvres Complètes de Laplace* **7**, pp. 531-80. Paris: Gauthier-Villars.

Lee, P. M., 1997. Bayesian statistics: An introduction. *Arnold*. Second edition. 344 pp.

Lemcke, C. and S. Kruizinga, 1988. Model output statistics forecasts: Three years of operational experience in the Netherlands. *Mon. Wea. Rev.*, **116**:1077–1090.

Lorenc, A., 1986. Analysis methods for numerical weather prediciton. *Q. J. R. Meteorol. Soc.*, **112**:1177–1194.

Lorenz, E. N., 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.*, **42**:433–471.

Marengo, J., 1992. Interannual variability of surface climate in the Amazon basin. *Int. J. Climatol.*, **12**:853–863.

Marengo, J. A., I. F. A. Cavalcanti, P. Satyamurty, I. Trosnikov, C. A. Nobre, J. P. Bonatti, H. Camargo, G. Sampaio, M. B. Sanches, A. Manzi, C. A. C. Castro, C. D'Almeida, L. P. Pezzi, and L. Candido, 2003. Assessment of regional seasonal rainfall predictability using the CPTEC/COLA atmospheric GCM. *Clim. Dynamics*, **21**(5-6):459–475.

Martis, A., G. J. van Oldenborgh, and G. Burgers, 2002. Predicting rainfall in the Dutch Caribbean – More than El Niño. *Int. J. Climatol.*, **22**:1219–1234.

Mason, S. J., 2004. On using "climatology" as a reference strategy in the Brier and Ranked Probability Skill Scores. *Mon. Wea. Rev.*, **132**(7):1891–1895.

Mason, S. J. and G. M. Mimmack, 2002. Comparison of some statistical methods of probabilistic forecasting of ENSO. *J. Climate*, **15**:8–29.

Mason, S. J. L., L. Goddard, N. E. Graham, E. Yulaeva, L. Sun, and P. A. Arkin, 1999. The IRI seasonal climate prediction system and the 1997/198 El Niño event. *Bull. Am. Meteorol. Soc.*, **80**:1853–1873.

McPhaden, M. J., 1995. The Tropical Atmosphere-Ocean array is completed. *Bull. Am. Meteorol. Soc.*, **76**:739–741.

McPhaden, M. J., A. J. Busalacchi, R. Cheney, J.-R. Donguy, K. S. Gage, D. Halpern, M. Ji, P. Julian, G. Meyers, G. T. Mitchum, P. P. Niiler, J. Picaut, W. Reynolds, N. Smith, and K. Takeuchi, 1998. The Tropical Ocean-Global Atmosphere observing system: A decade of progress. *J. Geophys. Res.*, **103**: 14169–14240.

McPhaden, M. J., H. P. Freitag, S. P. Hayes, B. A. Taft, and K. Wyrtki, 1988. The response of the equatorial Pacific Ocean to a westerly wind burst in May 1986. *J. Geophys. Res.*, **93**:10589–10603.

Mechoso, C. R., S. W. Lyons, and J. A. Spahr, 1990. The impact of sea surface temperature anomalies on the rainfall over Northeast Brazil. *J. Climate*, **3**:812–826.

Metzger, S., M. Latif, and K. Fraedrich, 2004. Combining ENSO forecasts: A feasibility study. *Mon. Wea. Rev.*, **132**:456–472.

Miller, L., R. E. Cheney, and B. C. Douglas, 1988. Geosat altimeter observations of Kelvin waves and the 1986-87 El Niño. *Science*, **239**:52–54.

Molteni, F., R. Buizza, P. T. N., and T. Petroliagis, 1996. The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.*, **122** (529):73–119.

Morse, A. P., F. J. Doblas-Reyes, M. B. Hoshen, R. Hagedorn, M. C. Thomson, and T. N. Palmer, 2005. First steps towards the integration of a dynamic malaria model within a probabilistic multi-model forecast system. *Tellus*. In press.

Moura, A. D. and S. Hastenrath, 2004. Climate prediction for Brazil's Nordeste: Performance of empirical and numerical modeling methods. *J. Climate*, **17**(13): 2667–2672.

