University of Exeter
College of Engineering, Mathematics and Physical Sciences

Statistical methods for post-processing ensemble weather forecasts

Robin Mark Williams

Supervised by Dr Christopher A. T. Ferro & Dr Frank Kwasniok

Submitted by Robin Mark Williams, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Mathematics , February 2016.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(signature) .....

Abstract

Until recent times, weather forecasts were deterministic in nature. For example, a forecast might state "The temperature tomorrow will be 20°C." More recently, however, increasing interest has been paid to the uncertainty associated with such predictions. By quantifying the uncertainty of a forecast, for example with a probability distribution, users can make risk-based decisions. The uncertainty in weather forecasts is typically based upon 'ensemble forecasts'. Rather than issuing a single forecast from a numerical weather prediction (NWP) model, ensemble forecasts comprise multiple model runs that differ in either the model physics or initial conditions. Ideally, ensemble forecasts would provide a representative sample of the possible outcomes of the verifying observations. However, due to model biases and inadequate specification of initial conditions, ensemble forecasts are often biased and underdispersed. As a result, estimates of the most likely values of the verifying observations, and the associated forecast uncertainty, are often inaccurate. It is therefore necessary to correct, or post-process ensemble forecasts, using statistical models known as 'ensemble post-processing methods'. To this end, this thesis is concerned with the application of statistical methodology in the field of probabilistic weather forecasting, and in particular ensemble post-processing. Using various datasets, we extend existing work and propose the novel use of statistical methodology to tackle several aspects of ensemble post-processing.

Our novel contributions to the field are the following. In chapter 3 we present a comparison study for several post-processing methods, with a focus on probabilistic forecasts for extreme events. We find that the benefits of ensemble post-processing are larger for forecasts of extreme events, compared with forecasts of common events. We show that allowing flexible corrections to the biases in ensemble location is important for the forecasting of extreme events. In chapter 4 we tackle the complicated problem of post-processing ensemble forecasts without making distributional assumptions, to produce recalibrated ensemble forecasts without the intermediate step of specifying a probability forecast distribution. We propose a latent variable model, and make a novel application of measurement error models. We show in three case studies that our distribution-free method is competitive with a popular alternative that makes distributional assumptions. We suggest that our distribution-free method could serve as a useful baseline on which forecasters should seek to im-

prove. In chapter 5 we address the subject of parameter uncertainty in ensemble post-processing. As in all parametric statistical models, the parameter estimates are subject to uncertainty. We approximate the distribution of model parameters by bootstrap resampling, and demonstrate improvements in forecast skill by incorporating this additional source of uncertainty in to out-of-sample probability forecasts. In chapter 6 we use model diagnostic tools to determine how specific post-processing models may be improved. We subsequently introduce bias correction schemes that move beyond the standard linear schemes employed in the literature and in practice, particularly in the case of correcting ensemble underdispersion. Finally, we illustrate the complicated problem of assessing the skill of ensemble forecasts whose members are dependent, or correlated. We show that dependent ensemble members can result in surprising conclusions when employing standard measures of forecast skill.

Contents

List of Tables

List of Figures

1 Introduction

Weather conditions have had wide-ranging effects on humanity, seemingly since the beginning of time. Most obviously, the weather is a crucial factor in determining crop yields, and adverse periods of weather can lead to humanitarian crises. In modern society, weather conditions have both social and economic impacts. For example, in recent years the United Kingdom has witnessed several instances of wide-spread flooding. These events have impacted upon the livelihoods of those affected, as well as posing new challenges to the insurance industry. Other businesses and industries that are sensitive to weather conditions include supermarkets, construction, shipping, aviation and tourism. Indeed, in recent times the economic significance of the weather has led to the development of financial instruments that can be purchased by businesses to hedge against, and thus limit their financial exposure to, adverse weather conditions.

With the above comments in mind, it is clearly desirable to be able to provide accurate forecasts of future weather conditions. The possibility of mathematical approaches to weather forecasting was recognised as long ago as 1922, in the founding work of Lewis Richardson, 'Weather prediction by numerical process'. Since then, the advent of computers and supercomputers has given rise to weather forecasts over the entire planet, for many weather variables, for forecast lead times up to and beyond two weeks. Weather forecasts are primarily based on numerical models of the atmosphere, which are derived from the field of fluid dynamics.

Until the last decade, weather forecasts were almost always deterministic in nature. For example, a forecaster might state 'The temperature at 12pm tomorrow will be 20°C'. However, despite years of research and the computational power available to the forecasters of today, weather forecasts are still subject to errors, and so we can not treat such deterministic forecasts as exact predictions that can be wholly relied upon. The uncertainty in weather forecasting has provided an opportunity for statisticians to determine systematic errors in weather forecasts that can be corrected, as well as to quantify the uncertainty in such forecasts. In other words, recent developments have led to the field of probability weather forecasts, rather than deterministic weather forecasts. By quantifying the uncertainty in the deterministic forecasts that are typically issued, users of probability forecasts can also estimate

the likelihood of the occurrence of weather events. For example, local councils may wish to estimate the joint likelihood of temperatures falling below 0°C and heavy rainfall, which would lead to the formation of ice. Such estimates can then be used to make risk-based decisions — in the given example, councils may choose to deploy road gritting services if the probability forecast of ice formation exceeds a certain threshold.

The uncertainty in weather forecasts of the future atmospheric state is based upon so-called 'ensemble forecasts'. An ensemble forecast is a collection of deterministic forecasts that differ in either the numerical model used to issue the forecasts or the initial atmospheric conditions that are supplied to the model. As the constituent members of an ensemble forecast will typically differ in their forecast values, an ensemble forecast provides a means of estimating the probability of certain weather events. For example, if six of nine ensemble members forecast the temperature, T, to fall below 0°C, then we could assign a probability forecast $\Pr(T \leq 0) = 2/3$. Furthermore, the width of the ensemble forecast (i.e. the difference in the largest and smallest member), could be used as an 80% prediction interval — that is, an interval within which we would expect temperature observations to fall 80% of the time in the long-run.

Unfortunately, however, ensemble forecasts do not provide reliable representations of the forecaster's uncertainty in the future, unknown observations. This is due to persistent errors in the numerical models used for the ensemble member forecasts, and uncertainty in how to select initial conditions with which to initialise the numerical models. For example, operational ensemble forecasts are often underdispersed, and so issuing probability forecasts of an event by the proportion of ensemble members that predict its occurrence leads to inaccurate probability forecasts — in the above example, the event $\{T \leq 0\}$, which was assigned probability 2/3, will typically not occur on two thirds of occasions. Similarly, the prediction intervals of operational ensemble forecasts are often too narrow, and so the verifying observations lie outside the intervals more often than one would like. This has led to the development of the field of so-called 'ensemble post-processing', which is the subject of this thesis. Ensemble post-processing methods can be thought of in two classes. Firstly, ensemble forecasts may be corrected, for example by the removal of systematic errors in the ensemble location and spread. The output from the post-processing method is another ensemble forecast, which (we should hope) has more desirable properties than the initial, so-called 'raw' ensemble forecast. Alternatively, and more commonly used in practice, is to use information contained in ensemble forecasts to issue probability forecasts, usually in the form of probability distributions. For example, a popular post-processing method that we use frequently throughout this thesis is to model the verifying observation as a Gaussian-distributed random variable, where the ex-

pectation and variance of the Gaussian distribution are modelled as linear functions of the sample mean and sample variance of the ensemble forecast. Well-known properties of the Gaussian distribution can then be used to issue probability forecasts and prediction intervals, as well as deterministic forecasts. For example, the expectation of the Gaussian probability forecast distributions has frequently been shown to be a more accurate deterministic forecast than either the individual ensemble members or the ensemble mean, while the interval bounded by the $100 \times \alpha/2\%$ and $100 \times (1 - \alpha/2)\%$ quantiles, where α is a constant in the interval $(0, 1)$ provides a $100 \times (1 - \alpha)\%$ prediction interval. Ensemble post-processing methods are typically statistical models, and provide an opportunity for novel applications of statistical methods. In this thesis we present work that tackles several problems in the field of ensemble post-processing using statistical methodology.

The remainder of this thesis is organised as follows. In chapter 2 we give a broad exposition of the background material that we make use of in later chapters. The chapter begins with an overview of ensemble forecasting, including early development and current operational practice. We then give a detailed introduction and discussion of ensemble post-processing methods, many of which we make use of in our new work. The chapter concludes with an overview and discussion of the various methods that we use for the verification of both deterministic and probability forecasts, and a description of the datasets that we use in exemplifying our new methodology. In chapter 3 we present an investigation into ensemble post-processing methods for extreme events. We illustrate that allowing additional flexibility in the statistical models used for ensemble post-processing produces significant improvements to the skill of probability forecasts of the form $\Pr(y \leq q)$, where y is the verifying observation, and q is an *extreme* threshold of interest. This work has been published in the literature [Williams et al., 2014]. In chapter 4 we introduce a novel post-processing method that leads to recalibrated ensemble forecasts, rather than probability forecasts. Our new method makes fewer assumptions than are usually required, and serves as a more useful baseline than the simple frequency-based approaches described above. We recommend that new ensemble post-processing methods should seek to improve upon our baseline method. In chapter 5 we address the issue of uncertainty in the parameter estimates of the statistical models used for ensemble post-processing, a topic that was hitherto largely neglected in the literature. The results presented have also been published [Siegert et al., 2015a]. We show that probability forecasts issued by ensemble post-processing methods are more reliable when they account for parameter uncertainty, and we propose a method of doing so that is easy to implement. Finally, in chapter 6, we introduce two ideas that we think worthy of further research. We illustrate how diagnostic plots, which are widely used in statistical modelling but are infrequently discussed in the post-processing literature, can be used to improve the specification of the statistical models issued

by ensemble post-processing methods. We then discuss the effect of dependencies between ensemble members, expressed through their correlation, on the conclusions that we may draw from commonly employed verification measures of forecast skill. Chapter 6 concludes with a summary of our findings presented during the thesis, and suggests directions for future research.

2 Ensemble weather forecasting and ensemble post-processing

2.1 Introduction

In this chapter we provide details of much of the material that forms the basis for our novel work presented in chapters 3–6. The chapter is organised as follows. In section 2.2 we provide a brief overview of numerical weather prediction, and give an outline of the methodology that is used in operational ensemble weather forecasting. In section 2.3 we discuss the notion of calibration for both probability and ensemble forecasts, in particular, what is meant by 'reliability', 'resolution' and 'forecast uncertainty'. In section 2.4 we provide an extensive review of many ensemble post-processing methods, several of which we make use during this thesis. We also give details of the routines that are used for estimating the parameters in the statistical models that are specified by ensemble post-processing methods. In section 2.5 we provide details of the graphical and quantitative methods that we use for the verification of out-of-sample forecasts in our new work, for both probability and ensemble forecasts. Finally, in section 2.6 we describe the datasets that are used to illustrate our new contributions in chapters 3–6.

2.2 Numerical weather prediction and ensemble weather forecasting

2.2.1 Introduction

In this section we provide a brief overview of the main components of an ensemble forecasting system. We begin with a discussion of numerical weather prediction (NWP) models, and explain the main sources of error that result from the many difficulties in constructing accurate models of the atmospheric state. In section 2.2.3 we introduce the practice of ensemble forecasting, which provides a means of acknowl-

edging and accounting for the aforementioned errors in the deterministic forecasts that are issued by NWP model forecasts. We recommend the text Kalnay [2003] for a far more detailed and complete exposition of the material that is presented in these two subsections. We conclude this section with a discussion of two interpretations of ensemble forecasts that are commonly employed by forecasters.

2.2.2 Numerical weather prediction

The twentieth century saw the rapid development of machines that were able to automate mathematical operations, which grew in to the computers and super-computers of today. These technological advances gave rise to the possibility of completing tasks that required large numbers of calculations for the first time. Not least among these was the ability to use numerical models of the atmosphere to issue meteorological forecasts. Prior to such an automated approach, the first attempts at weather forecasting (Richardson 1922, see Richardson [2007]) involved the laborious process of making calculations by hand and interpolating observations gathered at weather stations to an appropriate grid. While much of the fundamental process established by Richardson (and others) remains to this day, powerful computation allows high resolution forecasts to be automatically generated for multiple meteorological variables, forecast lead times and for increasingly fine grids in three dimensional space.

Modern-day operational weather forecasts are based on numerical weather prediction (NWP) models, which are an approximation to the physics of the atmosphere. The models are derived from the field of fluid dynamics, and take the form of a high-dimensional system of coupled partial differential equations (PDEs). These PDEs represent the evolution in time and the spatial dependence of the many meteorological variables comprised within the model, including the complicated relationships governing inter-variable interactions. Initially based on a small set of equations that approximated the most important properties of the dynamics of the atmosphere, known as the 'full equations of motion', years of research has seen NWP models develop in to high-dimensional systems that are used operationally for forecast lead times of two weeks and beyond. The models provide approximations to large-scale, slowly-varying features, such as the Atlantic jet stream, as well as to small-scale, localised phenomena such as cloud formations and associated localised precipitation. Historically, a deterministic weather forecast was issued as the output of a single run of an NWP model. However, due to the many difficulties in numerical weather prediction, these forecasts are imperfect. In the remainder of this subsection we discuss three of the most difficult areas of numerical weather prediction, each of which lead to errors in the deterministic forecasts that are issued.

As is the case for all dynamical systems, NWP models are initial value problems, in the sense that their evolution through time is dependent on the initial value supplied. When making a single, deterministic weather forecast, therefore, it is important that the initial condition supplied to the model is the best available estimate of the atmospheric state at that time. This estimate is known as the analysis, and is produced using one of a variety of methods from the field of data assimilation. Errors in the analysis forecast will therefore lead to errors in the resulting deterministic forecast, even in the idealised setting of a 'perfect' NWP model. In his founding work in the 1960s [Lorenz, 1963, 1965], Lorenz studied the temporal evolution of models that were perceived as realistic approximations to the atmosphere at the time. In a series of numerical experiments, Lorenz found that two runs of an idealised model started from initial conditions that differ only very slightly will, after sufficient time, appear to evolve independently of one another, as they might had the two model runs been started from very different initial conditions. This problem became known as the 'butterfly effect', which considered the effect of small disturbances (or perturbations) in the initial conditions on the long-run evolution of dynamical models. In terms of the predictability of the atmosphere, Lorenz suggested that even a perfect NWP model would lose all predictive skill for forecast lead times longer than approximately two weeks, due to errors in the initial conditions which, even if known perfectly, are subject to round-off when used as inputs to computer models. Analogously, small differences in the analyses issued by weather centres could result in markedly different forecasts, even with the use of a common NWP model. There is yet to be a unified approach to producing analyses and, therefore, model initial conditions are likely to differ across operational centres. Furthermore, Lorenz found that the predictability of nonlinear dynamical systems with instabilities, such as the atmosphere, is dependent on the state of the system itself — the system is more stable in certain states than in others and, therefore, the skill of NWP forecasts is likely to vary, depending on the atmospheric stability at the time.

Secondly, constructing accurate representations of the atmospheric physics is a highly complicated task, and one that even after many years of research is still problematic. Large-scale features are generally well understood and are therefore predictable, while small-scale phenomena remain problematic. In addition, understanding the complicated interactions between the various features of the planet, for example the differences in the behaviour of the atmosphere over the oceans and over land, and then casting these interactions in a viable mathematical form is a complicated research problem. Furthermore, NWP models must be 'tuned' to the observed atmosphere by adjusting the many parameters that control their evolution as specified by the aforementioned coupled system of PDEs.

Finally, approximating the solution of the NWP model, that is, the system of PDEs that describe the atmospheric physics and structure, is a highly complicated task. Due to the size and complexity of the system of PDEs, closed-form solutions are not available. It is therefore necessary to discretise the model and use numerical methods to approximate the model solution. It is necessary to discretise the model over a fixed grid, the structure of which forms a research problem in its own right. Popular choices of grid coordinates include Cartesian, Spherical and Gaussian. The model resolution refers to the density of grid points, with finer grids corresponding to models with higher resolution. The temporal evolution is also discretised, such that the model state is estimated at discrete, rather than continuous time steps. With the model suitably discretised, numerical methods, known as finite difference schemes, can be applied to estimate the model state at each gridpoint and time step. The choice of numerical scheme has considerable implications, not only in terms of the accuracy of the model solution, but also computational cost. The development of improved numerical methods is also an important field of research in its own right, and further discussion of this topic is beyond the scope of this thesis.

As mentioned above, NWP models typically provide accurate representations of large-scale features, while forecasters are less confident in their ability to represent localised events. The difficulty in predicting small-scale features is largely due to the fact that they occur within areas smaller than the 'grid boxes' that are enclosed by the grids on which NWP models are based. Inaccuracies in initial conditions and imperfect model physics, as well as the difficulties in approximating the solution of the model itself, all contribute to the inaccuracies that are observed in operational weather forecasts.

2.2.3 From deterministic forecasts to ensemble forecasts

In the previous subsection we highlighted several sources of uncertainty in the forecasts issued by NWP models:

- The sensitivity of the NWP model to the initial conditions, and consequently the effect of analysis error on the model evolution.
- Uncertainty in the physical parameterisations and the parameter values used.
- The loss of accuracy in the model solution due to the discretisation of the model over space and time.
- The stability of the atmospheric system itself — Lorenz showed that the predictability of unstable dynamical systems can vary.

Considering these fundamental problems of weather prediction in combination, therefore, forecasters are rightly uncertain as to the accuracy of the deterministic forecasts that are issued by NWP models. It is therefore natural to turn to a framework that enables a forecaster to both issue a deterministic forecast and to provide an assessment of their confidence in that forecast or, or in other words, to quantify their forecast uncertainty. The framework for quantifying forecast uncertainty is therefore probabilistic in nature. A more detailed discussion of forecast uncertainty is provided in section 2.3. In this subsection we provide an overview of ensemble weather forecasting, which is the framework upon which forecast uncertainty is based.

As described in chapter 1, an ensemble forecast is a collection of deterministic forecasts that, in general, differ in their forecast values. We define an ensemble forecast for a general forecast occasion as $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$, where M denotes the number of ensemble members, or the ensemble size. We now provide a brief overview of the history of ensemble forecasting, and describe the methods that are commonly employed in their generation.

Epstein [1969] proposed the idea of a so-called 'stochastic-dynamic' framework for weather forecasting. Epstein's idea was to develop a partial differential equation that approximated the evolution of a probability density function (PDF) for the future, verifying observation, based on NWP model forecasts. The idea was to sample the possible initial conditions, and to approximate this PDF with the resulting model runs. The approach involved a very large number of model runs, which proved computationally infeasible. Even after some simplifying assumptions to estimate the first two moments rather than the entire PDF, Epstein's approach was not applicable in settings beyond those of low-dimensional, 'toy' models.

Hoffman and Kalnay [1983] experimented with so-called 'lagged average forecasting', which generates ensemble forecasts whose members are deterministic forecasts of the same verifying observation, but initialised at different times. While the method obviates the need to generate perturbations to the initial conditions, it is necessary to weight the ensemble members according to their age, to accommodate the idea that their forecast skill will decrease with increasing lead time. Obtaining the ensemble member weights requires estimating the temporal evolution of the covariance matrix of forecast errors, and difficulties in doing so have resulted in limited applications of the method. There are also limitations on the size of ensembles that can be generated from lagged average forecasting, as large ensemble forecasts would require the inclusion of forecasts at prohibitively long lead times.

Leith [1974] introduced the 'Monte Carlo' approach to ensemble forecasting. The idea is to generate ensemble forecasts by perturbing the analysis, and using these perturbed analyses as initial conditions for the NWP model forecasts. The pertur-

bations are sampled at random from a multivariate distribution that is based upon the dependence structure of historical forecast errors and scaled such that their amplitude is equal to the estimated analysis error. The dependence structure is derived in the data assimilation cycle, and must reflect the statistical horizontal and vertical structure of the forecast errors. Among other findings, Leith showed that the ensemble mean of Monte Carlo ensemble forecasts is in general a more skilful deterministic forecast than the 'control forecast', the NWP model forecast initialised at the analysis.

An alternative but related approach to Monte Carlo ensemble forecasting is to choose perturbations that are not sampled at random, but instead include information that is pertinent to the current predictability of the atmosphere. This approach recognises Lorenz's finding that the stability of the atmosphere, and therefore its predictability, is subject to variation. In this approach, the amplitudes of the perturbations depend on the estimated analysis errors at that time, which vary in keeping with the predictability of the atmosphere. Larger perturbations are chosen for more difficult forecasts, and smaller perturbations are used when atmospheric conditions are more stable. This is in contrast to the approach of Monte Carlo forecasting, for which initial conditions are chosen at random and so do not contain information about the current predictability of the atmosphere.

So-called 'bred vectors' are commonly used to generate perturbations that contain current information for the atmospheric stability. After first initialising ensemble members with randomly sampled initial conditions, as in the Monte Carlo approach, the evolution of the ensemble members is updated by adding regular (for example, every six hours) perturbations that depend on the forecast errors of the NWP model at that time. Bred vectors are commonly employed in operational centres [Kalnay, 2003, chapter 6]. It is reported (see section 6.6.2 of Kalnay [2003]) that experiments conducted at the National Centre for Environmental Prediction (NCEP) demonstrated that the second type of perturbation grew much faster than the perturbations that were chosen for Monte Carlo ensembles, resulting in ensemble forecasts that exhibited greater spread.