Moura, A. D. and J. Shukla, 1981. On the dynamics of droughts in northeast Brazil: Observations, theory and numerical experiments with general circulation model. *J. Atmos. Sci.*, **38**:2653–2675.

Murphy, A. H., 1973. A new vector partition of the probability score. *J. Appl. Meteorol.*, **12**:595–600.

Mylne, K. R., R. E. Evans, and R. T. Clark, 2002. Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Q. J. R. Meteorol. Soc.*, **128**:361–384.

Neelin, J. D., D. S. Battisti, A. C. Hirst, F.-F. Jin, Y. Wakata, T. Yamagata, and S. E. Zebiak, 1998. ENSO theory. *J. Geophys. Res.*, **103**:14261–14290.

Newbold, P. and C. W. J. Granger, 1974. Experience with forecasting univariate time series and the combination of forecasts (with discussion). *Journal of the Royal Statistical Society, Series A*, **137**:131–149.

Nobre, P. and J. Shukla, 1996. Variations of sea surface temperature, wind stress, and rainfall over the tropical atlantic and south america. *J. Climate*, **9**:2464–2479.

Paegle, J. N. and K. C. Mo, 2002. Linkages between summer rainfall variability over South America and sea surface temperature anomalies. *J. Climate*, **15**:1389–1407.

Palm, F. C. and A. Zellner, 1992. To combine or not to combine? Issues of combining forecasts. *Journal of Forecasting*, **11**:687–701.

Palmer, T. N., 2000. Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.*, **63**:71–116.

Palmer, T. N., 2002. The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Q. J. R. Meteorol. Soc.*, **128**:747–774.

Palmer, T. N., A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, P. Délécluse, M. Déqué, E. Díez, F. J. Doblas-Reyes, H. Feddersen, R. Graham, S. Gualdi, J.-F. Guérémy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A. P. Morse, B. Orfila, P. Rogel, J.-M. Terres, and M. C. Thomson, 2004. Development of a European Multi-Model Ensemble System for Seasonal to Inter-Annual Prediction (DEMETER). *Bull. Am. Meteorol. Soc.*, **85**:853–872.

Palmer, T. N., C. Brankovic, and D. S. Richardson, 2000. A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Q. J. R. Meteorol. Soc.*, **126**:2013–2034.

Palmer, T. N. and J. Shukla, 2000. Editorial to DSP/PROVOST special issue. *Q. J. R. Meteorol. Soc.*, **126**:1989–1990.

Patt, A., 2000. Communicating probabilistic forecasts to decision makers: A case study of Zimbabwe. *Belfer Center for Science and International Affairs (BCSIA), Environment and Natural Resources Program, Kennedy School* *of Government, Harvard University*. Discussion paper 2000-19. Available from http://environment.harvard.edu/gea.

Pavan, V. and F. J. Doblas-Reyes, 2000. Multi-model seasonal hindcasts over the Euro-Atlantic: skill scores and dynamic features. *Clim. Dynamics*, **16**:611–625.

Pezzi, L., V. Ubarana, and C. Repelli, 2000. Forecast and performance of a statistical regional model for seasonal climate prediction in Southern Brazil. *Rev. Bras. Geof.*, **18**(2):129–146.

Philander, S. G. H., 1983. Anomalous El Niño of 1982-83. *Nature,* **305**:16.

Philander, S. G. H., 1985. El Niño and La Niña. *J. Atmos. Sci.*, **42**:2652–2662.

Philander, S. G. H., 1990. El Niño, La Niña and the Southern Oscillation. *Princeton University, New Jersey. Academic Press*. 291 pp.

Philander, S. G. H., T. Yamagata, and R. C. Pacanowski, 1984. Unstable air-sea interactions in the tropics. *J. Atmos. Sci.*, **41**:604–613.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002. Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.*, **130**:1792–1811.

Rasmusson, E. M. and T. H. Carpenter, 1982. Variations in tropical sea surface temperature and surface wind fields associated with the Southern Oscillation / El Niño. *Mon. Wea. Rev.,* **110**:354–384.