Two further approaches to producing ensemble forecasts are to use multiple data assimilation systems [Houtekamer et al., 1996] and to combine forecasts from multiple operational centres [Hou et al., 2001]. In the former system, random noise is added to the observations to reflect uncertainty in the analyses, and the NWP model parameterisations are also perturbed. The idea behind the system is that perturbing the NWP model will increase the extent to which the main contributory sources of uncertainty described in the previous subsection are sampled. In the second approach, the authors suggest that the forecast uncertainty will be well sampled by combining the analyses and state of the art NWP models from multiple weather

centres, whose NWP models and data assimilation processes are likely to differ.

2.2.4 Interpretations of ensemble forecasts

As mentioned in section 2.2.3, ensemble forecasts have been used successfully to improve the skill of deterministic forecasts and to estimate the associated forecast uncertainty. Other desirable applications include estimating the probability of binary events, such as the probability that the temperature will exceed a given threshold on a given day or, in a multivariate setting, that precipitation will exceed a certain threshold and the temperature will fall below 0°C, resulting in the likely formation of ice. As described in the previous chapter, a related application is the estimation of prediction intervals for the verifying observations.

With a variety of possible applications, therefore, there is evidently a need for clarity over how exactly to interpret the ensemble forecasts. For example, if a forecaster wishes to derive probability forecasts from ensemble forecasts produced using the lagged average forecasting scheme mentioned in section 2.2.3, it appears unnatural to assign equal importance, or weight, to the ensemble members, given we know that the forecasts differ in age and, consequently, that some members are likely to be more skilful than others. On the other hand, assigning equal weight to all members appears more acceptable in the Monte Carlo setting, where perturbations to the analysis at the forecast initialisation time are simulated randomly. Equally, the forecaster may need to think carefully when handling ensemble forecasts whose members differ due to perturbations in the NWP model and/or the initial conditions. We may expect those ensemble members issued with perturbed NWP model parameterisations to produce less skilful deterministic forecasts than the control forecast, given that the parameters of the NWP model are likely to be 'tuned' based on the control forecast.

From a practical perspective, it is often necessary for the forecaster to make a simplifying assumption when drawing inferences from ensemble forecasts or, in other words, to decide upon an interpretation of the ensemble members. One commonly employed interpretation is that the ensemble members represent an independent and identically distributed (IID) sample from an underlying probability distribution. For the remainder of this thesis we refer to this distribution as the *ensemble distribution*. Analogously, the verifying observation for an individual forecast occasion can be viewed as an IID draw from an underlying distribution, hereafter referred to as the *observation distribution*. As mentioned in the introductory chapter, a desirable property of an ensemble forecasting system is that the ensemble distribution, of which the ensemble forecast can be interpreted as representing a sample, and the observation distribution are equal.

The members of ensemble forecasts that we know, or choose to interpret as IID draws from an underlying distribution are said to be *exchangeable*. Furthermore, ensemble forecasts whose members are dependent, but can be viewed as a single draw from a symmetric multivariate distribution are also exchangeable. For example, an assumption of exchangeability is reasonable if a single NWP model is used and initial conditions are generated in a consistent manner, such as by random perturbations to the analysis. This leads to the following definition.

Definition 2.2.1 *The members of an ensemble forecast* $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$ *are* exchangeable *if the statistical properties of* $\boldsymbol{x}$ *are invariant to any relabelling of its constituent members. In other words, the members of an ensemble forecast are exchangeable if the forecaster is able to treat all ensemble members as statistically indistinguishable.*

A second interpretation of an ensemble forecast is that its empirical distribution function (EDF) represents a probability forecast distribution for the verifying observation, y. In this case the probability forecast distribution of y, conditional on the ensemble forecast $\boldsymbol{x}$, is

$$F(y \mid \boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{I}(x_m \leq y), \tag{2.1}$$

where $F(y)$ denotes the cumulative distribution function (CDF) for y, and $\mathbf{I}(\cdot)$ is the indicator function that takes the value 1 if its argument is true, and 0 if its argument is false. Equation (2.1) assumes equal weighting between ensemble members. Alternatively, the distribution for y could be constructed using weighted ensemble members such that

$$F(y \mid \boldsymbol{x}) = \sum_{m=1}^{M} w_m \mathbf{I}(x_m \leq y)$$

where

$$\sum_{m=1}^{M} w_m = 1,$$

and an appropriate statistical framework is needed to estimate the weights w_m, for $m = 1, 2, \ldots, M$. In section 2.5 we discuss how different interpretations of the ensemble members can affect how the skill of ensemble forecasts is assessed.

2.3 Forecast calibration and forecast uncertainty

2.3.1 Calibration for probability forecasts

As we describe in the next section (2.4), forecasters often use ensemble forecasts as a basis for issuing probability forecasts of the future, unknown atmospheric state. Depending on the nature of the predictand (the meteorological variable for which probability forecasts are issued), forecasters may issue continuous probability forecast distributions, discrete (either binary or categorical) distributions or, in a few special cases, a discrete-continuous mixture distribution. Probabilistic forecasters are particularly interested in two distinct properties of their forecasts, namely the forecast reliability, or calibration, and the forecast resolution. Forecast reliability refers to the ability of a probabilistic forecasting system to issue 'accurate' probability forecasts, in the sense that the value or event that materialises occurs with the relative frequency expected by the probabilistic forecasting system. Forecast resolution, meanwhile, is a measure of the information content of a forecast. The resolution can be viewed as the variability of the verifying observations, conditional on the probability forecasts. Forecasts with high resolution provide useful information to the forecast user, and so the conditional variability of the observations is large. On the other hand, forecasts with no resolution are unable to distinguish between the possible outcomes of the observations, and so there is no (conditional) variability in the observations. The notions of reliability and resolution are perhaps best exemplified with a discussion of probability forecasts of binary events, which now follows.

Let q denote a threshold of interest, and y the verifying observation, where y is unknown when the forecast is issued. Suppose the forecaster issues forecasts of the binary event $z = \mathbf{I}(y \leq q)$, which take the form $p = \Pr(y \leq q) = \Pr(z = 1)$. A probabilistic forecasting system for z is reliable if, among those occasions on which the event $z = 1$ is forecast to occur with probability p, the event does occur with relative frequency p, and this is true for all p. The forecast resolution of this system is a measure of its ability to distinguish between the outcomes $z = 0$ and $z = 1$. Observe that the constant forecast $p = \bar{z}$ is reliable — the long-run proportion of events that satisfy $z = 1$ is equal to the probability forecast p — but the forecast has no resolution — it does not tell the user anything that cannot be inferred from historical observations. In section 2.5.5 we provide expressions for two measures that are commonly employed to assess the reliability and resolution of probability forecasts of binary events.

A further property of probabilistic forecasts that is of interest is known as the 'sharpness', which provides a measure of the dispersion of the probability forecasts. The

sharpness of forecasts for binary predictands, z, is often measured as the variance of the Bernoulli distribution with probability $p = \Pr(z = 1)$, given by $p(1 - p)$. Similarly, the sharpness of forecasts for continuous predictands is often measured as the variance of the continuous forecast distribution. Forecasters prefer probability forecasts that are both reliable and sharp. In this thesis we refer to the guiding principle of Gneiting et al. [2007] and other papers by the same authors, who state that forecasts should be as sharp as possible, subject to reliability. Specifically, in a discussion of desirable properties of probability forecast distributions for continuous predictands, Gneiting et al. state: "The more concentrated the forecast PDF, the sharper the forecast, and the sharper the better, subject to calibration."

In this thesis we also make use of 'prediction intervals', which provide an insight in to both the sharpness and calibration of probability forecasts for continuous, rather than binary, observations. For a given probability forecast distribution, we define the $\alpha\%$ prediction interval as the interval within which the verifying observation lies with probability α. For example, for a Gaussian forecast distribution a 90% prediction interval for the verifying observation could be calculated as the interval $(q_{.05}, q_{.95})$, where the notation q_β refers to the β-quantile of the forecast distribution. The 'coverage' of prediction intervals refers to the actual relative frequency of observations that fall in such intervals. Prediction intervals for which the expected and observed coverage are not equal are indicative of probability forecast distributions that are not reliable. In keeping with the foregoing remarks concerning forecast reliability and sharpness, we would like prediction intervals to be as narrow as possible, subject to having accurate coverage.

We also make frequent use of the term 'forecast uncertainty' in the remainder of this thesis, the precise meaning of which depends on the context of the probabilistic forecast. For example, forecast uncertainty for a continuous predictand usually refers to the spread, or dispersion, of the probability forecast distribution, and can be viewed as a measure of uncertainty in deterministic forecasts that might also be inferred from the forecast distribution, such as its expectation or median. Forecasters are 'more uncertain' when issuing forecasts from probability distributions whose dispersion is large, compared with distributions whose dispersion is small. There are natural relationships between such forecasts and prediction intervals — prediction intervals are narrower for forecasts in which we are confident, or less uncertain.

2.3.2 Calibration for ensemble forecasts

Forecasters sometimes prefer to issue ensemble forecasts, rather than probability forecasts, and so it is also important to establish the notion of calibration for ensembles. As described in section 2.2.4, ensemble forecasts can be interpreted as

either a probability forecast for the verifying observation (via their EDF), or as a collection of IID realisations from an underlying 'ensemble distribution'. In the former case the comments given in the previous subsection apply. On the other hand, we view ensemble forecasts whose members are IID samples as being calibrated with the verifying observations if the observations also appear as IID realisations of the corresponding ensemble distributions.

Calibrated ensemble forecasts $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$ for continuous predictands, whose members are IID random draws, have the following appealing properties. The expectation of the ensemble mean, $\bar{x}$, is equal to the expectation of the observation, y, that is $E(\bar{x}) = E(y)$. Secondly, the range of the ensemble forecast, $\max(\boldsymbol{x}) - \min(\boldsymbol{x})$, forms a $100 \times (M-1)/(M+1)\%$ prediction interval with the correct long-run observed coverage.

Historically, and (we understand) still in some operational settings, the empirical distribution functions (EDFs) are used as probability forecast distributions for the verifying observations. For example, the EDF of an ensemble forecast might be used to issue a probability forecast for a binary predictand by calculating the frequency of ensemble members that predict the event to occur. As we discuss in section 2.4.2.1, however, probability forecasts derived from such frequency-based approaches have undesirable properties, even if the ensemble forecast is calibrated with the verifying observation in the sense described above.

2.4 An overview of ensemble post-processing methods

2.4.1 Introduction

While considerable effort has been devoted to the production of NWP models that accurately describe the physics of the atmosphere, as well as to the development of perturbations to the analysis that accurately represent the forecaster's uncertainty of the atmospheric state at the model initialisation time, it remains the case that the evolution of the atmosphere is insufficiently resolved, and that the growth of the perturbations does not accurately reflect the state-dependent predictability of the atmosphere. Often the growth rates of the perturbations are slower than the growth rates resulting from the instabilities of the true atmospheric flow, and therefore many operational ensemble forecasts are underdispersed [Hamill and Colucci, 1997, 1998].

As mentioned in chapter 1, therefore, forecasters typically can not rely on ensemble

forecasts as a basis for stating their beliefs about their uncertainty in the future verifying observations. Persistent errors, or biases, in the location (the forecast values of the deterministic ensemble members) and the spread of ensemble forecasts mean that, for example, the proportion of ensemble members that forecast the occurrence of an event, such as the binary event $\{y \leq q\}$, where y is an observation and q a threshold of interest, is not a reliable estimate of the probability of the event occurring. Typically the event will occur with relative frequency that is not equal to the proportion of ensemble members that predict it to do so or, in other words, ensemble forecasts are typically not well calibrated with the verifying observations. Similarly, prediction intervals, usually defined as the range of the ensemble forecasts (see section 2.3.2), are frequently too narrow — observations typically fall outside of the range of the ensemble forecasts more often than indicated by the nominal coverage of the prediction intervals.

Despite their deficiencies, however, ensemble forecasts often contain useful information that can be exploited to issue recalibrated forecasts of the future atmospheric state, either as probability forecasts or ensemble forecasts. For example, from as early as Leith [1974], the ensemble mean has frequently proven to be a more skilful deterministic forecast than the control forecast. Furthermore, the ensemble spread, which is usually measured by its sample variance, is often a useful predictor of the error in the ensemble mean forecast, despite the underdispersion typically observed; see, amongst many others Hamill and Colucci [1997, 1998]; Raftery et al. [2005]; Gneiting et al. [2005]. In other words, referring to our discussion in section 2.3.2, the ensemble variance is often a useful predictor of the forecast uncertainty. Ensemble forecasts with this property are said to exhibit 'spread-skill relationships'. Ensemble forecasts with large spread are often associated with ensemble means whose deterministic forecasting errors are larger than ensemble forecasts with small spread. Such spread-skill relationships are exploited frequently throughout this thesis.

As mentioned in chapter 1, ensemble post-processing methods often take the form of parametric statistical models that seek to quantify relationships (such as the bias) between the ensemble forecasts and observations, and specify probability forecast distributions for the future (unknown) verifying observations. However, an ensemble post-processing method could be as simple as, for example, removing a constant bias from each member. In this thesis we class any method that exploits relationships between the ensemble forecasts and observations to produce recalibrated forecasts as a post-processing method, whether in the form of probability forecast distributions or recalibrated ensemble forecasts. For example, in chapters 3 and 5, we investigate post-processing methods that construct continuous probability forecast distributions in the standard parametric statistical modelling framework, while in chapter 4 we introduce a new post-processing method that recalibrates ensemble forecasts only.

As is generally the case in statistical modelling, the majority of ensemble post-processing methods require the estimation of model parameters. These estimates are obtained from samples of historical ensemble forecasts and their verifying observations, which we refer to as 'training samples' throughout this thesis. The parameter estimation procedure is usually performed by optimising an objective function that is calculated over the training sample, as in chapters 3, 5 and 6, using numerical algorithms to find the optimal set of parameter estimates. Alternatively, parameter estimates are sometimes calculated directly using techniques such as the method of moments, as in chapter 4. So-called 'rolling' training samples are often employed to estimate model parameters for the next forecast occasion — that is, a training sample of the N previous ensemble forecasts and verifying observations is used to estimate the parameters for the next 'out-of-sample' forecast.

Having obtained parameter estimates from a training sample of ensemble forecasts and observations, the chosen ensemble post-processing method can be used to issue either probability forecasts or post-processed ensemble forecasts (depending on the post-processing method) of the future, unknown atmospheric state or, in other words, to issue out-of-sample forecasts. We make clear this distinction by denoting out-of-sample forecasts and observations with the subscript t, and within-sample forecasts and observations (that are used for parameter estimation) with the subscript i. The size of training samples is denoted by N, and the size of the dataset of out-of-sample forecasts is denoted by T. If a rolling training sample is used for parameter estimation, the parameter estimates will change with forecast occcasion t, where $t = 1, 2, \ldots, T$ indexes the out-of-sample forecasts in the test dataset. In certain studies, however, such as that presented in chapter 3, the same parameter estimates are used for each out-of-sample forecast. We make clear the distinction in our notation for the particular study at hand.

The remainder of this section is organised as follows. In section 2.4.2 we describe some simple methods for issuing probability forecasts for the binary event $\{y_t \leq q\}$, where q denotes a threshold of interest, that are related to the frequency-based approaches mentioned previously. In sections 2.4.3 and 2.4.4 we review several post-processing methods that were used in the publication Williams et al. [2014]. These post-processing methods, or adaptations thereof, have been successfully applied to a variety of problems in the forecasting of various meteorological variables, and are frequently used in chapters 3–6. We give an overview of the methods only, and defer more technical material such as the objective functions used for parameter estimation until they are required in chapter 3. In section 2.4.5 we outline some post-processing methods that tackle meteorological variables of renowned difficulty, such as wind direction and precipitation, and describe an extension to the logistic regression model described in section 2.4.4.3. We also outline some post-processing

methods that yield forecasts of multivariate quantities. We have not made use of many of the methods in section 2.4.5, either because we were not aware of them, they were not published until the later stages of the study, or they were not relevant due to being targeted at specific variables that were not under consideration. In section 2.4.6 we describe the approach of ensemble copula coupling (ECC), which is an increasingly popular method for producing post-processed ensemble forecasts of multivariate predictands, such as for spatial fields. Finally, in section 2.4.7 we give details of the procedures used in our work for the estimation of model parameters. We describe two popular objective functions and the numerical algorithms that are used to find the optimal parameter estimates.

2.4.2 Ad-hoc post-processing methods

2.4.2.1 Frequency-based probability forecasts

In this section we describe some simple frequency-based approaches for estimating probability forecasts of binary events. The climatology serves as the simplest probability forecast of the form $p = \Pr(y_t \leq q)$, where q is a threshold of interest to the user. The probability p is estimated from the empirical distribution function of the observed climatology as

$$\Pr(y_t \leq q) = F_{clim}(q) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}(y_i \leq q), \tag{2.2}$$

where $y_i, i = 1, 2, \ldots, N$ denote N historical observations. The climatology forecast is reliable in the sense described in section 2.3.1, but has no resolution — the same forecast is always issued, and so does not provide any useful information to the user beyond that that can be inferred from the historical observations.

Until the growth in popularity of more sophisticated methods, probability forecasts of the form $\Pr(y_t \leq q)$ were derived from ensembles using simple frequency-based calculations. These approaches are based on the assumption that the ensemble members are sampled from the 'true' probability density function (PDF) of the future verifying observation y_t. In this case, the proportion of ensemble members predicting the event $\{y_t \leq q\}$ is a consistent and unbiased estimator of the probability of the event occurring. In the simplest case, the proportion of members of the ensemble forecast $\boldsymbol{x}_t = (x_{1,t}, x_{2,t}, \ldots, x_{M,t})$ predicting the event $\{y_t \leq q\}$ can be used to estimate $\Pr(y_t \leq q)$

$$\Pr(y_t \leq q) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{I}(x_{m,t} \leq q). \tag{2.3}$$

This probability forecast suffers from the implications that probability forecasts of 0 (1) are issued when the threshold q is smaller (larger) than all ensemble members. An alternative estimator is

$$\Pr(y_t \leq q) = \text{Rank}(q)_t/(M+1), \tag{2.4}$$

where $\text{Rank}(q)_t = \sum_{m=1}^{M} \mathbf{I}(x_{m,t} \leq q) + 1$ is the rank of the threshold q when pooled together with the members of the ensemble forecast $\boldsymbol{x}_t$. Again, however, this estimator implies a probability of 1 when all ensemble members are smaller than q, i.e. when $\text{Rank}(q)_t = M + 1$.

Direct model output (DMO) is a further frequency-based estimator, that avoids probability forecasts of 0 or 1. Probability forecasts are given by

$$\Pr(y_t \leq q) = \frac{\text{Rank}(q)_t - 1/3}{M + 1 + 1/3}. \tag{2.5}$$

Unlike the other frequency-based approaches described previously (equations (2.3) and (2.4)), the probability forecasts returned by equation (2.5) do not attain either the undesirable values of 0 or 1. The probability forecasts can range from $2/(3M+4)$ (when $\text{Rank}(q)_t = 1$) to $(3M+2)/(3M+4)$ (when $\text{Rank}(q)_t = M+1$). The adjustments $-1/3$ to the numerator and $1/3$ to the denominator of equation (2.5) are one of several possible corrections to frequency-based approaches, such as those given in equations (2.3) and (2.4). Wilks [2006b, page 41] provides several other possibilities. We have chosen to show the DMO forecasts as they were used in the article by Wilks [2006a] and in our own work presented in chapter 3.

Note that the four methods of issuing probability forecasts $\Pr(y_t \leq q)$ described in this section should not be viewed as ensemble post-processing methods, since the forecasts are simply a function of the members of the ensemble forecast $\boldsymbol{x}_t$ — the ensemble is not post-processed, and the probability forecasts are independent of any historical ensemble forecasts.

2.4.2.2 Rank histogram recalibration

As highlighted in section 2.4.1, in practice ensemble forecasts suffer from biases in both their location and dispersion. This was discussed by Hamill and Colucci [1997] in an application to probability forecasts of precipitation. Hamill and Colucci [1997, 1998] proposed an ensemble post-processing method that we refer to as rank histogram recalibration (RHR), that attempts to account for the biases in ensemble location and dispersion. Unlike the frequency-based approaches described above, the RHR method makes use of historical training samples of ensemble forecasts

and observations as follows. Firstly, constant biases are removed from the ensemble forecasts $\boldsymbol{x}_i, i = 1, 2, \ldots, N$, to form bias-corrected ensembles $\widehat{\boldsymbol{x}}_i$, where

$$\widehat{x}_{im} = x_{im} + \frac{1}{M \times N} \sum_{i=1}^{N} \sum_{m=1}^{M} (y_i - x_{im}), \text{ for } i = 1, 2, \ldots, N \text{ and } m = 1, 2, \ldots, M, \tag{2.6}$$

so that the unconditional (over the entire training sample) sample mean of the bias-corrected ensemble members is equal to the sample mean of the observations $y_i, i = 1, 2, \ldots, N$. Then define a vector of weights,

$$w_j = \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}\left(\text{Rank}(y_i) = j\right) \text{ for } j = 1, 2, \ldots, M+1 \tag{2.7}$$

where $\text{Rank}(y_i) = 1 + \sum_{m=1}^{M} \mathbf{I}(\hat{x}_{im} \leq y_i)$ is the rank of the observation when pooled together with the members of the (de-biased) ensemble forecast $\hat{\boldsymbol{x}}_i$. The weights are therefore given by the relative frequency of the $M + 1$ possible ranks that can be taken by the N observations y_i over the training sample. This enables out-of-sample probability forecasts to be issued, as we now describe.

Firstly, any constant bias is removed from the out-of-sample ensemble forecast $\boldsymbol{x}_t$, using the same bias correction as was performed for the ensemble forecasts in the training sample. We then define the order statistics for the corrected ensemble forecast $\hat{\boldsymbol{x}}_t$, denoted $\hat{x}_t^{(m)}$ for $m = 1, 2, \ldots, M$, such that $\hat{x}_t^{(1)} < \hat{x}_t^{(2)} < \ldots < \hat{x}_t^{(M)}$. The probability forecast distribution issued by the rank histogram recalibration method is then constructed as follows. The distribution is assumed uniform between consecutive order statistics of the ensemble forecast, with each uniform distribution weighted by the relevant weight defined above. For example, the distribution between ensemble members $\hat{x}_t^{(1)}$ and $\hat{x}_t^{(2)}$ is assumed uniform, and is weighted by w_2, and the uniform distribution between ensemble members $\hat{x}_t^{(M-1)}$ and $\hat{x}_t^{(M)}$ is weighted by w_M. The probability distribution for values that are unbounded by the ensemble forecast must be specified by the forecaster. In an application to probability forecasts of quantitative precipitation, Hamill and Colucci [1997] assumed that observations y_t were uniformly distributed between 0 and the smallest ensemble member, $\hat{x}_t^{(1)}$, but fitted a Gamma distribution to the right hand tail, that is used for probability forecasts of values that lie above the largest ensemble member, $\hat{x}_t^{(M)}$. On the other hand, Wilks [2006a] fitted a Gaussian distribution to both tails of the RHR forecast distribution, with expectation and variance given by the ensemble mean and variance, respectively. Probability forecasts for observations in the lower and upper tails of the forecast distribution are weighted by w_1 and w_{M+1}, respectively. As described earlier, many operational ensemble forecasts are underdispersed, and so a relatively large proportion of observations fall in the tails of the probability forecast distributions. We denote by $g^{\text{RHR}}(\cdot)$ the parametric family of probability distribu-

tions that is chosen for the tails of the RHR forecast distribution, with cumulative distribution function $G^{\mathrm{RHR}}(\cdot)$. As with Wilks [2006a], this probability distribution is typically dependent on the ensemble forecast $\hat{\boldsymbol{x}}_t$. The RHR forecast distribution is therefore a weighted, disjoint mixture of uniform distributions, and the distributions that are chosen by the forecaster for quantities in the lower and upper tails that are unbounded by the ensemble forecast.

With the above comments in mind, probability forecasts of the binary event $\{y_t \leq q\}$, for a threshold of interest q are given by

$$\Pr(y_t \leq q) = \begin{cases} \sum_{j=1}^{k} w_j + w_{k+1} \frac{q - \hat{x}_t^{(k)}}{\hat{x}_t^{(k+1)} - \hat{x}_t^{(k)}} & \text{if } \hat{x}_t^{(1)} < q \leq \hat{x}_t^{(M)}, \\ w_1 \frac{G^{\mathrm{RHR}}(q)}{G^{\mathrm{RHR}}(\hat{x}_t^{(1)})} & \text{if } q \leq \hat{x}_t^{(1)}, \\ \sum_{j=1}^{M} w_j + w_{M+1} \frac{G^{\mathrm{RHR}}(q) - G^{\mathrm{RHR}}(\hat{x}_t^{(M)})}{1 - G^{\mathrm{RHR}}(\hat{x}_t^{(M)})} & \text{if } q > \hat{x}_t^{(M)}. \end{cases} \tag{2.8}$$

2.4.3 Ensemble dressing methods

We now introduce a class of post-processing methods that 'dress' the members of ensemble forecasts, in an attempt to correct the biases in dispersion that are often observed. The dressing can be in the form of adding to each member either additional ensemble members [Roulston and Smith, 2003] or continuous probability distributions, referred to as dressing kernels [Wang and Bishop, 2005; Raftery et al., 2005].

2.4.3.1 Best member dressing

Roulston and Smith [2003] proposed dressing each member of an ensemble forecast with an additional 'daughter ensemble', to form a so-called 'hybrid ensemble'. The post-processing method was based on the idea that residual uncertainty remains in operational ensemble forecasts, and that 'dressing' the ensemble members with additional samples will aid in reflecting uncertainty that is not present in the raw ensemble forecast. The dressing procedure involves sampling from a historical archive of forecast errors of the 'best member', defined as the ensemble member with the smallest error in forecasting the verifying observation, in a D-dimensional space. For example, the multi-dimensional space could refer to spatially gridded forecasts of a single variable, multiple variables at a single location, or a combination of these. Additional ensemble members are added to the raw ensemble forecast by repeatedly sampling from the historical archive of best member errors. The method therefore allows the construction of ensemble forecasts with additional members, the size of which can be predetermined. The method therefore provides a means of comparing

ensemble forecasting systems of different size, as hybrid ensemble forecasts of the same size can be constructed for each of the competing systems. There is, however, a finite limit on the size of the hybrid ensemble forecasts that can be attained by resampling, due to the finite size of the archive of best member errors.

Wang and Bishop [2005] developed the so-called best member dressing (BMD) post-processing method, which builds on the ideas of Roulston and Smith [2003]. Rather than adding additional samples to the ensemble forecasts, however, the BMD method dresses each member with a Gaussian distribution, or kernel, centred at the ensemble member. This post-processing method therefore results in the construction of continuous probability forecast distributions, rather than ensemble forecasts as in Roulston and Smith [2003]. Wang and Bishop derived an analytic expression for the variance of the Gaussian-distributed dressing kernels, denoted σ^{BMD^2}, which mathematically constrains the variance of the BMD forecast distribution to be equal to that of the expected squared distance between the observations and ensemble members in the training sample. In other words, if ensemble members were randomly sampled from the BMD forecast distributions, the average squared distances between the ensemble members and the observations should be equal to the average squared distances between the ensemble members, if the observations arise from the BMD probability forecast distribution. The dressing kernel variance, σ^{BMD^2}, is calculated from a training sample. The ensemble members are first corrected for seasonal biases, by subtracting the seasonally averaged error of the ensemble mean at each location and forecast lead time from all members of the corresponding raw ensemble forecast, in a similar vein to equation (2.6). For the 1-dimensional (univariate) case the dressing kernel variance is then given by

$$\sigma^{\text{BMD}^2} = \frac{1}{N}\sum_{i=1}^{N}\left\{(\overline{\hat{x}}_i - y_i)^2 - (1 + 1/M)s_i^2\right\}, \tag{2.9}$$

where $\overline{\hat{x}}_i = M^{-1}\sum_{m=1}^{M}\hat{x}_{im}$ and $s_i^2 = (M-1)^{-1}\sum_{m=1}^{M}(\hat{x}_{im} - \overline{\hat{x}}_i)^2$ are the ensemble mean and variance of the ith (bias corrected) ensemble forecast. In the 1-dimensional case, the dressing kernel variance is equal to the difference of the average variance of the forecast errors of the ensemble mean and a slightly inflated average ensemble variance.

The out-of-sample probability forecast distribution for the observation y_t conditional on bias-corrected ensemble member $\hat{x}_{m,t}$ is then $y_t \mid \hat{x}_{m,t} \sim N(\hat{x}_{m,t}, \sigma^{\text{BMD}^2})$, and the forecast PDF for observation y_t is

$$f^{BMD}(y_t \mid \hat{\boldsymbol{x}}_t) = \frac{1}{M \times \sigma^{\text{BMD}}}\sum_{m=1}^{M}\phi\left(\frac{y_t - \hat{x}_{m,t}}{\sigma^{\text{BMD}}}\right), \tag{2.10}$$

where $\phi(\cdot)$ denotes the PDF of the standard Gaussian distribution. Probability forecasts of the binary event $\{y_t \leq q\}$ are given by

$$\Pr(y_t \leq q) = \frac{1}{M} \sum_{m=1}^{M} \Phi\left(\frac{q - \hat{x}_{m,t}}{\sigma^{\mathrm{BMD}}}\right), \tag{2.11}$$

where $\Phi(\cdot)$ denotes the CDF of the standard Gaussian distribution.

The method of Wang and Bishop [2005] can also be used to produce multivariate dressing kernels for forecasts of multivariate predictands, such as temperature forecasts over spatial fields. However, the method can only correct the ensemble dispersion for variables in which the ensemble forecasts are underdispersed — as the method adds uncertainty to each ensemble member, it cannot be used to correct the dispersion of ensemble forecasts that are overdispersed. In this case, it is necessary to adjust the dispersion of the ensemble members through a rescaling of the ensemble forecasts. A scheme that facilitates such a correction is introduced in chapter 3.

Wang and Bishop found the skill of the approach of Roulston and Smith [2003] to be dependent on the multi-dimensional space used to identify the 'best ensemble member', and that the dispersion of the hybrid ensembles remained an inaccurate measure of the forecast uncertainty. By contrast, the authors found that their BMD method yielded forecasts that were reliable under the second moment, that is, the forecast variance was well calibrated with the squared errors of the ensemble mean forecast. However, the authors noted that their method was unlikely to be adequate for variables such as precipitation and windspeed, for which non-Gaussian dressing kernels are likely to be preferable.

2.4.3.2 Bayesian model averaging

Raftery et al. [2005] made the first application of Bayesian Model Averaging (BMA), a popular method in the economics literature, to probability forecasts of meteorological variables. Each of the M (possibly bias corrected) ensemble members is dressed with a kernel that reflects the uncertainty about the forecasts of that member, in an analogous manner to the BMD method described in section 2.4.3.1. The key difference is that the BMA dressing kernel variance is estimated as a model parameter, rather than being calculated explicitly from the training sample with equation (2.9). The dressing kernel variance is estimated by optimisation of an objective function that is calculated over the training sample, for which we give details in chapter 3. Furthermore, in the full specification [Raftery et al., 2005], each ensemble member is weighted and has its own dressing kernel variance, with the weights determined

by the forecast skill of the individual members over the training sample. As with the rank histogram recalibration method the weights are constrained to sum to 1.

We denote by g_m^{BMA} and G_m^{BMA} the PDF and CDF of the dressing kernel associated with the mth ensemble member, respectively. In the full model specification [Raftery et al., 2005] probability forecasts of the binary event $\{y_t \leq q\}$ are given by

$$\Pr(y_t \leq q) = \sum_{m=1}^{M} w_m G_m^{\text{BMA}}(q \mid \hat{x}_{m,t}). \tag{2.12}$$

If individual, or groups of ensemble members can be distinguished, for example as in multi-model ensembles, then it seems appropriate to use member-dependent (or group-dependent) kernels and weights as in equation (2.12). On the other hand, if the ensemble members are exchangeable then it is appropriate to use a simplified forecast distribution in which all ensemble members are assigned equal weight, and where each dressing kernel has the same variance. In this case probability forecasts are given by

$$\Pr(y_t \leq q) = \frac{1}{M} \sum_{m=1}^{M} G_m^{\text{BMA}}(q \mid \hat{x}_{m,t}). \tag{2.13}$$

The choice of distribution for the dressing kernels, $G_m^{\text{BMA}}(\cdot)$, is dependent on the meteorological variable to be forecast. Raftery et al. [2005] successfully applied Gaussian-distributed kernels centred on the ensemble members,

$$y_t \mid \hat{x}_{m,t} \sim N(\hat{x}_{m,t}, \sigma_m^{\text{BMA}^2}), \text{ for } m = 1, 2, \ldots, M, \tag{2.14}$$

with separate dressing kernel variances $\sigma_m^{\text{BMA}^2}$ estimated for each ensemble member. As noted above, a simplifying assumption appropriate for exchangeable ensembles is $\sigma_m^{\text{BMA}^2} = \sigma^{\text{BMA}^2}$ for all $m = 1, 2, \ldots, M$. This model is also more parsimoneous, and is subject to less sampling error than the full specification. In this case, probability forecasts of the binary event $\{y_t \leq q\}$ take the same form as the corresponding BMD forecasts,

$$\Pr(y_t \leq q) = \frac{1}{M} \sum_{m=1}^{M} \Phi\left(\frac{q - \hat{x}_{m,t}}{\sigma^{\text{BMA}}}\right). \tag{2.15}$$

In section 2.4.5 we outline some alternative formulations of the BMA method that have been developed for meteorological variables whose distribution cannot be reasonably approximated by a mixture of Gaussian distributions. Such flexibility has seen the BMA method become perhaps the most popular method in the post-processing literature.

2.4.4 Regression methods

We now describe a second class of ensemble post-processing methods, that can be broadly classified as regression models. Unlike dressing methods, which add uncertainty to the raw ensemble forecast while preserving the location of the individual ensemble members, these regression methods construct forecast distributions using summary statistics of the ensemble forecasts as model covariates. In a sense regression methods therefore use less information about the raw ensemble forecast than dressing methods, as in general the relative location of the ensemble members is less influential on the probability forecast distributions for the verifying observations.

2.4.4.1 Model output statistics

Model output statistics (MOS) is a simple post-processing method that is more commonly used for post-processing ensemble forecasts on longer time-scales, such as for seasonal and climate forecasts [Tippett et al., 2005; Glahn et al., 2009], although less so in the post-processing of short-range weather forecasts as considered in this thesis. The method is simply a linear regression model of the form

$$y = a + b\bar{x} + \epsilon, \tag{2.16}$$

where ϵ is a Gaussian-distributed random variable with expectation 0 and constant variance c^2. As for all regression models, the model parameters (here a, b and c) are estimated from training samples. Further details of the parameter estimates are discussed later in section 5.2.1. Out-of-sample probability forecasts for the binary event $\{y_t \leq q\}$ are given by

$$\Pr(y_t \leq q) = \Phi\left(\frac{y_t - \mu_t}{c}\right), \tag{2.17}$$

where again $\Phi(\cdot)$ denotes the CDF of the Gaussian distribution, and $\mu_t = a + b\bar{x}_t$ is the expectation of the MOS probability forecast distribution for the verifying observation y_t.

2.4.4.2 Nonhomogeneous Gaussian Regression

Gneiting et al. [2005] introduced the post-processing method known as Nonhomogeneous Gaussian regression (NGR), which is appropriate when forecasting variables whose distribution, conditional on the ensemble forecasts, can be reasonably modelled by a Gaussian distribution. The NGR method extends the MOS forecasts

discussed previously to account for the possible existence of spread-skill relationships between the spread of the ensemble forecasts and the magnitude of the errors of the NGR forecast mean.

For the out-of-sample forecast at time t, the expectation and variance of the Gaussian probability forecast distribution issued by the NGR post-processing method are given by linear functions of the ensemble mean and variance $\bar{x}_t$ and s_t^2, respectively. In the full model specification proposed by Gneiting et al. [2005] the expectation, μ_t, is a weighted sum of the members of $\boldsymbol{x}_t$,

$$\mu_t^{\mathrm{NGR}} = a + \sum_{m=1}^{M} w_m x_{m,t}, \tag{2.18}$$

where the weights $w_1, w_2, \ldots, w_M$ reflect the deterministic forecast skill of the ensemble members in the training sample used for parameter estimation, and the parameter a is a constant offset. It is appropriate to assign equal weight to each ensemble member in the case of exchangeable ensembles, in which case the NGR forecast mean reduces to

$$\mu_t^{\mathrm{NGR}} = a + b\bar{x}_t \tag{2.19}$$

where the parameters a and b are estimated from the training sample. As is also true for the MOS post-processing method described previously, the expectation of the NGR forecast distribution, μ_t^{NGR}, is therefore a bias-corrected deterministic forecast, where the bias correction is assumed to be a linear function of the ensemble mean, rather than simply an additive constant as for the RHR, BMD and BMA post-processing methods (see equation (2.6)).

Similarly, at time t the NGR forecast variance is given by

$$\sigma_t^{\mathrm{NGR}^2} = c + ds_t^2, \tag{2.20}$$

where $s_t^2 = (M-1)^{-1}\sum_{m=1}^{M}(x_{m,t} - \bar{x}_t)^2$ is the sample variance of the ensemble forecast $\boldsymbol{x}_t$. As with the parameters a and b, estimates for c and d are obtained by optimisation of an objective function over a training sample, details of which are provided in chapter 3. Forecast probabilities of the binary event $\{y_t \leq q\}$ are thus given by

$$\Pr(y_t \leq q) = \Phi\left(\frac{q - \mu_t^{\mathrm{NGR}}}{\sigma_t^{\mathrm{NGR}}}\right). \tag{2.21}$$

The forecast uncertainty, as represented by the NGR forecast variance $\sigma_t^{\mathrm{NGR}^2}$, is not fixed to a constant value as is the case in the more simplistic MOS post-processing method. Rather, the NGR method exploits spread-skill relationships between the uncertainty inherent in the ensemble forecasts and the predictability of the verifying

observations, and also provides a correction to the biases in forecast dispersion that are often found in operational settings. If spread-skill relationships do not exist we may reasonably expect the parameter d to tend to 0, and thus to recover the MOS forecast distributions in which the errors of the forecast mean, μ_t^{NGR} (equation (2.19)), are normally distributed with constant variance.

Gneiting et al. [2005] used the NGR post-processing method to issue probability forecasts of sea level pressure. The authors reported that the NGR forecasts were better calibrated with the verifying observations than the forecast derived from the raw ensemble forecasts. For example, the coverage of prediction intervals was more accurate. The authors reported that the estimate of the parameter d was negligibly small on several of the forecast occasions considered. This questions the necessity of the NGR model for that particular dataset — perhaps the simple MOS model would have been adequate. However, if spread-skill relationships do exist, we may reasonably expect the NGR forecasts to be more skilful than those given by the MOS post-processing method.

Coelho et al. [2004] also used the ensemble variance as a measure of forecast uncertainty in a Bayesian model for December ENSO forecasts. Prior distributions for the observations were given by a linear regression, where the observation for the previous July was used as a predictor variable, that is

$$y_{\mathrm{Dec},t} \sim N(a_0 + b_0 y_{\mathrm{Jul},t}, \sigma_{0t}^2), \tag{2.22}$$

where y_{Dec} and y_{Jul} denote the December and July observations, respectively, and the subscript t indexes the T observations in the dataset. The prior variance, σ_{0t}^2, is given by

$$\sigma_{0t}^2 = \sigma_0 \left[1 + \frac{1}{T} + \frac{\left(y_{\mathrm{Dec},t} - \bar{y}_{\mathrm{Dec}} \right)^2}{T s_{y_{\mathrm{Dec}}}^2} \right], \tag{2.23}$$

where $\bar{y}_{\mathrm{Dec}}$ and $s_{y_{\mathrm{Dec}}}^2$ denote the sample mean and variance of the T observations y_{Dec} in the dataset. In a second stage, the likelihood is obtained by regressing the ensemble means $\bar{x}_{\mathrm{Dec}}$ on the corresponding observations y_{Dec}, where the model variance is equal to the ensemble variance rescaled by a constant parameter, denoted γ. That is,

$$\bar{x}_{\mathrm{Dec}} \sim N(a_1 + b_1 y_{\mathrm{Dec}}, \gamma s_{\mathrm{Dec}}^2), \tag{2.24}$$

where $\bar{x}_{\mathrm{Dec}}$ and s_{Dec}^2 are the ensemble mean and variance of the ensemble forecasts for the December observations. As the authors had only a small dataset available, the parameters for the prior and likelihood distributions given above were calculated for the T observations using cross-validation, that is, each of the T forecasts and observations was left out, and the remaining $T - 1$ forecasts and observations were

used for parameter estimation. Applying Bayes Theorem, analytic results then show that the precision of the posterior distribution for the observation $y_{\text{Dec},t}$, $1/\sigma^2_{\text{Dec},t}$, is given by

$$\frac{1}{\sigma^2_{\text{Dec},t}} = \frac{1}{\sigma^2_{0t}} + \frac{b_1^2}{\gamma s^2_{\text{Dec},t}}, \tag{2.25}$$

the sum of the precisions of the prior distribution and the ensemble forecasting system for observation $y_{\text{Dec},t}$. Similarly, the mean of the posterior distribution, $\mu_{\text{Dec},t}$, is given by

$$\frac{\mu_{\text{Dec},t}}{\sigma^2_{\text{Dec},t}} = \frac{\mu_{0t}}{\sigma^2_{0t}} + \frac{b_1^2}{\gamma s^2_{\text{Dec},t}} \left(\frac{\bar{x}_{\text{Dec},t} - a_1}{b_1} \right), \tag{2.26}$$

where μ_{0t} is the expectation of the prior distribution for observation $y_{\text{Dec},t}$. The representation of forecast uncertainty using sums of precisions differs from that of the NGR model (see equation (2.20)). Inverting the expression for the precision leads to an estimate of the forecast variance, which is not a linear function of the ensemble variance. It would be interesting to compare the skill of the two methods in larger scale studies.

2.4.4.3 Logistic regression

Logistic regression (LR) is an alternative model for probability forecasts of binary events. Out-of-sample forecasts are given by

$$\Pr(y_t \leq q) = \frac{e^{\eta_t}}{1 + e^{\eta_t}}, \tag{2.27}$$

where η_t is the so-called linear predictor, a function that is linear in the model covariates. The method was successfully applied to the post-processing of ensemble forecasts in to probability forecasts in Hamill et al. [2004], who proposed the linear predictor

$$\eta_t = a + b\bar{x}_t + cs_t^2, \tag{2.28}$$

where the parameters a, b and c are estimated from a training sample. Hamill et al. found that including the ensemble variance did not improve the skill of probabilistic forecasts of precipitation, although Wilks [2006a] found the opposite for the Lorenz 1996 system [Lorenz, 1996].