Rebert, J., J. Donguy, G. Eldin, and K. Wyrtki, 1985. Relations between sea level, thermocline depth, heat content, and dynamic height in the tropical Pacific Ocean. *J. Geophys. Res.*, **90**:11719–11725.

Reid, D. J., 1968. Combining three estimates of gross domestic products. *Economica*, **35**:431–444.

Reid, D. J. *A comparative study of time series prediction techniques on economic data*. PhD thesis, University of Nottingham, 1969.

Reynolds, R. W., N. A. Rayner, D. C. Smith, T. M. Stockes, and W. Wang, 2002. An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**(13): 1609–1625.

Robert, C. P., 2001. The Bayesian choice. *Springer-Verlag*. 604 pp.

Robertson, A. W., L. Upmanu, S. E. Zebiak, and L. Goddard, 2004. Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.* In press.

Rocha, A. and I. Simmonds, 1997. Interannual variability of south-eastern African summer rainfall. Part I: Relationships with air-sea interaction processes. *Int. J. Climatol.*, **17**:235–265.

Ropelewski, C. F. and M. S. Halpert, 1986. North American precipitation and temperature associated with the El Niño Southern Oscillation (ENSO). *J. Climate*, **114**:2352–2362.

Ropelewski, C. F. and M. S. Halpert, 1987. Global and regional scale precipitation patterns associated with El Niño/Southern Oscillation. *Mon. Wea. Rev.*, **115**: 1606–1626.

Ropelewski, C. F. and M. S. Halpert, 1989. Precipitation patterns associated with high index phase of Southern Oscillation. *J. Climate*, **2**:268–284.

Sanders, F., 1963. On subjective probability forecasting. *J. Appl. Meteorol.*, **2**: 191–201.

Seber, G. A. F., 1977. Linear regression analysis. *John Wiley and Sons*. New-York. 465 pp.

Sivillo, J. K., A. J. E., and Z. Toth, 1997. An ensemble forecasting primer. *Weather and Forecasting*, **12**:809–818.

Smith, T. M. and R. E. Livezey, 1999. GCM systematic error correction and specification of the seasonal mean Pacific-North America region atmosphere from global SSTs. *J. Climate*, **12**:273–288.

Stael von Holstein, C.-A, S., 1971. An experiment in probabilistic weather forecasting. *J. Appl. Meteorol.*, **10**:635–645.

Stefanova, L. and T. N. Krishnamurti, 2002. Interpretation of seasonal climate forecast using Brier Skill Score, the Florida State University superensemble, and the AMIP-I dataset. *J. Climate*, **15**:537–544.

Stephenson, D. B. and F. J. Doblas-Reyes, 2000. Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus*, **52A**:300–322.

Stigler, S. M., 1973. Laplace, Fisher and the discovery of the concept of sufficiency. *Biometrika*, **60**:439–445.

Stockdale, T. N., 1997. Coupled ocean-atmosphere forecasts in the presence of climate drift. *Mon. Wea. Rev.*, **125**:809–818.

Stockdale, T. N., D. L. T. Anderson, J. O. S. Alves, and M. A. Balmaseda, 1998. Global seasonal rainfall forecasts using a coupled ocean-atmosphere model. *Nature*, **392**:370–373.

Stoeckenius, T., 1981. Interannual variations of tropical precipitation patterns. *Mon. Wea. Rev.*, **109**:1233–1247.

Straus, D. M. and J. Shukla, 2000. Distinguishing between the SST-forced variability and internal variability in mid-latitudes: Analysis of observations and GCM simulations. *Q. J. R. Meteorol. Soc.*, **126**:2323–2350.

Swets, J. A., 1988. Measuring the accuracy of diagnostic systems. *Science*, **240**: 1285–1293.

Taylor, J. W. and R. Buizza, 2003. Using weather ensemble predictions in electricity demand forecasting. *Int. Journal of Forecasting*, **19**:57–70.

Taylor, J. W. and D. W. Bunn, 1999. Investigating improvements in the accuracy of prediction intervals for combinations of forecasts: A simulation study. *Int. Journal of Forecasting*, **15**:325–339.