The expression given in equation (2.27) is derived from the logit link function, that specifies the linear predictor η as a function of the probability forecast $p = \Pr(y \leq q)$, for a general predictor η and observation q. The Logit link function is

$$\log\{p/(1-p)\} = \eta, \tag{2.29}$$

and ensures that the forecast probabilities given by equation (2.27) are bounded by the interval $(0, 1)$, which would not be the case, for example, if using the identity link function $p = \eta$. Another popular choice of link function is the probit link, $\Phi^{-1}(p) = \eta$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

As with other post-processing methods, such as NGR, BMA and BMD, for which probability forecasts are issued with the respective CDFs, the probability forecasts issued by equation (2.27) result from evaluating the CDF of the logistic distribution at the threshold q. For the logit link function, the logistic CDF is given by

$$F(q) = \frac{1}{1 + e^{-(q-\mu_t^{\mathrm{LR}})/\sigma_t^{\mathrm{LR}}}}, \quad -\infty < q < \infty, \tag{2.30}$$

where μ_t^{LR} and $\sigma_t^{\mathrm{LR}^2}$ denote the expectation and variance of the logistic distribution at forecast time t. The subscript t indicates the possible dependence of μ_t^{LR} and $\sigma_t^{\mathrm{LR}^2}$ on statistics of the ensemble forecast $\boldsymbol{x}_t$.

Unfortunately, unlike other post-processing methods, such as NGR, the LR model does not provide an intuitive way for the user to exploit possible spread-skill relationships. This results from the fact that inclusion of the ensemble variance in the linear predictor η_t serves to alter the location, rather than the dispersion, of the logistic distribution. To see this, first observe that the form of equation (2.30) for $F(q)$ is recovered by multiplying the numerator and denominator of equation (2.27) by $e^{-\eta_t}$ and making the substitution

$$\eta_t = (q - \mu_t^{\mathrm{LR}})/\sigma_t^{\mathrm{LR}}.$$

For a fixed threshold q, therefore, changes in the linear predictor η_t correspond to changes in the quantity $(q-\mu_t^{\mathrm{LR}})/\sigma_t^{\mathrm{LR}}$. Inclusion of the ensemble variance, s_t^2, in the linear predictor does therefore not directly affect the standard deviation of the Logistic distribution, σ_t^{LR}, unlike the NGR model (see equation (2.20)). The inclusion of covariates in the linear predictor η_t does not have an intuitive interpretation in terms of adjusting the location, μ_t^{LR}, and variance, $\sigma_t^{\mathrm{LR}^2}$, of the Logistic distribution. In section 2.4.5 we describe an extension to the logistic regression model [Messner et al., 2014] that allows for a more intuitive use of model covariates in the context of ensemble post-processing.

Unlike the statistical models that are employed for continuous predictands, it is necessary to estimate separate LR model parameters for each threshold, q, of interest. This requirement introduces the undesirable possibility of inconsistent probability forecasts, that is, for two thresholds q_a and q_b with $q_a < q_b$, logistic regression parameters may imply $\Pr(y \leq q_a) > \Pr(y \leq q_b)$. The need to estimate separate parameters for each threshold of interest is somewhat burdensome.

2.4.5 Miscellaneous post-processing methods

Several variants of the post-processing methods described above have been developed with specific applications in mind. For probability forecasts of wind speed, Thorarinsdottir and Gneiting [2010] applied the truncated normal distribution with location and scale parameters governed by the ensemble mean and variance, in an analogous manner to the standard NGR post-processing method (see section 2.4.4.2). Friederichs and Thorarinsdottir [2012] applied the generalised extreme value (GEV) distribution to issue probability forecasts of daily maximum wind speeds, and Lerch and Thorarinsdottir [2013] proposed a regime-switching approach that used either of the two aforementioned models, depending on the value of the median of the ensemble forecast. The three methods all improve considerably on probability forecasts derived from the raw ensemble forecasts, such as the frequency of members forecasting the occurrence of an event. The regime-switching model [Lerch and Thorarinsdottir, 2013] appears to improve slightly on the single model approaches.

An alternative BMA approach was implemented for probabilistic forecasts of wind direction [Bao et al., 2010], where the dressing kernels take the form of the von Mises distribution. The post-processing of bivariate wind vectors has been studied by Pinson [2012] and Schuhen et al. [2012] in the context of bivariate Nonhomogeneous regression, and with a bivariate extension to BMA [Sloughter et al., 2013].

Due to its local (small scale) nature and highly skewed distribution, precipitation is a notoriously difficult variable to forecast, both deterministically and probabilistically. As described in section 2.4.2.2, the rank histogram recalibration method [Hamill and Colucci, 1997, 1998] improved the calibration of ensemble forecasts of precipitation using ideas based on the rank histogram. More recently, a variant of the BMA method [Sloughter et al., 2007] has yielded promising results. The distribution is a discrete-continuous mixture, with a point probability mass at 0 and a continuous probability distribution over the positive real axis. The probability $\Pr(y = 0)$ is estimated using logistic regression, and the remaining (strictly positive) forecast distribution is a mixture of M Gamma-distributed dressing kernels centred at the ensemble members. Alternatively, Wilks [2009] introduced an extension of logistic regression (see section 2.4.4.3), termed ELR, that overcomes the need to estimate separate parameters for multiple thresholds of interest. By including a monotonically increasing function of the threshold q, say $\omega(q)$ in the linear predictor (see equation (2.28)), fully continuous distributions are specified over the possible range of q. The extension therefore circumvents the need to fit the LR model for each threshold of interest. The difficulty lies in specifying an appropriate function $\omega(\cdot)$, although the cube root of the threshold value ($\omega(q) = q^{1/3}$) is said to be appropriate for precipitation forecasts. In addition, Messner et al. [2014] introduced a

further extension of Logistic regression that enables the user to separately adjust the location and variance of the Logistic forecast distribution, by specifying nonlinear, rather than linear predictors, with terms that correspond to the location μ^{LR} and variance σ^{LR^2}. This method is termed heteroscedastic extended logistic regression (HELR), and is appropriate for forecasts of binary predictands of many meteorological variables. Finally, Scheuerer [2014] proposed a regression-based post-processing method, modelling the distribution of precipitation observations conditionally on the ensemble forecast by exploiting the properties of the 3-parameter (location, scale and shape) GEV distribution. The forecast distribution is left-censored at 0, meaning that all probability mass on the negative real axis is shifted to a point mass at 0. Scheuerer concluded that the GEV-based method improved slightly on ELR, although we note that the method was not compared to HELR.

Stephenson et al. [2005] proposed a post-processing method known as Forecast Assimilation (FA) which, unlike the methods described previously, adopts the Bayesian philosophy. The authors suggest that, just as data assimilation is used to map information from the observed atmosphere to the NWP model initial conditions, so-called forecast assimilation should be used to infer the observations and the associated uncertainty from the available ensemble forecasts. Probability forecasts are constructed through the application of Bayes' theorem. In Stephenson et al. [2005] the probability forecasts are restricted to either univariate or multivariate Gaussian distributions. Due to its Bayesian nature, it is necessary to issue prior distributions for the parameters of interest. Stephenson et al. [2005] accomplish this by using separate datasets for the estimation of priors and, subsequently, the recalibration of operational forecasts. The so-called 'forecast operator' is used to link the ensemble forecasts to the verifying observations, and in Stephenson et al. [2005] is analogous to a standard linear regression approach, such as the MOS method (section 2.4.4.1). Indeed, if the same dataset is used for the estimation of prior distributions and forecast recalibration, Stephenson et al. [2005] state that forecast assimilation and MOS are equivalent. The probability forecast distributions produced by the FA post-processing method are therefore a function of both the prior specification of the model parameters, which are associated with the data assimilation process, and the parameters of the forecast operator, which is analogous to the regression-type approaches to ensemble post-processing described in this section.

In addition to the post-processing methods for bivariate wind vectors mentioned previously, several parametric approaches have been proposed that facilitate the post-processing of ensemble forecasts of multivariate quantities. The multivariate quantity may represent a spatial field of a single variable [Berrocal et al., 2007, 2008; Scheuerer and Büermann, 2014], or a field of multiple meteorological variables [Möller et al., 2013]. The methods proposed by Berrocal et al. [2007, 2008]

and Scheuerer and Büermann [2014] combined well-known geostatistical methods with state-of-the-art post-processing methods, such as BMA, to produce probabilistic forecast distributions of the multivariate quantities. On the other hand, Möller et al. [2013] uses a Gaussian copula to combine the marginal distributions of univariate meteorological variables, estimated with BMA, in to a multivariate distribution function. The Gaussian copula approach requires only the estimation of a covariance matrix that provides an estimate of the correlation between the various variables of interest. In the published literature, inclusion of a multivariate structure has proved beneficial to measurements of forecast skill of both univariate and multivariate quantities. Put simply, this is seemingly because more information contained within the NWP-based ensemble forecasts is utilised in the post-processing stage, and so we should expect improved probability forecasts of multivariate quantities. An approach for producing ensemble forecasts, rather than probability forecasts, of multivariate predictands is described next.

2.4.6 Ensemble copula coupling

Rather than producing multivariate probability forecasts with post-processing methods such as those described at the end of the previous subsection, the user may instead require multivariate ensemble forecasts. This is likely to be necessary for forecasts of high-dimensional variables, for which multivariate probability forecasts require specification of the (often complicated) dependence structure of the marginal variables. For example, even in the relatively simple case of multivariate Gaussian forecast distributions, it is necessary to state the high-dimensional multivariate covariance matrix, which is impractical for forecasts of predictands such as gridded temperature over a large area.

One might reasonably hope that a 'good' ensemble forecasting system will contain useful information about the dependencies between the many weather variables, as well as the spatial and temporal structure of the future atmospheric state. If this is indeed the case, then it follows that we should exploit the dependence structure of the ensemble forecasts in order to improve our forecasts of multivariate predictands. In this thesis we are concerned with forecasts over spatial fields, for which we provide examples in chapters 4 and 5.

Before constructing ensemble forecasts of multivariate variables, in our case for spatial fields, it is necessary to first obtain post-processed ensemble forecasts, rather than probability forecasts, for each of the R marginal variables. Fortunately this is readily achieved by drawing samples of the required ensemble size from the probability forecast distributions that are constructed by ensemble post-processing methods, such as those introduced in this section. That is, a standard ensemble post-

processing method is applied to the raw ensemble forecasts, in order to produce recalibrated probability forecast distributions from which recalibrated, M-member ensemble forecasts are sampled. The question is then how to sample the ensemble members? Schefzik et al. [2013] proposed the three following sampling schemes.

1. Random sampling — at each margin, ensemble forecasts $\widetilde{\boldsymbol{x}} = (\tilde{x}_{1,r}, \tilde{x}_{2,r}, \ldots, \tilde{x}_{M,r})$ for $r = 1, 2, \ldots, R$ are sampled as independent and identically distributed (IID) draws from the probability forecast distribution issued by the ensemble post-processing method at location r. This is achieved by setting

$$\tilde{x}_{m,r} = F_r^{-1}(u_m) \text{ for } m = 1, 2, \ldots, M, \tag{2.31}$$

   where $F_r^{-1}(\cdot)$ is the inverse of the CDF of the probability forecast distribution at location r, and the u_m are each IID realisations of a uniform-distributed random variable on the interval $[0, 1]$.

2. Quantile sampling — ensemble members $\tilde{x}_{m,r}$ are chosen as equidistant quantiles of the probability forecast distribution at location r, that is

$$\tilde{x}_{m,r} = F_r^{-1}\left(\frac{m}{M+1}\right) \text{ for } m = 1, 2, \ldots, M. \tag{2.32}$$

   As we shall see later in this chapter (section 2.5.4), this choice of ensemble members is close to optimal in the sense of optimising a commonly used measure of forecast skill.

3. Transformational sampling — Firstly, a probability distribution is fitted to the initial, raw ensemble forecasts, and the quantile values of each member are calculated. For example, a Gaussian distribution with mean $\bar{x}$ and variance s^2, where $\bar{x}$ and s^2 denote the ensemble mean and variance, could be fitted to the raw ensemble forecasts. The CDF values for each ensemble member, $q_m = \Phi((x_m - \bar{x})/s)$ are then calculated for all $m = 1, 2, \ldots, M$. The post-processed ensemble forecast is then given by:

$$\tilde{x}_{m,r} = F_r^{-1}(q_m). \tag{2.33}$$

The idea of exploiting the dependence structure of ensemble forecasts was proposed by Flowerdew [2012], in an article that concerned spatial forecasts of precipitation. The author pointed out that, in post-processing the ensemble forecasts at each location, the dependence structure between forecast locations is lost. If the raw ensemble forecasts do indeed contain useful information about the dependence structure of the verifying, multivariate observations, then we might reasonably expect to improve the skill of the post-processed multivariate ensemble forecasts by incorporating the

dependence in to our recalibrated, post-processed forecasts. To quote Flowerdew [2012]: "The key to preserving spatial, temporal and inter-variable structure is how this set of values is distributed between ensemble members. One can always construct ensemble members by sampling from the calibrated PDF, but this alone would produce spatially noisy fields lacking the correct correlations. Instead, the values are assigned to ensemble members in the same order as the values from the raw ensemble: the member with the locally highest rainfall remains locally highest, but with a calibrated rainfall magnitude."

The methodology known as ensemble copula coupling (ECC), introduced by Schefzik et al. [2013] is in essence a generalisation of the ideas proposed by Flowerdew [2012]. The ECC methodology uses the empirical copula of the raw, multivariate ensemble forecasts to combine the dependence structure of the raw forecasts with recalibrated ensemble members that are sampled from probability distributions as described above. The dependence structure of the raw ensemble forecasts at multiple locations is represented by the rank correlation structure, or empirical copula, which can be viewed as an empirical representation of the multivariate copulas discussed in the previous subsection. The ECC method proceeds as follows. Firstly, the empirical copula of the raw, multivariate ensemble forecasts is calculated. Given an M-member ensemble forecast of an R-dimensional field, calculation of the empirical copula simply involves calculating the rank order of the M ensemble members for each of the R marginal variables. For each marginal variable, therefore, the rank order is a permutation of the integers $\{1, 2, \ldots, M\}$. We denote the permutations by π_r, for $r = 1, 2, \ldots, R$,

$$\pi_r = (\text{rank } x_{1,r}, \text{rank } x_{2,r}, \ldots, \text{rank } x_{M,r}), \tag{2.34}$$

where $\text{rank } x_{m,r} = \sum_{l=1}^{M} \mathbf{I}(x_{l,r} \leq x_{m,r})$ is the rank of the ensemble member $x_{m,r}$ among the M members of the ensemble forecast $\boldsymbol{x}_r$. The R-dimensional ensemble members, $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_M$ are then obtained by applying the permutation π_r to the ensemble forecasts sampled at each of the R margins. That is

$$\widehat{\boldsymbol{x}}_r = \pi_r(\widetilde{\boldsymbol{x}}_r), \text{ for } r = 1, 2, \ldots, R, \tag{2.35}$$

where the $\widetilde{\boldsymbol{x}}_r$ are ensemble forecasts sampled from probability forecast distributions, as described above. By reordering the post-processed ensemble members to the order observed in the raw ensemble forecasts, the rank dependence structure of the raw, multivariate ensemble forecast is recovered, while the marginal ensemble forecasts should be better calibrated with the verifying observations, in keeping with the foregoing quote from Flowerdew [2012].

Schefzik et al. [2013] illustrated the use of the ECC methodology with an example

of spatial forecasts of near-surface temperature and air pressure over a spatial field defined by three German airports. It was shown that by retaining the rank dependence structure of the raw ensemble forecasts, the skill of multivariate forecasts was improved compared to ensemble forecasts that were only recalibrated at the margins.

2.4.7 Parameter estimation

In this section we give a brief overview of two objective functions that are used for parameter estimation in many of the examples used throughout this thesis. We also review the method of moments, which we use in chapter 4. We provide an outline only, and defer details for specific post-processing methods until the relevant sections in the subsequent chapters. As mentioned earlier in this chapter, the objective functions are calculated over training samples of N historical ensemble forecasts and verifying observations. The section concludes with a description of two numerical optimisation routines that are used to find the 'optimal parameter estimates', which are those estimates that minimise the objective function over the training sample.

2.4.7.1 Parameter estimation by objective function minimisation

We make frequent use of the well known likelihood framework for parameter estimation. Let ψ denote the vector of parameters that are to be estimated in the statistical model, and denote the estimate of ψ by $\widehat{\psi}$. For example, the parameter vector for the NGR method (see section 2.4.4.2) is $\psi = (a, b, c, d)'$, and its estimate is $\widehat{\psi} = (\hat{a}, \hat{b}, \hat{c}, \hat{d})'$. The likelihood function for the statistical model under consideration is a function of ψ, conditional on the training sample, in our case the ensemble forecasts and observations $(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, N$. In this thesis we work with the negative logarithm of the likelihood (negative log-likelihood, NLL). The logarithmic likelihood is often easier to work with for reasons of algebraic simplicity, and we negate it so that the parameter estimate $\widehat{\psi}$ is found by minimising the objective function. The general form of the NLL is

$$\text{NLL}(\widehat{\psi} \mid \text{ data }) = -\sum_{i=1}^{N} \log f(y_i \mid \boldsymbol{x}_i, \widehat{\psi}), \tag{2.36}$$

where $f(\cdot)$ denotes the PDF of the forecast distribution for the observations, which as before is conditional on the corresponding ensemble forecasts and parameter vector. The optimal parameter estimates are those that minimise equation (2.36), and are referred to as the 'likelihood parameter estimates' hereafter. Likelihood parameter estimates have several appealing properties. Not least, under correct model

assumptions, the parameter estimates are asymptotically distributed as multivariate Gaussian random variables, with mean vector ψ (the 'true' parameter vector), and covariance matrix that can be calculated explicitly from the likelihood function. The terms in the covariance matrix are typically inversely proportional to the training sample size, N, meaning that parameter uncertainty decreases asymptotically as the size of the training sample increases.

In chapter 3 we make particular use of parameter estimates obtained by minimising an alternative objective function, namely the continuous ranked probability score (CRPS, Matheson and Winkler [1976]). The CRPS is discussed in the context of verifying the skill of probability forecasts in section 2.5.3, for which it is more commonly employed. The general form for the CRPS, calculated over a training sample is

$$\text{CRPS}(\hat{\psi} \mid \text{ data }) = \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{\infty} \left\{ F(u \mid \boldsymbol{x}_i, \hat{\psi}) - \mathbf{I}(y_i \leq u) \right\}^2 du. \tag{2.37}$$

Here u is a dummy variable, and $F(u \mid \boldsymbol{x}_i, \hat{\psi})$ is the cumulative distribution function (CDF) issued by the statistical post-processing model under consideration, which is conditional on ensemble forecast $\boldsymbol{x}_i$ and the parameter estimate $\hat{\psi}$. The integral given in equation (2.37) can be written in a closed form for several probability distributions used in the ensemble post-processing literature, examples of which are given in chapter 3. The parameter estimates $\hat{\psi}$ that minimise equation (2.37) are hereafter referred to as the 'CRPS parameter estimates'.

The CRPS was first used as an objective function for parameter estimation for the NGR model [Gneiting et al., 2005]. The authors state that the resulting probability forecast distributions are more skilful and sharper than those issued with NLL parameter estimates. In chapter 3 we provide a more comprehensive comparison of the skill of probability forecasts using the NLL and CRPS parameter estimates for several post-processing methods.

2.4.7.2 Parameter estimation by the method of moments

In chapter 4 we use the method of moments to obtain parameter estimates of a newly developed ensemble post-processing method. The method of moments derives parameter estimates by equating the expectations of summary statistics of the training data with their realised values, where the expectations are defined according to the assumed model for the observations, conditional on the ensemble forecasts. The parameter estimates are found by solving the resulting system of equations. It is therefore necessary to derive at least as many equations as unknown param-

eters, and to ensure that the system gives a unique solution. As suggested by its name, the method of moments requires only assumptions about the moments of the forecasts and observations, rather than distributional assumptions, as are required when optimising an objective function such as the NLL or CRPS. Therefore, the method of moments is, in a sense, less restrictive than other parameter estimation routines. However, the properties of moment-based estimators are less appealing — for example, the uncertainty in moment-based estimators is generally larger than their likelihood-based counterparts [Gillard, 2014].

2.4.7.3 Numerical optimisation routines

In general, closed-form expressions do not exist for parameter estimates in the likelihood and minimum CRPS estimation frameworks, and we must resort to numerical optimisation routines. These routines seek to find the minimum of the multidimensional objective function surface that is enclosed by the space described by the possible parameter values, known as parameter space. The parameter estimates are then the set of parameters that minimise the objective function surface. Numerical optimisation forms a research field in its own right, and a detailed discussion is beyond the scope of this thesis. In subsequent chapters we make frequent use of two algorithms: the Nelder-Mead [Nelder and Mead, 1965] and Broyden-Fletcher-Goldfarb-Shanno (BFGS), published simultaneously by the aforementioned authors in 1984, which are outlined below. We typically use the Nelder-Mead algorithm to find initial parameter estimates, and then initialise the BFGS algorithm with these estimates to obtain the final parameter estimates. We use the in-built implementations available in the R language [R Core Team, 2015].

The Nelder-Mead algorithm is a 'simplex' algorithm, that does not rely on derivatives of the objective function and so is often suitable for finding the minima of non-differentiable functions. Briefly, a simplex of $D + 1$ 'test points' in $\mathbb{R}^D$ is formed, where D is the dimensionality of the parameter space. The algorithm identifies the worst test point, the point at which the objective function attains the largest value among the $D + 1$ test points. This worst point is reflected through the centroid of the test points, the idea being that smaller values of the objective function will exist in the area of parameter space occupied by the other test points. Depending on the value of the objective function at the reflected point, the simplex is either expanded or contracted in the direction of 'steepest descent' of the objective function surface. The process is iterated until the algorithm can identify no smaller values of the objective function, at which time the algorithm stops and reports the best test point as the parameter estimates. The Nelder-Mead algorithm is susceptible to finding local, rather than global minima, and so care is needed to ensure that

the parameter estimates are not influenced by the starting values supplied to the algorithm, which are typically chosen by the user.

The BFGS algorithm is a quasi-Newton method that relies on derivatives of the objective function. The derivatives can either be supplied in closed form by the user, or are otherwise estimated by the numerical routine. The derivative of the objective function is necessarily 0 at the minima, with the exception of minima that lie on the boundary of the parameter space. The BFGS algorithm therefore seeks points in parameter space that correspond to roots of the derivative of the objective function. The method requires an approximation to the Hessian matrix, the square matrix of second partial derivatives of the objective function with respect to the parameters, which is then evaluated at the parameter estimates $\hat{\psi}$. Having obtained this estimate, say B, using a Taylor series expansion of the scalar-valued objective function $O : \mathbb{R}^D \to \mathbb{R}$ leads to the following equation:

$$B^{(k)}\Delta\psi = -\nabla O(\psi^{(k)}).$$

Solving this equation for $\Delta\psi$ yields the new 'search direction' for the optimal parameters. Here the superscript refers to the kth iterate of the algorithm, and ∇ denotes the vector of partial derivatives. The value $\hat{\psi}^{(k)}$ is updated according to

$$\hat{\psi}^{(k+1)} \propto \hat{\psi}^{(k)} + \Delta\psi.$$

The update $\Delta\psi$ is in fact scaled by a constant, the technical details of which are omitted here. The algorithm is deemed to have converged when the magnitude of the vector of derivatives, say $\|\nabla\psi^{(k)}\|$, is smaller than some pre-defined tolerance.

2.5 Forecast verification

2.5.1 Introduction

In this section we give details for a variety of techniques and methods that we use for the assessment of ensemble and probability forecasts throughout chapters 3–6 of this thesis. Rigorous assessments of forecast skill are important, not just to rank competing forecasts, but to aid in determining deficiencies in current post-processing methods that can subsequently be improved upon. Graphical tools can aid in the diagnosis of the strengths and weaknesses of the statistical post-processing models, and can suggest potential areas for improvement. Scoring rules, on the other hand, provide quantitative measures of a forecast's performance, usually in the form of a single number that summarises the skill of the forecast. Scoring rules reward

forecasts for both their reliability and resolution (see section 2.3) or, equivalently, penalise forecasts that are not well-calibrated. Different scores penalise certain characteristics of a forecast more heavily than others, as we discuss in section 2.5.3.2. In this thesis we use two classes of scoring rules: proper scoring rules [Gneiting and Raftery, 2007], which are appropriate for the evaluation of probability forecasts, such as those outlined in sections 2.4.3.2, 2.4.4.2 and 2.4.5, and fair scores [Ferro, 2014], which are appropriate for the evaluation of ensemble forecasts that are interpreted as IID samples from underlying ensemble distributions.

In this thesis we consider out-of-sample forecasts — that is, probability and/or ensemble forecasts that are not contained within the training sample used for parameter estimation (see section 2.4.7). As with the examples provided in section 2.4, we make clear this distinction in our notation, by calculating verification scores for verifying observations y_t, where $t = 1, 2, \ldots, T$ indexes the forecasts and observations contained in a test dataset of size T.

2.5.2 Graphical assessments of forecast skill

2.5.2.1 Diagnostic plots using model residuals

Assessing the calibration of statistical models by inspection of the residuals is a widely used exploratory technique in applied statistics. Let r_t denote the residual on forecast occasion t,

$$r_t = \hat{y}_t - y_t, \tag{2.38}$$

where $\hat{y}_t$ is the predicted (deterministic) forecast of y_t given by the ensemble post-processing method — for example, in section 2.4.4.2 we explained that the expectation of the NGR forecast distribution, say μ_t, can be used as a deterministic forecast (see equation (2.19)). Plotting model residuals against covariates serves as a useful aid both in checking the validity of model assumptions and, if appropriate, identifying possible model improvements.

It is often useful to supplement the scatter plots described below with a line of best fit, or 'scatter plot smoother'. These lines approximate the empirical expectation of the dependent variable, such as the residuals, throughout the plotted range of the independent variable, and thus provide a basis for suggesting model improvements. We achieve this by using the Loess method as implemented in the R language [R Core Team, 2015]. The Loess method estimates a line of best fit using local polynomial regression. At each point in the scatter plot, the Loess line is estimated using a local neighbourhood of points, where the size of the neighbourhood is controlled by the user and influences the amount of smoothing that is performed. The influence

of each point in the neighbourhood is governed by a weight function. In our usage in this thesis, the weights are inversely proportional to the cubed distance of the points in the neighbourhood from the point at which the line is to be estimated.

We now provide some examples of how such scatter plots might be used, although the appropriate choice of plot should be determined by the situation at hand. Scatter plots of the residuals, r, as a function of the post-processed forecast mean should, if the deterministic forecasts (the post-processed mean) are well calibrated, appear as white noise distributed around the line $r = 0$. Systematic departures from the 0 line are indicative of misspecification of the forecast mean, and may provide evidence for alternative specifications of the mean of the probability forecast distributions.

Scatter plots of model residuals can also be used to assess the variance of probability forecast distributions. Define the squared standardised residuals,

$$\text{ssr}_t = \frac{r_t^2}{\sigma_t^2}, \tag{2.39}$$

where again σ_t^2 is the forecast variance. If the post-processing model is correctly specified, the squared standardised residuals $\text{ssr}_t, t = 1, 2, \ldots, T$ have expectation 1. Plots of ssr_t against the forecast variance can therefore be used to assess the calibration of the forecast variance — in this case, the Loess curve should follow the line ssr = 1 throughout the range of the forecast variance.

The residuals can also be used to assess the distributional assumptions of ensemble post-processing methods. Let $\text{sr}_t = r_t/\sigma_t$ denote the standardised residuals, for $t = 1, 2, \ldots, T$. If the probability forecast distributions are well-calibrated, the standardised residuals should each have expectation 0 and unit variance, with distribution that is assumed by the statistical model. For example, for the Gaussian forecasts issued by the NGR post-processing method, the standardised residuals should appear as IID draws from the standard normal distribution $N(0, 1)$. So-called quantile-quantile (or Q-Q) plots are commonly employed to assess the distributional assumptions. This is a plot of the pairs $(q_{t/(T+1)}, \text{sr}^{(t)})$ for $t = 1, 2, \ldots, T$, where $q_{t/(T+1)}$ denote the T equidistant quantiles of the (standardised) probability forecast distribution, and $\text{sr}^{(t)}$ denotes the order statistics of the standardised residuals. Statistical models that are calibrated with the verifying observations yield Q-Q plots that lie on the diagonal line that has intercept 0 and gradient 1.

2.5.2.2 Reliability diagrams

Reliability diagrams provide a graphical assessment of forecast calibration for probability forecasts of binary events, i.e. forecasts of the form $\Pr(y \leq q)$, where y is

the verifying observation and q is a threshold of interest. Such forecasts are the subject of chapter 3. Reliability diagrams are also related to the reliability component of the decomposition of the Brier score, discussed in section 2.5.5 below. Firstly, probability forecasts are divided in to mutually exclusive bins that cover the interval $(0, 1)$, for example $(0, 0.05], (0.05, 0.1], \ldots, (0.95, 1)$. We denote by K the number of bins, and $I_k, k = 1, 2, \ldots, K$ the indices of forecasts that fall in bin k. Now let $z_t = \mathbf{I}(y_t \leq q)$ denote the binary observation of the event $\{y_t \leq q\}$, for $t = 1, 2, \ldots, T$. The reliability diagram is a plot of $\bar{z}_k = \frac{1}{N_{I_k}} \sum_{t \in I_k} z_t$, the arithmetic mean of the binary observations, against $\bar{p}_k = \frac{1}{N_{I_k}} \sum_{t \in I_k} p_t$, the arithmetic mean of the probability forecasts in each bin, for all $k = 1, 2, \ldots, K$, where N_{I_k} denotes the number of forecasts in bin k.

Reliable probability forecasts therefore lie on the diagonal line (within sampling variation). The shape of the reliability diagram can further inform the user as to possible deficiencies in the calibration of the probability forecasts. For example, reliability curves that consistently lie below or above the diagonal are indicative of biased probability forecasts. S-shaped reliability curves are indicative of probability forecasts that are either overconfident or underconfident, depending on the orientation of the curve. For example, a curve that lies above the diagonal for small forecast probabilities, but below for large forecast probabilities is indicative of an overconfident forecasting system — events forecast to occur with small probability realise more often than they should, while events forecast to occur with high probability realise less often than they should, if the probability forecasts $\Pr(y \leq q)$ were correctly calibrated.

2.5.2.3 Rank and PIT histograms

While the reliability diagrams described above are a useful diagnostic tool for probability forecasts of binary events, they do not assist in the diagnosis of the strengths and deficiencies of continuous forecast distributions. In this section, therefore, we describe the rank and PIT histograms, that aid in assessing the calibration of ensemble forecasts and continuous probability forecast distributions, respectively. The rank and PIT histograms are typically calculated over the test dataset of out-of-sample forecasts and observations.

The probability integral transform (PIT) evaluates the forecast CDF, denoted F, at the verifying observation y. If y is indeed a draw from F, as in the idealised case, then the PIT values $F_t(y_t), t = 1, 2, \ldots, T$ follow a uniform distribution. To see this

observe that

$$\begin{aligned}\Pr(F_t(y_t) \leq u) &= \Pr(y_t \leq F_t^{-1}(u)) \\ &= F_t(F_t^{-1}(u)) \\ &= u\end{aligned}$$

where each F_t is taken to be uniquely invertible and monotonically increasing. Note that if the CDF of y_t is in fact $G_t \neq F_t$, then the quantity $G_t(F_t^{-1}(u))$ provides a measure of the discrepancy between F_t and G_t. Therefore, plotting a histogram of PIT values calculated over all available forecast-observation pairs, referred to as the PIT histogram, provides a graphical assessment of the departure of the distribution of PIT values from uniformity, and is thus useful for assessing the assumption that the observations, $y_t, t = 1, 2, \ldots, T$, have CDFs F_t. The interpretation of such histograms is discussed below.

An analogous graphical tool is available for assessing ensemble forecasts. In section 2.4.2.2 we discussed how the distribution of the rank of the observations when pooled with the ensemble forecasts in the training sample can be used to produce recalibrated out-of-sample probability forecasts [Hamill and Colucci, 1997, 1998]. Equation (2.7) gives the formula for calculating the relative frequency of the observation ranks, which can take values in the set $\{1, 2, \ldots, M + 1\}$, where M denotes the ensemble size. As well as being used to recalibrate the raw ensemble forecasts, a histogram plot of the values $w_1, w_2, \ldots, w_{M+1}$, where here the w_m are calculated for the post-processed ensemble forecasts and verifying observations in the test dataset, provides a useful graphical assessment of the calibration of the ensemble forecasting system [Anderson, 1996; Hamill and Colucci, 1997]. A calibrated ensemble forecasting system yields uniform (or flat) rank histograms with bin heights of $1/(M + 1)$ (within sampling variation).

The shapes of PIT and rank histograms can be informative guides to the strengths and deficiencies of the calibration of probability and ensemble forecasts, respectively. U-shaped histograms are indicative of forecast underdispersion — they imply that a larger than expected proportion of observations fall in the tails of probability distributions (PIT histograms), or outside the range of the ensemble forecasts (rank histograms). Similarly, inverted U-shaped histograms are indicative of overdispersed forecasts. Sloped histograms are indicative of systematic forecast bias in either the location of probability forecast distributions (PIT histograms) or the ensemble members (rank histograms). This follows from the fact that biases in forecast location result in observations falling more often in one tail of the distribution than the other. For example, more than half of forecasts issued by a probabilistic forecasting system whose median is on average larger than the observations will lead to more than half

of the observations falling in the lower tail of the forecast distributions, and therefore the resulting PIT histogram will exhibit a negative gradient.

While the uniformity of PIT and rank histograms is a necessary condition for calibrated forecasts, it is not sufficient. This was illustrated by Hamill [2001], who provided examples of uncalibrated ensemble forecasts that still resulted in flat rank histograms. The paper highlights the fact that conditional biases in forecast calibration within individual strata of the historical forecast data can be overlooked by histograms of the entire data, which essentially smooth the rank distribution over all cases. Hamill suggested stratifying the data and calculating rank histograms for each stratum, in order to assess the calibration of forecasts within separate strata. The author suggested that the stratification could be performed, for example, by conditioning on covariates such as the ensemble mean and variance.

Gneiting et al. [2008] provided a broad and informative discussion on the verification of forecasts of multivariate predictands. The authors introduced the multivariate rank histogram, which is an extension of the rank histogram (described above) for assessing the calibration of ensemble forecasts of multivariate predictands. The MRH maps D-dimensional ensemble members $x_{t1}, x_{t2}, \ldots, x_{tM} \in \mathbb{R}^D$ and observations $y_t \in \mathbb{R}^D$ to 'multivariate rank values', $\mathrm{MR}_t \in \{1, 2, \ldots, M+1\}$. The MRH is then the histogram of multivariate rank values MR_t for $t = 1, 2, \ldots, T$. More formally, the MRH is calculated as follows. To ease notation, let x_0 temporarily denote a general multivariate observation y. Firstly, for ensemble members x_j and $x_k \in \mathbb{R}^D$, define

$$x_j \preceq x_k \Leftrightarrow x_{j,l} \leq x_{k,l} \quad \forall l = 1, 2, \ldots, D. \tag{2.40}$$

In words, $\mathbf{I}(x_j \preceq x_k) = 1$ if and only if the vector x_j lies below the vector x_k in all elements of the D-dimensional Euclidean space. The 'pre-ranks' of the observation and the M ensemble members x_m are calculated as

$$\mathrm{PR}_m = \sum_{k=0}^{M} \mathbf{I}(x_k \preceq x_m) \text{ for } m = 0, 1, \ldots, M. \tag{2.41}$$

Finally, the multivariate rank is the rank of the pre-rank of the observation, PR_0, with any ties resolved at random. Following the notation of Gneiting et al. [2008], if $s^< = \sum_{m=0}^{M} \mathbf{I}(\mathrm{PR}_m < \mathrm{PR}_0)$ and $s^= = \sum_{m=0}^{M} \mathbf{I}(\mathrm{PR}_m = \mathrm{PR}_0)$, then the multivariate rank is chosen randomly from the set $\{s^< + 1, \ldots, s^< + s^=\}$. If $s^= = 1$ (in the case of no ties in the pre-ranks of the observation and ensemble members), then the multivariate rank is simply $\mathrm{MR} = s^< + 1$. The multivariate ranks take values in the set $\{1, 2, \ldots, M+1\}$. The multivariate rank histogram is the histogram of the multivariate ranks MR_t for $t = 1, 2, \ldots, T$. Observe that in the special case of $D = 1$, the multivariate rank histogram reduces to the rank histogram for univariate

predictands.

Gneiting et al. [2008] stated that interpretations of the multivariate rank histogram are the same as those for its univariate counterpart. Plainly the distribution of the multivariate ranks $\mathrm{MR}_t, t = 1, 2, \ldots, T$ is uniform on the set $\{1, 2, \ldots, M+1\}$ if the multivariate observation and ensemble members are indeed IID realisations of the same multivariate distribution. Ensemble members whose vector-valued forecasts are consistently biased will yield pre-ranks that, on average, are smaller (if the ensemble 'underforecasts' the observation) or larger (if the ensemble 'overforecasts' the observation) than the pre-rank of the observation, leading to skewed, or 'sloped' rank histograms. In our view, interpreting characteristics of ensemble dispersion from multivariate rank histograms is less intuitive. However, intuition can be gleaned by considering the hypothetical case in which the (assumed) multivariate ensemble distribution is of the same statistical form as the observation distribution, but where the covariance structure is scaled by a constant, say c. If $c < 1$, it follows that the pre-rank of the observation will populate the outer values in the set $\{1, 2, \ldots, M+1\}$ more often than one would like, with the opposite being the case for $c > 1$. This feature yields U-shaped (inverse U-shaped) multivariate rank histograms for $c < 1$ ($c > 1$), analagous to the shapes of univariate rank histograms for underdispersed (overdispersed) ensemble forecasts.

Gneiting et al. [2008] recommended that the multivariate rank histogram should only used for fairly low-dimensional predictands. Higher dimensions result in an excessive number of pre-rank ties, as instances of one forecast lying below the other in all D dimensions are rare. Indeed, in chapter 4 we use the multivariate rank histogram for a 4-dimensional predictand, and find that tied pre-ranks occur often, even in this low-dimensional setting.

2.5.2.4 Quantile regression

Quantile regression [Koenker, 2005] enables the estimation of specific quantiles of a distribution as a function of model covariates, rather than just its expectation as in standard regression models. While the PIT and rank histograms detailed above provide diagnostics of forecast calibration over all forecast cases, we also consider the idea of assessing distributions of verification measures, such as the PIT values, as a function of important covariates. For example, in the idealised setting, the deciles of the distribution of PIT values are exactly equal to $0.1, 0.2, \ldots, 0.9$, independently of any covariates, and significant departures from the idealised deciles are indicative of forecast misspecification, possibly as a result of an incorrect usage of the covariate under consideration. Further details are given in chapter 6, in which we discuss some preliminary results of using quantile regression.

2.5.3 Scoring rules for probability forecasts

2.5.3.1 The notion of propriety

As before, let f and F denote the PDF and CDF of a forecast distribution on a general forecast occasion, and let y denote the verifying observation. The probability forecast distribution and verifying observation may be univariate or multivariate. An important principle in the verification of probability forecasts is the notion of propriety. Proper scoring rules [Gneiting and Raftery, 2007] reward forecasters who are honest when issuing their probability forecast distributions. In other words, a forecaster who is honest when issuing their probability forecast will not want to have issued an alternative forecast if they subsequently learn that a proper scoring rule is to be used for forecast verification. Proper scores are often used as a means of ranking competing forecasts.

More formally, we define proper scoring rules as follows.

Definition 2.5.1 *Let $h(f, y)$ be a real-valued function of the forecast distribution, f, and verifying observation, y, and let q denote the probability distribution for y. The scoring rule $h(f, y)$ is proper with respect to any class of probability distributions, Q, if*

$$E_q\{h(q, y)\} \leq E_q\{h(f, y)\} \text{ for all } f \text{ and } q \in Q, \tag{2.42}$$

where in the above equation the expectations are calculated with respect to y. The score $h(f, y)$ is strictly proper if its expectation is uniquely optimised when $f = q$.

In this thesis we use proper scores that are negatively orientated, that is, smaller values are preferred. This is in keeping with the notation of definition 2.5.1.

2.5.3.2 Examples of proper scores

Many proper scoring rules have been proposed in the literature. The choice of score is influenced by the type of forecast under consideration — for example, whether the forecast variable is binary or continuous, the variable's dimensionality (univariate or multivariate), as well as which properties of the forecast are of most interest to the user. For example, certain scores penalise outlying observations more heavily than others, while others penalise forecasts more severely for biases in their location. All proper scores, however, reward forecasts for their reliability and resolution. Here we highlight some commonly applied proper scores that we make use of throughout subsequent chapters of this thesis.

The quadratic, or Brier score [Brier, 1950] is widely used for the assessment of

probability forecasts of binary predictands. In this thesis we use the Brier score to assess the skill of probability forecasts of the form $p_t = \Pr(y_t \leq q)$, where y_t denotes a (usually continuous) observation, and q is a threshold of interest. The score is given by

$$h_{Brier}(p_t, y_t) = (p_t - z_t)^2 , \tag{2.43}$$

where (as in section 2.5.2.2) $z_t = \mathbf{I}(y_t \leq q)$ denotes the binary observation of the event $\{y_t \leq q\}$. The limiting values of h_{Brier} are $h_{Brier}(0,0) = h_{Brier}(1,1) = 0$ and $h_{Brier}(0,1) = h_{Brier}(1,0) = 1$, which are attained if the forecaster correctly (incorrectly) issues forecasts of complete certainty about the event $\{y_t \leq q\}$. The quadratic nature of the Brier score implies that the reward for confident, or sharp, calibrated forecasts grows in a quadratic manner, while sharp, uncalibrated forecasts — for example, small forecast probabilities for events that occur with high frequency — are similarly penalised.

In section 2.4.7.1 we introduced the continuous ranked probability score (CRPS, see equation (2.37)) in the context of parameter estimation. However, the CRPS is more commonly used as a proper scoring rule for assessing the skill of probability forecasts for (usually continuous) predictands. As can be seen from equation (2.37) and (2.44) (below), the CRPS has the appealing interpretation as the integral of the Brier score over all possible thresholds q. For an out-of-sample probability distribution of the observation y_t, with CDF F_t we have

$$h_{CRPS}(F_t, y_t) = \int_{-\infty}^{\infty} \{F_t(q) - \mathbf{I}(y_t \leq q)\}^2 \, dq. \tag{2.44}$$

Just as the Brier score is a measure of distance of the probability forecast of $\{y_t \leq q\}$ from the verifying binary observation, it is easy to infer from equation (2.44) that the CRPS is, in a sense, a measure of the disparity between the forecast CDF F_t and the CDF that would be issued if the forecaster had perfect knowledge about the observation y_t, namely the indicator function $\mathbf{I}(y_t \leq q)$. Like the Brier score, the CRPS favours forecasts that are sharp by penalising the disparity between the CDFs of the forecast distribution and the 'truth' in a quadratic manner. Indeed, it is straightforward to show that the CRPS reduces to the Brier score for binary predictands (see section 3.3.3.3).