Thompson, P. D., 1977. How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.*, **105**:228–229.

Tracton, M. S. and E. Kalnay, 1993. Operational ensemble prediction at the National Meteorological Center - Practical aspects. *Weather Forecast*, **8**:379–398.

Trenberth, K. E., 1998. Development and forecasts of the 1997/98 El Niño: CLIVAR scientific issues. *International CLIVAR Project Office. Exchanges; Newsletter of the Climate Variability and Predictability Program (CLIVAR). Available from http://www.clivar.org*.

Trenberth, K. E., G. W. Branstator, D. Karoly, A. Kumar, N.-C. Lau, and C. Ropelewski, 1998. Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.*, **107**(C7):14291–14324.

Trenkler, G. and B. Gotu, 2000. Combination of forecasts: A Bibliography. *Available from http://www.sfb475.uni-dortmund.de/berichte/tr20-98.ps*.

Uvo, C. R. B., C. A. Repelli, S. E. Zebiak, and Y. Kushnir, 1998. On the relationships between tropical Pacific and Atlantic SST in northeast Brazil monthly precipitation. *J. Climate*, **13**:287–293.

van den Dool, H. M., J. Hoopingarner, E. O'Lenic, A. J. Wagner, A. G. Barnston, R. E. Livezey, D. Unger, A. Artusa, and R. Churchill, 1998. Third annual review of skill of CPC real time long lead predictions. *Proceedings of the* 23^{rd} *Annual Climate Diagnostics and Prediction Workshop. American Meteorological Society*, pages 9–12.

van Oldenborgh, G. J., M. A. Balmaseda, L. Ferranti, T. N. Stockdale, and D. L. T. Anderson, 2003. Did the ECMWF seasonal forecast model outperform a statistical model over the last 15 years? *European Centre for Medium-Range Weather Forecasts Technical Memorandum number 418*.

Velasco, I. and J. M. Fritsch, 1987. Mesoscale convective complexes in the america. *J. Geophys. Res.*, **92**(D8):9591–9613.

Vitart, F., M. Balmaseda, L. Ferranti, and D. Anderson, 2003. Westerly wind events and the 1997/98 El Niño event in the ECMWF seasonal forecasting system: A case study. *J. Climate*, **16**:3153–3170.

von Storch, H., 1999. On the use of inflation in statistical down-scaling. *J. Climate*, **12**:3505–3506.

von Storch, H. and F. W. Zwiers, 1999. Statistical Analysis in Climate Research. *Cambridge University Press*. ISBN 0521 450713. 484 pp.

Wallace, J. M. and D. S. Gutzler, 1981. Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.*, **109**:785–812.

Wilks, D. S., 1995. Statistical methods in the atmospheric sciences: An introduction. *Academic Press*. First edition. 467 pp.

Wilks, D. W., 2000. Diagnostic verification of the Climate Prediction Center long-lead outlooks, 1995-1998. *J. Climate*, **13**:2389–2403.

Williams, E. J., 1969. A note on regression methods in calibration. *Technometrics*, **11**:189–192.

Winkler, R. L., A. H. Murphy, and R. W. Katz, 1977. The consensus of subjective probability forecasts: Are two, three, ..., heads better than one? pages 57–62. Preprint volume. AMS fifth conference on probability and statistics in atmospheric sciences - Las Vegas, Nevada.

Wu, D-H., D. L. T. Anderson, and M. K. Davey, 1994. ENSO prediction experiments using a simple ocean-atmosphere model. *Tellus*, **46A**:465–480.

Wyrtki, K., 1975. El Niño – the dynamic response of the equatorial Pacific Ocean to atmospheric forcing. *J. Phys. Oceanography*, **5**:572–584.

Yang, Y., 2004. Combining forecasting procedures: Some theoretical results. *Econometric Theory*, **20**:176–222.

Zebiak, S.E. and M. A. Cane, 1987. A model of El Niño-Southern Oscillation. *Mon. Wea. Rev.*, **115**:2262–2278.

Ziehmann, C., 2000. Comparison of a single model EPS with a multi-model ensemble consisting of a few operational models. *Tellus*, **52A**:280–299.