Gneiting and Raftery [2007] showed that the CRPS can alternatively be written with the following, appealing construction:

$$h_{CRPS}(F_t, y_t) = E_{F_t}(|y_t - x_t| - \frac{1}{2}|x_t - x_t'|), \tag{2.45}$$

where x_t and x_t' are independent copies of a random variable with distribution function F_t, and $|\cdot|$ denotes the Euclidean norm. In the above equation the expectations

are calculated with respect to x_t and x'_t. Equation (2.45) shows that the CRPS is measured in the same units as the verifying observation y_t. Due to the nature of equations (2.44) and (2.45), closed forms for the CRPS do not generally exist, although expressions have been derived for several families of probability distributions. We defer giving closed forms of the CRPS for specific distributions until the relevant later sections.

The ignorance, or logarithmic score [Good, 1952] provides an alternative measure of forecast skill.

$$h_{ign}(f_t, y_t) = -\log f_t(y_t), \tag{2.46}$$

where often the logarithm is taken to the base 2. The ignorance score is a so-called 'local score', as it is determined completely by the value of the forecast PDF at the observation. The ignorance score issues harsh penalties to outlying, unlikely observations, due to the rapid growth rate of the logarithmic function as the probability density function (PDF) $f_t(y_t)$ tends to 0.

Finally we introduce the energy score, which is used for assessing probability forecasts of multivariate variables, such as spatial fields of near-surface temperature. The propriety of this score was proven in Székely [2003], and is discussed in the context of weather forecasting by Gneiting and Raftery [2007]. Here the cumulative distribution functions F_t are multivariate, as are the verifying observations y_t. In its most general form, the energy score is defined as

$$h_{ES}(F_t, y_t) = E_{F_t}\left(\|y_t - x_t\|^\beta - \frac{1}{2}\|x_t - x'_t\|^\beta\right), \tag{2.47}$$

where $\beta \in (0, 2)$, $\|\cdot\|$ denotes the Euclidean norm, x_t and x'_t are independent copies of random variables with distribution function F_t, and the expectations are taken with respect to x_t and x'_t. We follow Gneiting and Raftery [2007] and set $\beta = 1$. Observe that in the univariate case (with $\beta = 1$) the energy score reduces to the continuous ranked probability score — equation (2.47) reduces to (2.45).

2.5.4 Assessing ensemble forecasts with fair scoring rules

In section 2.5.3 we defined proper scoring rules for the assessment of probability forecasts, and stated (see definition 2.5.1) that their use encourages forecasters to be honest when stating their beliefs. Unfortunately, the situation is more complicated when assessing the skill of ensemble forecasts, due to their nature as finite samples. Unlike probability distributions, it is possible to attain improved score values by hedging an ensemble forecast. For example, let $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$ denote a M-member ensemble forecast with verifying observation y, and suppose we interpret the

EDF of $\boldsymbol{x}$ as a probability forecast distribution for y. The CRPS for this distribution, which we denote by E_y, is given by

$$\mathrm{CRPS}(E_y, y) = \frac{1}{M}\sum_{m=1}^{M}|y - x_m| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{n=1}^{M}|x_m - x_n|\,. \tag{2.48}$$

Bröcker [2012] showed that resetting the ensemble members x_m to the quantile values

$$\hat{x}_m = F^{-1}((2m-1)/(2M)) \text{ for } m = 1, 2, \ldots, M$$

minimises the expectation of equation (2.48), where the function $F(\cdot)$ denotes the CDF of y. In other words, the value of the CRPS given by equation (2.48) could be improved by resetting the ensemble members to the above quantiles, even if the ensemble forecast was calibrated in the sense described in section 2.3.2. In practice F is unknown but, nonetheless, this result demonstrates the idea that ensemble forecasts can be 'hedged' to improve a score, even if the properties of the adjusted ensemble forecast appear less desirable. It is appropriate to evaluate ensemble forecasts whose EDFs are interpreted as probability distributions for the observations using proper scoring rules.

Ferro [2014] introduced so-called 'fair scores', which are appropriate for ensemble forecasts that we know, or choose to interpret, as IID random samples from an underlying ensemble distribution. *Given* that an ensemble forecast is an IID sample, the expectations of fair scores are optimised when the ensemble and observation distributions are equal. More formally, fair scores are defined as follows.

Definition 2.5.2 *Let y denote an observation with distribution q.* Given *that the members of an ensemble forecast $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$ are independent and identically distributed realisations of an underlying ensemble distribution, p, the scoring rule $h^{\mathrm{fair}}(\boldsymbol{x}, y)$ is fair with respect to a class of distributions, Q if*

$$E_{q,q}\{h^{fair}(\boldsymbol{x}, y)\} \leq E_{p,q}\{h^{fair}(\boldsymbol{x}, y)\} \text{ for all } p \text{ and } q \in Q, \tag{2.49}$$

where the expectations in the above equation are calculated with respect to both the ensemble forecast $\boldsymbol{x}$ and observation y. The scoring rule is strictly fair if its expectation is uniquely optimised when $p = q$.

Ferro [2014] showed that if ensemble forecasts are verified with proper, rather than fair scores, for example by treating the proportion of ensemble members that forecast the occurrence of an event as a probability forecast, then the optimal value of the proper score corresponds to ensemble distributions $p \neq q$, where q denotes the distribution for the observation, y. By definition 2.5.2, however, the optimal value of fair scores corresponds to $p = q$. Ferro showed that proper scores fail to account

for the bias inherent in using finite-member ensemble forecasts to infer probability forecasts, a bias that is removed by fair scores. For example, consider the CRPS for an ensemble forecast whose EDF is interpreted as a probability forecast distribution for y, given in equation (2.48). The second term on the right hand side of equation (2.48) is a biased estimator of the expectation $E(|x - x'|)$, where x and x' are IID random variables distributed according to the ensemble distribution, since we count M instances of the dependent case (when $x_m = x_n$). In order to assess the ensemble forecast under the interpretation that the ensemble members are an IID sample from the ensemble distribution, this bias should be removed. The fair analog of the CRPS (FCRPS) is

$$\text{FCRPS}(\boldsymbol{x}, y) = \frac{1}{M}\sum_{m=1}^{M}|y - x_m| - \frac{1}{2M(M-1)}\sum_{m=1}^{M}\sum_{n=1}^{M}|x_m - x_n|\,, \tag{2.50}$$

where the denominator $M(M-1)$ of the second term removes the aforementioned bias. Fair scores should therefore be used for ensemble forecasts that we know, or choose to interpret as IID samples from an ensemble distribution [Ferro, 2014]. However, unlike proper scores for probability forecasts, it is possible to attain improved fair scores by choosing non-random ensemble members. This is because the expectation of the fair score with respect to q, the probability distribution for the verifying observation, y, is a function of the ensemble forecast $\boldsymbol{x}$ only, and so it is possible to explicitly calculate the ensemble members that optimise the fair score, in the same manner as the example provided by Bröcker [2012]. Therefore, fair scores *do not* elicit IID samples — rather, they are applicable for ensembles whose members are interpreted as IID samples.

Ferro [2014] also discussed fair scores for ensemble forecasts whose members are exchangeable, but not independent. In general, the existence of fair scores for such ensembles depends on the nature of the member dependence. Even if fair scores do exist, situations in which the member dependence is known exactly are rare, and so the dependence structure must be estimated. In this thesis, therefore, we interpret ensemble forecasts as either probability forecast distributions for y (using their EDF) or as IID samples from an underlying ensemble distribution, in keeping with the discussion in section 2.2.4.

2.5.5 The decomposition of proper scoring rules

In section 2.3 we introduced the notion of calibration for both probability and ensemble forecasts. As noted in that section, two desirable properties of probability forecasts are reliability and resolution, where reliability pertains to the 'accuracy' of the probability forecasts, and resolution refers to the ability of the probabilistic

forecasting system to provide useful information to the forecast user. An appealing result [Bröcker, 2009] is that proper scores can be decomposed in to components that quantify the reliability and resolution of a probabilistic forecasting system. In this subsection, we provide an empirical estimate of the decomposition of the Brier score, which was given by Murphy [1973]. The decomposition provides the forecaster with a measure of the reliability and resolution of the probabilistic forecasting system for forecasts of binary events $\{y \leq q\}$, where y is the verifying observation and q is a threshold of interest. Hersbach [2000] provided a decomposition for the continuous ranked probability score, although we do not consider its use in this thesis.

To calculate the decomposition of the Brier score we proceed as follows. As for reliability diagrams (see section 2.5.2.2), we divide the interval $(0, 1)$ in to K equally spaced bins. In this thesis we choose $K = 20$, so that the bins are given by the intervals $(0, 0.05], (0.05, 0.1], \ldots, (0.95, 1)$. Again let I_k denote the set of indices t of forecasts in bin k, for $k = 1, 2, \ldots, K$. We calculate the arithmetic mean of the probability forecasts in each bin, that is

$$\bar{p}_k = \frac{1}{N_{I_k}} \sum_{t \in I_k} z_t,$$

for all k, where N_{I_k} denotes the cardinality of the set of indices I_k. Let $\bar{z}_k, k = 1, 2, \ldots, K$, denote the mean of the verifying binary observations in bin k, that is

$$\bar{z}_k = \frac{1}{N_{I_k}} \sum_{t \in I_k} z_t,$$

where (as for the reliability diagrams) $z_t = \mathbf{I}(y_t \leq q)$. Also let $\bar{z} = T^{-1} \sum_{t=1}^{T} z_t$ denote the mean of the binary observations. The reliability, resolution and uncertainty components of the Brier score decomposition given by Murphy [1973] are

$$\text{REL} = \sum_{k=1}^{K} \frac{N_{I_k}}{T} \left(\bar{p}_k - \bar{z}_k\right)^2, \tag{2.51}$$

$$\text{RES} = \sum_{k=1}^{K} \frac{N_{I_k}}{T} \left(\bar{z}_k - \bar{z}\right)^2, \tag{2.52}$$

$$\text{UNC} = \bar{z}(1 - \bar{z}). \tag{2.53}$$

Equation (2.51) provides a measure of the reliability of a probabilistic forecasting system in issuing probability forecasts of binary observations. The reliability component is related to the reliability diagram (see section 2.5.2.2) — it is a weighted average of the squared distance of the reliability curve from the diagonal line, where the weights correspond to the relative frequency of probability forecasts in each of the K bins. Equation (2.52) provides a measure of the forecast resolution — the

equation can be viewed as the sample variance of the quantities $\bar{z}_k$ for $k = 1, 2, \ldots, K$ centred on the mean of the binary observations, $\bar{z}$. This is in keeping with the explanation provided in section 2.3, where we commented that the forecast resolution pertains to the variability of the observations, conditional on the forecast distributions. The verifying observations for a probabilistic forecasting system with no resolution would satisfy $\bar{z}_k = \bar{z}$ for all k, in which case equation (2.52) returns a resolution score of 0. With these comments in mind, it follows that small reliability and large resolution scores are preferred.

For the pooled forecasts — that is, where each of the T forecasts is reassigned to the probability $\bar{p}_k$ of the bin within which it falls, the Brier score [Brier, 1950] is given by

$$\text{Brier} = \text{REL} - \text{RES} + \text{UNC}. \tag{2.54}$$

To see this, observe that

$$\begin{aligned}
\text{REL} - \text{RES} + \text{UNC} &= \sum_{k=1}^{K} \frac{N_{I_k}}{T} \left\{ (\bar{p}_k - \bar{z}_k)^2 - (\bar{z}_k - \bar{z})^2 \right\} + \bar{z}(1 - \bar{z}) \\
&= \sum_{k=1}^{K} \frac{N_{I_k}}{T} \left\{ \bar{p}_k^2 - 2\bar{p}_k \bar{z}_k + 2\bar{z}_k \bar{z} - \bar{z}^2 \right\} + \bar{z}(1 - \bar{z}) \\
&= \sum_{k=1}^{K} \frac{N_{I_k}}{T} \left\{ \bar{p}_k^2 - 2\bar{p}_k \bar{z}_k \right\} + \bar{z}, \qquad (2.55)
\end{aligned}$$

where equation (2.55) follows from the result

$$\sum_{k=1}^{K} \frac{N_{I_k}}{T} \bar{z}_k = \bar{z}.$$

Now let r_t temporarily denote the pooled forecast on forecast occasion t, that is $r_t = \bar{p}_k$, depending on which of the K bins the original forecast, p_t, falls. Now observe that $\sum_{t \in I_k} r_t^2 = N_{I_k} \bar{p}_k^2$, since $r_t^2 = \bar{p}_k^2$ for all $t \in I_k$, for all k. It therefore follows that

$$\frac{1}{T} \sum_{t=1}^{T} r_t^2 = \sum_{k=1}^{K} \frac{N_{I_k}}{T} \bar{p}_k^2.$$

Similarly we have

$$\begin{aligned}
\sum_{t \in I_k} r_t z_t &= \sum_{t \in I_k} \bar{p}_k z_t \\
&= \bar{p}_k \sum_{t \in I_k} z_t \\
&= N_{I_k} \bar{p}_k \bar{z}_k,
\end{aligned}$$

since $\sum_{t \in I_k} z_t = N_{I_k} \bar{z}_k$ (by definition of $\bar{z}_k$). It therefore follows that

$$\frac{1}{T} \sum_{t=1}^{T} r_t z_t = \sum_{k=1}^{K} \frac{N_{I_k}}{T} \bar{p}_k \bar{z}_k.$$

Finally, since the observations z_t take the values either 0 or 1, we have that

$$\begin{aligned} \bar{z} &= \frac{1}{T} \sum_{t=1}^{T} z_t \\ &= \frac{1}{T} \sum_{t=1}^{T} z_t^2. \end{aligned}$$

Substituting these results in to equation (2.55) gives

$$\text{REL} - \text{RES} + \text{UNC} = \frac{1}{T} \sum_{t=1}^{T} (r_t - z_t)^2, \tag{2.56}$$

which is the Brier score (see equation (2.43)) for the pooled probability forecasts, as claimed.

2.6 Data

2.6.1 The Lorenz 1996 system

The Lorenz 1996 system [Lorenz, 1996], hereafter referred to as L'96, acts as a surrogate, or 'toy model' of the atmosphere. Studies that utilise the L'96 system include Lorenz [1996]; Roulston and Smith [2003]; Wilks [2005, 2006a]; Williams et al. [2014]. The system comprises both slow and fast moving variables, representing large scale atmospheric features such as the Atlantic jet stream, and small scale phenomena, such as localised precipitation events, that are often inadequately resolved by NWP models. The governing equations of the system are

$$\frac{dX_j}{dt} = X_{j-1}(X_{j+1} - X_{j-2}) - X_j + F - \frac{HC}{B} \sum_{k=1}^{K} Y_{j,k}, \tag{2.57}$$

for $j = 1, 2, \ldots, J$ and

$$\frac{dY_{j,k}}{dt} = CBY_{j,k+1}(Y_{j,k-1} - Y_{j,k+2}) - CY_{j,k} + \frac{HC}{B} X_j, \tag{2.58}$$

for $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, K$. In this thesis we set $J = 8$ and $K = 32$, to give a system of eight slow-moving X variables and 256 fast-moving Y variables.

As in Wilks [2006a] and Williams et al. [2014], we set the constant parameters to be $H = 1, B = 10$ and $C = 10$, and the 'forcing parameter' to be $F = 20$. The boundary conditions are cyclical, so that $X_{-1} = X_{J-1}$ and $Y_{-1} = Y_{K-1}$. Observe that the state of each X_j is determined in part by the summation of the K associated Y variables, and similarly the state of the Y variables is affected by the associated X_j.

We use the L'96 system to simulate data to which we apply ensemble post-processing methods. We use the variable X_1 as the predictand for which we make forecasts — in the following text, therefore, the observations y are realisations of the variable X_1 of the 'true' system governed by equations (2.57) and (2.58). To simulate an imperfect NWP model for the $X_j, j = 1, 2, \ldots, 8$, we replace the final term in equation (2.57) by a quartic polynomial in X_j, denoted $U(X_j)$ as follows:

$$U(X_j) = 0.262 - 1.262X_j + 0.004608X_j^2 + 0.007496X_j^3 - 0.0003226X_j^4. \quad (2.59)$$

In other words, we pretend that the dynamics of the Y variables given in equation (2.58) are unknown, and estimate their effect on the evolution of the X_j variables by $U(X_j)$. The imperfect NWP model is therefore given by

$$\frac{dX_j^*}{dt} = X_{j-1}^*(X_{j+1}^* - X_{j-2}^*) - X_j^* + F + U(X_j^*), \quad (2.60)$$

for $j = 1, 2, \ldots, J$, where the notation X_j^* denotes the approximation of X_j by X_j^*. The function $U(X_j)$ was determined by regressing a quartic polynomial in X_j on the true state of the system. Both the true system (equations (2.57) and (2.58)) and the imperfect NWP model (equation (2.60)) are integrated forwards in time using a simple forward Euler scheme, with a time step of 10^{-4} time units.

In order to represent an ensemble forecasting system, initial conditions for each X_j^* were randomly sampled from a Gaussian distribution centred at the 'true' X_j, $N(X_j, 0.1^2)$. Ensemble forecasts were then constructed by integrating each of these initial conditions forwards in time, using the imperfect model given by equation (2.60). The standard deviation of the initial conditions yields ensemble forecasts that retain a mixture of forecast skill while providing reasonable spread at the lead times considered in subsequent chapters of this thesis. Because the initial conditions are IID random draws, the resulting ensemble forecasts can be justifiably interpreted as IID draws from underlying ensemble distributions.

Ensemble forecasts of varying size were produced, as will be reported when appropriate. Forecasts and observations were stored at lead times denoted $t = 1, 2, \ldots, 5$, where each lead time corresponds to 0.2 time units of the system's evolution, or 2000 iterates of the forward Euler scheme. Two separate datasets of ensemble forecasts

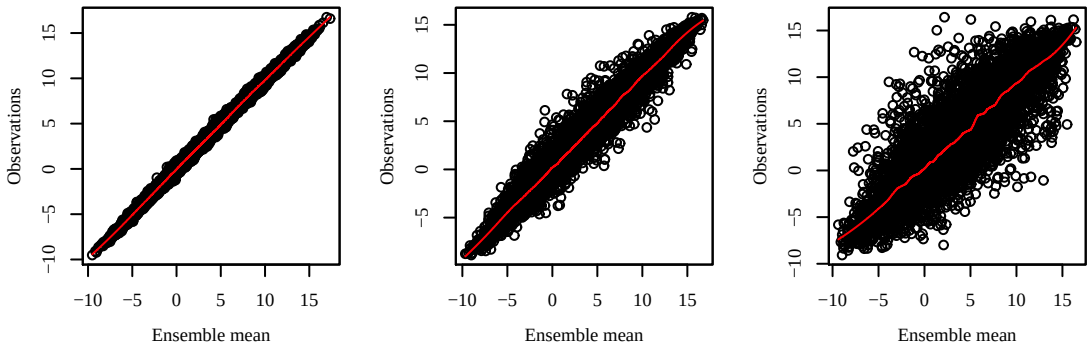

Figure 2.1 Observations y as a function of the ensemble mean $\bar{x}$ for forecast lead times $1, 3$ and 5. A nonparametric estimate to the observations is shown in red.

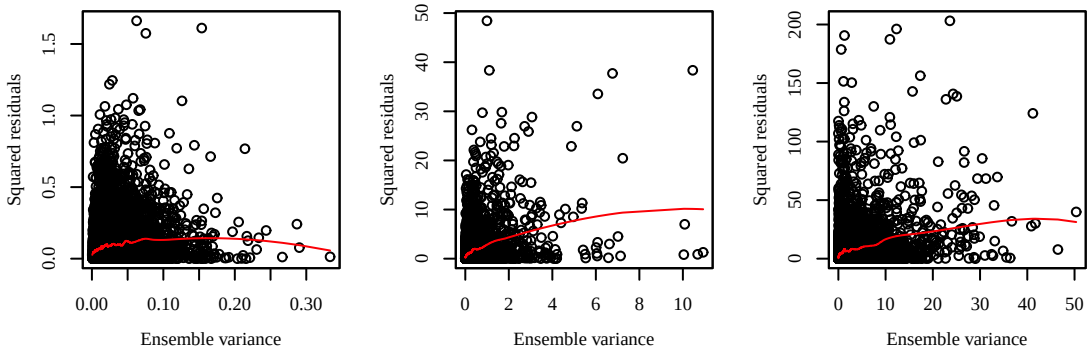

Figure 2.2 Plots of the squared residuals r^2 as a function of the ensemble variance s^2 for forecast lead times $1, 3$ and 5. A nonparametric estimate to the expectation of the squared residuals is shown in red.

and observations were produced. A training dataset of size $N = 500\,000$ is used for parameter estimation in Williams et al. [2014] and in other examples provided later in this thesis. The forecasts and observations are initialised every 0.5 time units, and are therefore temporally correlated in keeping with real world scenarios. The second dataset, which was used for forecast verification in Williams et al. [2014], is of size $T = 190\,000$, where each of the T forecasts and observations are initialised at time points separated by 50 time units, and are therefore effectively independent of other forecasts and observations in the dataset.

Figure 2.1 shows the observations as a function of the ensemble mean, for various forecast lead times, while figure 2.2 shows the corresponding plots of the squared residuals, $r^2 = (\bar{x} - y)^2$, as a function of the ensemble variance, s^2. We show only a random sample of the data from the second of the described datasets — it was necessary to sample from the dataset due to its size. We also show a nonparametric estimate to the expectation of the dependent variable (either the observations or residuals), using the Loess function implemented in the R language that was described in section 2.5.2.1.

Figure 2.1 shows that the ensemble mean is a 'good' predictor of the verifying ob-

servations at all lead times shown, although the relationship is not exact — the observations do not lie on the diagonal line, particularly for lead time 5. More interestingly, figure 2.2 indicates a fairly weak, but nonetheless significant relationship between the ensemble variance and the squared residuals, particularly for lead times 3 and 5 — the sample correlation coefficients for lead times $1, 3$ and 5 are $0.165, 0.328$ and 0.287. Figure 2.2 indicates the presence of a spread-skill relationship that might be exploited in modelling the variance of probability forecast distributions as a function of the ensemble variance.

2.6.2 The GEFS reforecast project

We also use hindcast data from the GEFS reforecasts project hosted by the Earth Systems Research Laboratory at the National Oceanic and Atmospheric Association, USA (`http://www.esrl.noaa.gov/psd/forecasts/reforecast2/`). A detailed description of the dataset is provided in Hamill et al. [2013]. In brief, the project comprises retrospective forecasts, or hindcasts, from 1 December 1984 – present, using an ensemble forecasting system that is similar to the operational Global Ensemble Prediction System (GEFS) operated by the National Centre for Environmental Prediction (NCEP). We use the 0 hour control forecast for the verifying observations, which corresponds to the reanalyses (the retrospective analyses) at the verification time. Ensemble forecasts of eleven members are available on a global grid. In our investigations (see chapters 4 and 5), however, we do not use the 'control forecast', the forecast that is initialised at the reanalysis. The remaining ten members are generated by five pairs of Bred vectors (see section 2.2.2 for a brief description). In this thesis we consider two datasets that were both taken from the available archive. In chapter 4 we consider ensemble forecasts over a 17×18 grid that approximately covers the United Kingdom. Forecasts were collected for the period 1 December 2011 – 14 December 2014 inclusive. In chapter 5, we use a longer time series of forecasts and observations at a single gridpoint, located near to New York City, USA (40 degrees North, 70 degrees West).

As for the Lorenz 1996 system described in the previous subsection, diagnostic plots (not shown) indicate a strong linear relationship between the ensemble mean and verifying observations. For the dataset of forecasts and observations located to New York, the sample correlation coefficient between the squared residuals and the ensemble variance is 0.176. While this does not indicate a 'strong' relationship, our investigations (see later chapters) demonstrated the value of including the ensemble variance as a covariate in the statistical models used for ensemble post-processing.

3 A comparison of post-processing methods for extreme events

3.1 Introduction

In this chapter we present an extensive comparison study of the probabilistic forecasting skill of several ensemble post-processing methods. Our interest is in probability forecasts of the form $\Pr(y \leq q)$, where y is a verifying observation, and q is a threshold of interest. We pay particular interest to extreme thresholds, for which occurrences of the binary event $\{y \leq q\}$ are rare. As highlighted in the previous chapter (see section 2.4), considerable and valuable efforts have been made towards the development of ensemble post-processing methods that issue calibrated probability forecasts, for a variety of meteorological variables. However, a surprisingly small amount of the literature has evaluated the probabilistic forecasting skill of these methods in issuing probability forecasts of rare, or extreme events. Such events are often those of most interest to forecast users and the general public — for example, it is well known that meteorological extremes such as heat waves and flooding have profound humanitarian and economic repercussions. Therefore, the ability to produce calibrated probability forecasts of extreme events is presumably of interest to a broad cross-section of users. The study presented in this chapter has been published in the literature [Williams et al., 2014].

Our new work was motivated by the article of Wilks [2006a], which provided a comparison of the probabilistic forecasting skill of several ensemble post-processing methods in the Lorenz 1996 system (see section 2.6.1 and Lorenz [1996]). Wilks concluded that Logistic Regression (LR), nonhomogeneous Gaussian regression (NGR), Bayesian model averaging (BMA) and best member dressing (BMD) were the most promising post-processing methods. In this chapter we also use data from the Lorenz 1996 system. We introduce extensions to the probability forecast distributions that are specified by the BMA and BMD post-processing methods, by introducing so-called 'ensemble adjustment schemes' that provide more advanced corrections to the biases in location and dispersion of the ensemble forecasts than were permitted in the founding papers for those methods (see [Wang and Bishop, 2005] and Raftery

et al. [2005]). We investigate the extent to which our more sophisticated bias corrections for the BMA and BMD statistical models improve the skill of probability forecasts, for both common and rare events. We also investigate the effect of other features of the ensemble post-processing methods on forecast skill, such as the choice of objective function, and the size of training sample, that are used for estimating the model parameters.

The remainder of this chapter is organised as follows. In section 3.2 we review the article by Wilks [2006a]. In section 3.3 we introduce our hierarchy of so-called 'ensemble adjustment schemes', that facilitate bias corrections to the location and dispersion of ensemble forecasts, and use these adjustment schemes to extend the statistical models specified by the BMA and BMD ensemble post-processing methods. We show how these model extensions allow for an extension of the comparison study of Wilks [2006a]. We also give details of the objective functions (see section 2.4.7) that are used for parameter estimation for the BMA, BMD, NGR and LR models, and provide related comments for the method of rank histogram recalibration (RHR, Hamill and Colucci [1997, 1998]). Section 3.4 provides details of the setup of our comparison study, and the verification measures that are used to assess the skill of probability forecasts. We present our results in section 3.5, and finish the chapter with our conclusions and a discussion in section 3.6.

3.2 A review of Wilks [2006a]

In this section we provide an overview of the article by Wilks [2006a]. Wilks used the Lorenz 1996 system (see section 2.6.1) as a surrogate, or 'toy' model of the atmosphere. The slow-varying X variables (see section 2.6.1 for further details) were used as the 'true state', for which probability forecasts of the form $\Pr(y \leq q)$, where y is the verifying observation and q is a threshold of interest, were issued using one of several ensemble post-processing methods. As we also describe in section 2.6.1, ensemble forecasts for the 'true state' were issued using an approximation to the system of slow-varying X variables, that was representative of an imperfect numerical weather prediction (NWP) model. The thresholds used in the study were the lower decile, the lower tercile, the median, the upper tercile and the upper decile of the climatology of the X variables in the Lorenz 1996 system. Wilks compared the skill of eight post-processing methods (listed below). He also investigated the effect of the training sample size, N, and the ensemble size, M, on the resulting out-of-sample probability forecasts, for forecast lead times of $1, 2, \ldots, 5$. The time units of the lead times were the evolution of the Lorenz system through $1/h$ iterates of the numerical scheme used to approximate the evolution of the system, where h denotes the step size (see section 2.6.1 for details). Two separate datasets were used

for the estimation of model parameters and forecast verification, in a similar vein to the description of the data used in this thesis (see section 2.6.1). The ensemble post-processing methods under consideration were

- Direct model output (DMO, see section 2.4.2.1)
- Rank histogram recalibration (RHR, see section 2.4.2.2)
- Multiple implementations of deterministic MOS equations (MIDME Erickson [1996])
- Best member dressing (BMD, see section 2.4.3.1)
- Bayesian model averaging (BMA, see section 2.4.3.2)
- Forecast assimilation (FA, see section 2.4.5)
- Logistic regression (LR, see section 2.4.4.3)
- Nonhomogeneous Gaussian regression (NGR, see section 2.4.4.2).

The approach of multiple implementations of deterministic model output statistics equations involves applying corrections derived for deterministic forecasts to each ensemble member, resulting in supposedly recalibrated ensemble forecasts. However, as explained by Wilks [2006a, section 2.1.3], this approach reduces, rather than increases, the ensemble dispersion as forecast lead times increase, and in the limit places all mass at the climatological mean. This behaviour is the opposite of what should be expected in a coherent ensemble forecasting system, namely that the ensemble dispersion increases with forecast lead time in keeping with the increasing forecast uncertainty.

Wilks implemented the post-processing methods as specified by the authors in their founding papers. The setup of the ensemble forecasting system involved a 'control member', with the remaining ensemble members considered as exchangeable. For the BMA and BMD forecasts, therefore, Wilks estimated two weights — a weight w_1 for the control forecast, and w_2 that was appropriate for weighting the remaining $M-1$ ensemble members. Parameter estimation was carried out on a single training sample of size N. The data in the training sample exhibited temporal correlation, similar to that of the first dataset described in section 2.6.1 that we use for parameter estimation in our own study. Out-of-sample probability forecasts were then issued for a second, test dataset, for which the data were temporally uncorrelated.

When necessary, the parameters of the statistical models specified by the ensemble post-processing methods were estimated using the objective function that was recommended by the method's authors. In particular, parameters for the logistic

regression (LR) and Bayesian model averaging (BMA) models were estimated by maximisation of the log-likelihood function (equivalent to minimising the negative log-likelihood function given in section 2.4.7), and the NGR model parameters were estimated with the continuous ranked probability score (CRPS, see equation (2.37) on page 48), as suggested in the founding article by [Gneiting et al., 2005]. The variance of the Gaussian-distributed dressing kernels that form the mixture components of the best member dressing (BMD) probability forecast distributions were estimated explicitly from the ensemble forecasts and observations in the training sample, using equation (2.9) [Wang and Bishop, 2005]. Therefore, the BMD statistical model did not require the use of an objective function in the study by Wilks [2006a]. Similarly, the rank histogram recalibration model (RHR, see section 2.4.2.2) did not require the use of an objective function. The weights w_m, for $m = 1, 2, \ldots, M + 1$, that are used in the equations for probability forecasts of the form $\Pr(y \leq q)$ (see equation (2.8) on page 33) were calculated as the proportion of observations that fall in the $M + 1$ rank histogram bins, where the rank histogram was calculated over the training sample (see the description in section 2.4.2.2 for details). The other post-processing methods considered in Wilks [2006a] were, to use the author's own words, 'early, ad-hoc approaches', that did not require the use of parameter estimation. Wilks used the DMO forecasts (see section 2.4.2.1) as a baseline against which the skill of probability forecasts given by the other post-processing methods was compared.

The skill of the probability forecasts issued by the various statistical models was assessed using the rank probability skill score (RPSS), which is proportional to the sum of the Brier scores (see section 2.5.3.2) evaluated at the five thresholds under consideration. The RPSS therefore provides an assessment of the overall skill of a probabilistic forecasting system. To gain further understanding of the performance of the post-processing methods at individual thresholds, the Brier scores and their decomposition (see section 2.5.5), and reliability diagrams (see section 2.5.2.2) were also used.

The key findings of the paper were the following. Under all considered measures of forecast skill, the more 'sophisticated' post-processing methods (LR, NGR, BMA, BMD and FA) improved upon the 'ad-hoc' methods (DMO, MIDME, and RHR), although RHR was competitive for short forecast lead times and large training samples. The relative skill of the 'sophisticated' methods was found to be dependent on the forecast lead time, the training sample size, and the threshold q for which the probability forecasts $\Pr(y \leq q)$ were issued. For short forecast lead times, the NGR forecasts were most skilful, although the BMA and BMD methods also exhibited significant improvements in forecast skill compared with the DMO forecasts (section 2.4.2.1). In contrast, the LR forecasts were the most skilful for longer forecast

lead times and large training samples. However, the skill of the LR forecasts was found to degrade rapidly with training sample size, to the extent that the DMO forecasts (which do not make use of training data) were equally skilful for the smallest training sample size considered, $N = 50$. The skill of the LR forecasts was also found to degrade with threshold extremity — the forecast skill at the most extreme thresholds considered, the upper and lower deciles, was poor relative to the skill of forecasts for more common events. This feature is exacerbated for smaller training samples, and was said by Wilks [2006a] to be due to poor parameter estimates. Overall, the BMA forecasts were found to be slightly less skilful than their BMD counterparts, particularly for the probability forecasts at the upper and lower deciles, and to a lesser extent for the upper and lower terciles. The reliability diagrams indicated that the BMA forecasts are overdispersed, i.e. the BMA probability forecast distributions overcompensated for the underdispersion in the raw ensemble forecasts. At the longer forecast lead times considered, only the LR and NGR post-processing methods were found to produce more skilful probability forecasts than the DMO method. This was achieved by restricting the probability forecasts to a narrow range around the corresponding climatalogical frequencies, so that the forecasts were reliable, although had little skill in terms of forecast resolution.

3.3 Extending the study of Wilks [2006a]

3.3.1 The aims of our study

The study [Wilks, 2006a] provided an extensive comparison of the skill of probability forecasts issued by several popular ensemble post-processing methods. The primary focus of the study was to suggest the methods that are most promising for probabilistic forecasting and, aside from some fairly general comments, detailed reasoning as to the relative performance of the post-processing methods was not provided. We now turn to a description of our comparison study, which builds on that of Wilks [2006a].

Firstly, as mentioned in section 3.1, in this chapter we primarily focus on the BMA, BMD and NGR post-processing methods. Attempts to extend the rank histogram recalibration (RHR) method proved unsuccessful, as we describe in section 3.3.3.4. Unlike Wilks [2006a], we did not consider the method of forecast assimilation (FA, Stephenson et al. [2005]). As described in section 2.4.5, this Bayesian approach to ensemble post-processing requires a preliminary step of specifying prior distributions for the model parameters from training data, a step that is associated with the data assimilation process. In this chapter, and indeed for the remainder of this thesis, we

use ensemble forecasts whose members can justifiably be considered exchangeable (see definition 2.2.1 in section 2.2.4).

Wilks implemented the post-processing methods as specified by their authors in the literature. As can be seen from the description of various methods provided in section 2.4, the statistical post-processing models that are used to issue probability forecasts differ in several of their 'features' — the family of probability distributions (such as Gaussian distributions) that are used, in how predictor variables (typically ensemble statistics such as the ensemble mean and variance) are related to properties of these distributions (such as their expectations and variances), and in how the parameters describing these relationships are estimated. A key aim of this chapter is to establish those features of the statistical post-processing models that are the most influential factors on probabilistic forecasting skill. We pay particular attention to the effect of different specifications of the expectation and variance of the probability forecast distributions on the skill of the out-of-sample probability forecasts. More specifically, we achieve this insight by proposing alternative specifications for the expectations and variances of the BMA, BMD and NGR statistical models. In the following subsection, we introduce a hierarchy of so-called 'ensemble adjustment schemes', that enable the user to specify the expectation and variance of the BMA and BMD forecast distributions in an analogous manner to the linear functions of the ensemble mean and variance that are used for the expectation and variance of the NGR forecast distributions (see equations (2.19) and (2.20) on page 38). This is achieved by allowing the bias in ensemble location to be corrected with a linear function of the ensemble mean, and the bias in ensemble dispersion to be corrected by rescaling the ensemble forecast, where the degree of rescaling is proportional to the ensemble variance.

Furthermore, we also compare the probabilistic forecasting skill of ensemble post-processing methods for parameter estimates that are obtained by minimising the negative log-likelihood (NLL) and the continuous ranked probability score (CRPS). Previously, with the exception of NGR [Gneiting et al., 2005], parameter estimation for those post-processing methods for which it is necessary has been conducted in the likelihood framework.

A second key aim of this chapter is to report on the skill of ensemble post-processing methods in issuing probability forecasts $\Pr(y \leq q)$ for extreme, as well as common thresholds q, for which occurrences of the binary event $\{y \leq q\}$ are rare. In doing so, we can determine whether certain features of the statistical models used for ensemble post-processing are particularly important for issuing skilful probability forecasts of rare events.

To motivate the idea of extending the BMA and BMD statistical models by allowing

alternative forms for their expectation and variance (as mentioned above), it helps to first consider the expectation of the probability forecast distributions issued by those methods, and the expectation of the NGR forecast distributions. For NGR, the expectation, μ^{NGR} on a general forecast occasion is a linear function of the ensemble mean (see equation (2.19)), that is

$$\mu^{\text{NGR}} = a + b\bar{x},$$

where $\bar{x} = M^{-1}\sum_{m=1}^{M} x_m$ is the sample mean of the M-member ensemble forecast $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$. On the other hand, for the exchangeable ensemble members considered in this thesis the expectation of the corresponding BMA forecast distribution is simply the mean of the (possibly de-biased) ensemble forecasts, that is

$$\mu^{\text{BMA}} = \bar{x}.$$

This follows immediately from the specification of the BMA forecast distribution for exchangeable ensemble members, which is an equally weighted mixture of M Gaussian-distributed dressing kernels, each with expectation x_m, for $m = 1, 2, \ldots, M$, and where we recall that the weights are constrained to sum to 1 (see section 2.4.3.2). If ensemble members x_m, for $m = 1, 2, \ldots, M$ are indeed first corrected for a constant bias, say a, then the expectation of the BMA forecast distribution is simply $\mu^{\text{BMA}} = a + \bar{x}$. The same is also true for the expectation of the probability forecast distributions issued by the flavour of the BMD post-processing method considered in this work [Wang and Bishop, 2005]. Clearly, therefore, the NGR forecast distributions have more flexibility in the specification of their expectation than the corresponding BMA and BMD forecast distributions. The NGR post-processing method assumes that a linear function of the ensemble mean is appropriate for modelling the verifying observations, whereas the BMA and BMD methods assume that the (possibly de-biased) ensemble mean is an adequate predictor of the observations or, in other words, that the bias in ensemble location is independent of the forecast values (as summarised by the ensemble mean).

Similar comments apply for the variance of the BMA, BMD and NGR probability forecast distributions. Recall from equation (2.20) that the variance of a general NGR forecast distribution is

$$\sigma^{\text{NGR}^2} = c + ds^2,$$

where $s^2 = (M-1)^{-1}\sum_{m=1}^{M}(x_m - \bar{x})^2$ is the sample variance of the ensemble forecast $\boldsymbol{x}$. It can easily be shown that the variance of the corresponding BMA forecast distribution is given by

$$\sigma^{\text{BMA}^2} = c^{\text{kernel}} + (M-1)s^2/M, \tag{3.1}$$

where c^{kernel} denotes the variance of the Gaussian-distributed dressing kernels. The same expression also holds for the variance of BMD forecast distributions. Therefore, analogous comments to the above for the expectation of the BMA and BMD forecast distributions apply for the variance — the NGR method allows for greater flexibility in the specification of its forecast variance than the BMA and BMD post-processing methods, in the implementations specified by Raftery et al. [2005] and Wang and Bishop [2005], respectively.

As mentioned earlier in this subsection, two objective functions (the negative log-likelihood (NLL) and the continuous ranked probability score (CRPS)) have been used in the literature for estimating the parameters of the statistical post-processing models. Gneiting et al. [2005] claimed that the CRPS is preferable for NGR forecasts, while other authors have only considered likelihood-based parameter estimation. In this work, therefore, we estimate the parameters of BMA, BMD and NGR models using both the NLL and CRPS objective functions, and report on any differences in the skill of the resulting probability forecasts. The specific forms of the NLL and CRPS functions for the three models, as well as comments on logistic regression (LR) and rank histogram recalibration (RHR) models, are provided in section 3.3.3.

3.3.2 A hierarchy of models for ensemble post-processing methods

We now introduce our hierarchy of 'ensemble adjustment schemes'. As mentioned in the previous subsection, these schemes allow us to specify more flexible parametric functions for the expectation and variance of the BMA and BMD probability forecast distributions, and thus facilitate a fair and coherent comparison with NGR forecast distributions. Each ensemble adjustment scheme specifies a bias correction for the ensemble forecast — a constant correction for the CC scheme, a linear correction (as a linear function of the ensemble mean) for the LC scheme, and a linear correction with the additional possibility of ensemble rescaling (the LCR scheme). These schemes are parametric, and the parameters must be estimated by optimisation of an objective function. In the following, we define the ensemble adjustment schemes for an ensemble forecast $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$. Specifically, the three adjustment schemes are defined as follows.

Constant correction (CC) This scheme removes constant bias from the ensemble members x_{im}, for $m = 1, 2, \ldots, M$ to give ensemble members $\hat{x}_{im} = a + x_{im}$, where a is a constant parameter that is to be estimated. This ensemble adjustment scheme implicitly assumes that the bias of ensemble members is constant regardless

of their forecast values. For CC, therefore, the expectation and variance of the ith BMA and BMD forecast distributions are $a + \bar{x}_i$ and $c^{\text{kernel}} + (M-1)s_i^2/M$, respectively. The CC ensemble adjustment scheme is incorporated in to the BMA/BMD post-processing models by substituting the adjusted ensemble members $\hat{x}_{im}$ given above in to equations (2.10) and (2.11) for the PDF (CDF) of the BMD forecast distributions, and similarly for the BMA forecast distributions. This results in a revised statistical model — the parameter a that is used for the bias correction must then be estimated by optimising an objective function.

Linear correction (LC) This ensemble adjustment scheme incorporates a bias correction to the ensemble members that is linear in the ensemble mean, that is

$$\hat{x}_{im} = a + b\bar{x}_i + x_{im} - \bar{x}_i,$$

where here both a and b are parameters that are to be estimated. The bias correction of the ensemble members therefore differs according to the value of the ensemble mean. Under the LC scheme the forecast mean and variance of the ith BMA and BMD mixture distributions are $a + b\bar{x}_i$ and $c^{\text{kernel}} + (M-1)s_i^2/M$, respectively.

Linear correction with rescaling (LCR) This scheme incorporates the LC linear bias correction, in addition to a rescaling of the ensemble members by replacing x_{im} by

$$\hat{x}_{im} = a + b\bar{x}_i + \sqrt{d}(x_{im} - \bar{x}_i),$$

where a, b and d are again parameters that are to be estimated. The LCR scheme not only allows for bias corrections that depend linearly on the ensemble mean, but for the idea that the skill of the BMA and BMD probability forecasts may be improved if the ensemble forecasts are rescaled. The rescaling is achieved through the final term in the above expression for $\hat{x}_{im}$ — ensemble members are moved either further from (if $d > 1$) or nearer to (if $d < 1$) the ensemble mean. The expectation and variance of the ith BMA and BMD forecast distributions are then $a + b\bar{x}_i$ and $c^{\text{kernel}} + d \cdot (M-1)s_i^2/M$, respectively.

Some brief notes on the three ensemble adjustment schemes now follow. To our knowledge, the CC scheme is typical of operational practice. However, the bias correction is usually accomplished by subtracting the mean error of the ensemble members over the training period from each ensemble member, rather than estimating the model parameter, a, by means of an objective function. For BMA, the dressing kernel variance is subsequently estimated by optimising the likelihood function over the training sample of historical ensemble forecasts and verifying observations, and is calculated explicitly using equation (2.9) [Wang and Bishop, 2005]

	Expectation (μ_i^{NGR})	Variance ($\sigma_i^{\text{NGR}^2}$)
CC	$a + \bar{x}_i$	$c + (M-1)s_i^2/M$
LC	$a + b\bar{x}_i$	$c + (M-1)s_i^2/M$
LCR	$a + b\bar{x}_i$	$c + d \cdot (M-1)s_i^2/M$

Table 3.1 The parametric form of the expectation and variance, μ_i^{NGR} and $\sigma_i^{\text{NGR}^2}$, of the ith NGR forecast distribution under the CC, LC and LCR ensemble adjustment schemes.

for BMD. We write the LC ensemble adjustment scheme as $\hat{x}_{im} = a + b\bar{x}_i + x_{im} - \bar{x}_i$ rather than the alternative $\hat{x}_{im} = a + bx_{im}$. The second expression yields the same mean for the adjusted ensemble, $a + b\bar{x}_i$, but has ensemble variance $\hat{s}_i^2 = b^2 s_i^2$. Our implementation of the LC scheme enables the forecaster to correct for bias in the ensemble location while leaving the ensemble variance unchanged. Finally, the LCR ensemble adjustment scheme enables the forecaster to exploit possible spread-skill relationships (see section 2.4.1) between the ensemble variance and the magnitude of the forecast errors of the ensemble mean, in the same vein as the NGR forecast variance discussed in the previous subsection, and given in equation (2.20). As we described for the CC adjustment scheme, the LC and LCR schemes are incorporated in to the BMA/BMD statistical models by using the adjusted ensemble members $\hat{x}_{im}$ in place of the raw ensemble members x_{im} in the original formulations of those methods described in sections 2.4.3.1 and 2.4.3.2. The parameters of these new models — a and b for LC, and a, b and c for LCR, must then be estimated by optimising an objective function.

In section 3.5 we compare the skill of probability forecasts given by the BMA and BMD forecast distributions for the CC, LC and LCR schemes. We also consider the equivalent hierarchy of NGR models, that is models in which the expectation and variance of the corresponding NGR forecast distributions, μ_i^{NGR} and $\sigma_i^{\text{NGR}^2}$, take the same form as the BMA and BMD forecast distributions as specified by the CC, LC and LCR schemes. In doing so, we are able to make a fair and coherent comparison of the NGR, BMA and BMD post-processing methods across the three ensemble adjustment schemes. Table 3.1 shows the form of μ_i^{NGR} and $\sigma_i^{\text{NGR}^2}$ for the NGR statistical model under the three ensemble adjustment schemes.

Note that in the above NGR models we have multiplied the ensemble variance, s_i^2, by the constant $(M-1)/M$, in order to conduct a fully fair comparison with the BMA and BMD forecast distributions, by ensuring the same parametric form for the forecast variance. For the LCR scheme, the constant $(M-1)/M$ is simply absorbed in to the parameter d. In addition, we consider a fourth parameterisation of the NGR model, denoted NGR^0, in which c is set to 0, such that $\sigma_i^{\text{NGR}^2} = (M-1)s_i^2/M$ for CC and LC, and $d \cdot (M-1)s_i^2/M$ for LCR. This allows us to investigate the importance of including a constant parameter, c, in the forecast variance.

Sadly, it is not possible to make such a coherent comparison of probability forecasts using the logistic regression (LR) post-processing method. As described in section 2.4.4.3, unlike the NGR, BMA and BMD forecast distributions, changes in the linear predictor, η_i, do not result in intuitive changes to the location and variance of the Logistic distribution. It was therefore decided to only consider the two LR models with $\eta_i = a + b\bar{x}_i$ and $\eta_i = a + b\bar{x}_i + ds_i^2$, as in Hamill et al. [2004] and Wilks [2006a]. In section 3.5 we report the skill of the resulting probability forecasts under the labels of LC and LCR, respectively, and do not show any results for LR under the CC scheme.

In theory, our ensemble adjustment schemes can also be applied to the rank histogram recalibration method. Rank histograms are calculated for the adjusted ensemble forecasts $\widehat{\boldsymbol{x}}_i$, where the $\hat{x}_{im}$ denote the members of the adjusted ensemble forecasts under either the CC, LC, or LCR adjustment scheme. Out-of-sample probability forecasts of the form $\Pr(y_t \leq q)$ are then issued using equation (2.8), by substituting the adjusted ensemble forecast, $\widehat{\boldsymbol{x}}_t$ in place of the raw forecast, $\boldsymbol{x}_t$. For a given ensemble adjustment scheme, the expectation of the resulting probability forecast distributions is as given above. However, the forecast variance is not equal to that of the BMA, BMD and NGR distributions. Rather, the forecast variance of the rank histogram distributions is a function of the weights $w_m, m = 1, 2, \ldots, M+1$, the distances between the (ordered) members of $\widehat{\boldsymbol{x}}$, as well as the choice of probability distribution that is used for extrapolation of quantities that are not bounded by the ensemble forecast. It therefore seems infeasible to obtain an equivalent parametric form for the RHR forecast variance as those specified for the BMA, BMD and NGR methods using the model hierarchy described earlier in this subsection. In any case, it turns out that parameter estimation for the RHR method is highly problematic (see section 3.3.3.4), and so this method is not pursued in our study.

As mentioned earlier in this subsection, in section 3.5 we compare the skill of probability forecasts issued by the NGR, BMA, and BMD forecasts for the hierarchy of models that specify the form of the expectation and variance of the forecast distributions of those methods, as well as forecasts issued by the LR method. The probability forecasts are 'out-of-sample'. Firstly, parameter estimates are obtained from training samples of a pre-specified size (see the next subsection). For example, the BMA and BMD post-processing methods under the LC ensemble adjustment scheme require estimates of the parameters a, b and c^{kernel}, denoted $\hat{a}, \hat{b}$ and $\hat{c}^{\text{kernel}}$. Probability forecast distributions for the out-of-sample, verifying observations, say y_t, where the subscript t indexes the forecasts and observations in the test dataset, are then issued by substituting the parameter estimates in to the BMA/BMD statistical models along with the (out-of-sample) ensemble forecast, $\boldsymbol{x}_t$.

3.3.3 Parameter estimation

As discussed in section 3.3.1, two objective functions (the likelihood and CRPS) have been employed for the post-processing methods that require parameter estimation. In order to establish whether the objective function is an influential factor on the forecast skill, we shall compare the skill of the various post-processing methods using both choices. We now proceed by giving the negative log-likelihood (NLL) and CRPS functions for the various post-processing methods under consideration. The objective functions are calculated from training samples of N forecasts and verifying observations. The parameter estimates are those that minimise the (multi-dimensional) objective function surface, and are obtained using the Nelder-Mead and BFGS algorithms described in section 2.4.7.

3.3.3.1 Parameter estimation for NGR

In the following, ϕ and Φ denote the PDF and CDF of a standardised Gaussian random variable. The NLL and CRPS functions for the NGR post-processing method are

$$\mathrm{NLL}^{\mathrm{NGR}} = \sum_{i=1}^{N} \left(\frac{\log 2\pi}{2} + \log \sigma_i + \frac{1}{2} z_i^2 \right), \tag{3.2}$$

and

$$\mathrm{CRPS}^{\mathrm{NGR}} = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \left[z_i \{ 2\Phi(z_i) - 1 \} + 2\phi(z_i) - \frac{1}{\sqrt{\pi}} \right], \tag{3.3}$$

where $z_i = (y_i - \mu_i)/\sigma_i$ is a standardised observation which, if the NGR model is correctly specified, follows a standard Gaussian distribution. As described earlier, the form of μ_i and σ_i depends on the scheme (CC, LC, LCR or NGR^0) that is of interest (see table 3.1).

3.3.3.2 Parameter estimation for BMA and BMD

Recall that, for exchangeable ensemble members (such as those considered in this study), the BMA and BMD forecast distributions on the ith forecast occasion are equally weighted mixtures of Gaussian distributions centred on the ensemble members, $\hat{x}_{im}$, where each component, or dressing kernel, has variance c^{kernel}. As described in section 3.3.2, the ensemble members $\hat{x}_{im}$ are parametric functions of the original, 'raw' ensemble forecasts, where the parametric form depends on the choice of the ensemble adjustment scheme. The NLL and CRPS functions given below therefore depend on c^{kernel}, as well as one, two or three additional parameters for the CC, LC and LCR adjustment schemes, respectively. The NLL for the mixture

distributions is

$$\text{NLL}^{BMA,BMD} = -\sum_{i=1}^{N} \log \left\{ \frac{1}{M \times \sqrt{c^{\text{kernel}}}} \sum_{m=1}^{M} \phi \left(\frac{y_i - \hat{x}_{im}}{\sqrt{c^{\text{kernel}}}} \right) \right\}. \tag{3.4}$$

The CRPS for a mixture of Gaussian random variables is given in closed form by Grimit et al. [2006], and in the case of exchangeable ensemble members reduces to the following.

$$\text{CRPS}^{BMA,BMD} = \frac{1}{N} \sum_{i=1}^{N} \text{crps} \left[\frac{1}{M \times \sqrt{c^{\text{kernel}}}} \sum_{m=1}^{M} \phi(\frac{y_i - \hat{x}_{im}}{\sqrt{c^{\text{kernel}}}}) \right], \tag{3.5}$$

where

$$\begin{aligned} &\text{crps} \left\{ \frac{1}{M \times \sqrt{c^{\text{kernel}}}} \sum_{m=1}^{M} \phi \left(\frac{y_i - \hat{x}_{im}}{\sqrt{c^{\text{kernel}}}} \right) \right\} \\ &= \frac{1}{M} \sum_{m=1}^{M} A(y - \hat{x}_{im}, c^{\text{kernel}}) - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{k=1}^{M} A(\hat{x}_{im} - \hat{x}_{ik}, 2c^{\text{kernel}}) \end{aligned} \tag{3.6}$$

and the function $A(\mu, \sigma^2) = 2\sigma\phi(\mu/\sigma) + \mu\{2\Phi(\mu/\sigma) - 1\}$ gives the expectation of the absolute value of a Gaussian-distributed random variable with mean μ and variance σ^2. Equation (3.5) can be derived by considering the kernel representation of the CRPS (see equation (2.45) on page 58).

For BMA, the dressing kernel variance c^{kernel} is estimated as a parameter in addition to those of the ensemble adjustment scheme by optimising either the NLL or CRPS objective function. For BMD, however, it is calculated explicitly using equation (2.9), while the remaining model parameters, that adjust the ensemble forecasts according to the choice of ensemble adjustment scheme, are estimated with the objective function. For each iteration of the objective function for BMD, therefore, the ensemble adjustment parameters are updated, and c^{kernel} is subsequently calculated using equation (2.9), which depends on the adjusted ensemble forecasts and observations. The parameter estimates are those that are obtained when the numerical algorithm is deemed to have converged.

3.3.3.3 Parameter estimation for LR

Recall from section 2.4.4.3 that separate parameters must be estimated for the LR model for each threshold, q, of interest. For a fixed threshold q, the NLL function

for logistic regression is given by

$$\text{NLL}^{LR} = -\sum_{i=1}^{N} \left\{ \mathbf{I}(y_i \leq q)\eta_i - \log\left(1 + e^{\eta_i}\right) \right\}, \tag{3.7}$$

where η_i is the linear predictor on the ith forecast occasion. As discussed in section 3.3.2, in this chapter we consider two linear predictors, $\eta_i = a + b\bar{x}_i$ and $\eta_i = a + b\bar{x}_i + ds_i^2$.

It is easily shown that the CRPS for Logistic regression reduces to the Brier score over the training data, that is

$$\text{CRPS}^{LR} = \frac{1}{N} \sum_{i=1}^{N} \left\{ p_i - \mathbf{I}(y_i \leq q) \right\}^2, \tag{3.8}$$

where $p_i = e^{\eta_i}/(1 + e^{\eta_i})$ is the probability forecast $\Pr(y_i \leq q)$. To see this, let z_i temporarily denote the binary event $\mathbf{I}(y_i \leq q)$. The distribution of z_i is $z_i \sim \text{Ber}(p_i)$, where $\text{Ber}(p_i)$ indicates a Bernoulli distribution with parameter p_i, such that $\Pr(z_i = 1) = p_i$. The CDF of z_i, say $F(u) = \Pr(z_i \leq u)$, is

$$F(u) = \begin{cases} 0 & u < 0 \\ 1 - p_i & 0 \leq u < 1 \\ 1 & u \geq 1, \end{cases}$$

where u is a dummy variable that can take any value on the real line. The verifying binary observation, z_i, represents the 'perfect forecast' for which all probability mass is placed on $u = 0$ (if $y_i > q$) and on $u = 1$ (if $y_i \leq q$). Finally, recall the integral representation of the CRPS (equation (2.44), see page 58). For binary predictands the integral can be written as

$$\text{crps}(y_i, q) = \int_{-\infty}^{\infty} \left\{ F(u) - \mathbf{I}(z_i \leq u) \right\}^2 du,$$

which here reduces to

$$\text{crps}(y_i, q) = \int_{0}^{1} \left\{ F(u) - \mathbf{I}(z_i \leq u) \right\}^2 du,$$

where $F(u)$ is as given above. Evaluation of this integral gives

$$\text{crps}^{LR}(y_i, q) = \begin{cases} p_i^2 & z_i = 1 \\ (1 - p_i)^2 & z_i = 0, \end{cases}$$

which is equal to the Brier score for probability forecasts of the binary event z_i.

Unfortunately, in practice we found that parameter estimation by minimising the CRPS was numerically unstable for the LR model — the parameter estimates were sensitive to small changes in their starting values, and often the numerical algorithms did not converge. The resulting out-of-sample probability forecasts given by the CRPS parameter estimates were significantly less skilful than their NLL counterparts, and were often less skilful than the baseline DMO forecasts. In section 3.5, therefore, we do not show or comment on the skill of LR probability forecasts with CRPS parameter estimation.

3.3.3.4 Parameter estimation for RHR

The rank histogram recalibration (RHR) post-processing method, as specified in the founding papers of Hamill and Colucci [1997, 1998], do not require the use of an objective function for parameter estimation (see the discussion in section 2.4.2.2). The user merely needs to calculate the weights, w_m for $m = 1, 2, \ldots, M+1$, that are given by the proportion of observations that lie in the $M+1$ rank histogram bins over the training sample. In this work, however, we investigated applying our hierarchy of ensemble adjustment schemes, so as to improve the corrections to the biases in location and/or dispersion of the ensemble forecasts. In doing so, therefore, it is necessary to estimate the parameters of the ensemble adjustment schemes by minimisation of either the NLL or CRPS objective function. In both cases this has proven to be problematic, as we now explain. Recall from section 2.4.2.2 and equation (2.8) that probability forecasts of the form $\Pr(y_i \leq q)$ are given by either a continuous probability distribution, if the threshold q is unbounded by the ensemble forecast, $\boldsymbol{x}_i$, or by a weighted mixture of non-overlapping uniform distributions, if q is bounded by the ensemble. Clearly, therefore, adjustments to the ensemble forecasts in the training dataset, such as those that result from the CC, LC and LCR schemes, will result in changes to the weights w_m. Specifically, for an (adjusted) ensemble forecast, $\widehat{\boldsymbol{x}}_i = (\hat{x}_{i1}, \hat{x}_{i2}, \ldots, \hat{x}_{iM})$, the contribution of the ith forecast and observation to the negative log-likelihood is

$$\mathrm{NLL}_i^{\mathrm{RHR}} = \begin{cases} -\log\left\{\frac{w_1}{G^{\mathrm{RHR}}(\hat{x}_i^{(1)})} \cdot g^{\mathrm{RHR}}(y_i)\right\} & \text{if } y_i < \hat{x}_i^{(1)}, \\ -\log\left\{\frac{w_{m+1}}{\hat{x}_i^{(m+1)} - \hat{x}_i^{(m)}}\right\} & \text{if } \hat{x}_i^{(m)} < y_i < \hat{x}_i^{(m+1)}, \\ -\log\left\{\frac{w_{M+1}}{1 - G^{\mathrm{RHR}}(\hat{x}_i^{(M)})} \cdot g^{\mathrm{RHR}}(y_i)\right\} & \text{if } y_i > \hat{x}_i^{(M)}, \end{cases}$$

where $g^{\mathrm{RHR}}(\cdot)$ and $G^{\mathrm{RHR}}(\cdot)$ denote the PDF and CDF of the distribution used for extrapolation of unbounded quantities, and $\hat{x}_i^{(m)}, \hat{x}_i^{(m+1)}$ denote consecutive order statistics of the adjusted ensemble $\hat{\boldsymbol{x}}$.

Similarly, referring to equations (2.44) (page 58) for the CRPS and (2.7) (page 32) for the weights, we see that the CRPS is a function of the weights $w_1, w_2, \ldots, w_{M+1}$, the distances between ensemble members, as well as the tails of the rank histogram distribution. For each iteration of the objective function, the chosen ensemble adjustment scheme results in a change to the location of the ensemble members which, in turn, results in a change to the weights w_m, for $m = 1, 2, \ldots, M+1$. As the weights take discrete, rather than continuous values, it is intuitive to think that the effect of the ensemble adjustment scheme on the objective function surface is not continuous. In addition, the tails of the rank histogram distribution, which must be used to extrapolate probability forecasts for quantities that lie below (above) the smallest (largest) ensemble member, pose further numerical difficulties. For example, the contribution to the CRPS of the lower tail of the RHR forecast distribution is

$$\begin{aligned}
&\mathrm{CRPS}^{\mathrm{Lower}} \\
&= \begin{cases} \frac{w_1}{G^{\mathrm{RHR}}(\hat{x}_i^{(1)})} \int_{-\infty}^{\hat{x}_i^{(1)}} \left\{G^{\mathrm{RHR}}(u)\right\}^2 \, du & \text{if } y_i > \hat{x}_i^{(1)} \\ \frac{w_1}{G^{\mathrm{RHR}}(\hat{x}_i^{(1)})} \left[\int_{-\infty}^{y_i} \left\{G^{\mathrm{RHR}}(u)\right\}^2 \, du + \int_{y_i}^{\hat{x}_i^{(1)}} \left\{G^{\mathrm{RHR}}(u) - 1\right\}^2 \, du\right] & \text{if } y_i < \hat{x}_i^{(1)}. \end{cases}
\end{aligned}$$

Similar expressions hold for the contribution to the CRPS of the upper tail of the forecast distribution. In general such integrals do not have closed forms, and so we must resort to numerical approximations. This is problematic for the calculation of the CRPS, both in terms of accuracy and computational expense.

We found parameter estimates for the RHR post-processing method obtained by minimising both the NLL and CRPS to be highly unstable — small changes in the starting values result in large changes to the final parameter estimates. In many cases the numerical algorithm did not converge, despite the maximum permitted number of iterations being increased significantly above the defaults implemented in the 'optim' function implemented in the R language [R Core Team, 2015]. This was the case for both small and large training samples. Experiments were conducted using Gaussian-distributed tails, for which numerical integration was necessary in calculating the CRPS, and an alternative assumption of exponentially-distributed tails that were chosen so as the integrand of the lower and upper tails could be evaluated in a closed form. The problems of numerical instability persisted for the second case. The numerical method known as 'simulated annealing', which is widely used for optimisation problems in which the objective function surface is 'rough', was used in place of the Nelder-Mead algorithm without success.

Our results suggest that the mixture of non-overlapping uniform distributions is not conducive to stable parameter estimation. Contour plots of the objective function surface show that the surface is 'rough', that is, the surface does not vary in a

continuous manner as a function of the parameter values, and there are several local minima. After considerable experimentation it was decided to cease investigating the rank histogram recalibration method as a viable post-processing method, since we believe that the parameters of any reasonable method should be stable and converge without difficulty in the majority of cases. Parameter estimates that are highly unstable are indicative of misspecified statistical models, particularly in the light of the relatively stable convergence of the parameter estimates for the other post-processing methods discussed previously. The probability forecasts obtained using the parameter estimates given by the optimisation of the NLL or CRPS were significantly less skilful than either those derived with the standard RHR method [Hamill and Colucci, 1997; Wilks, 2006a], and indeed the baseline DMO probability forecasts.

3.4 Ensemble post-processing in the Lorenz 1996 system

3.4.1 Training and verification data

The skill of probability forecasts that are issued by the BMA, BMD and NGR statistical models under our ensemble adjustment schemes and choice of objective function are assessed using data from the Lorenz 1996 system, which is described in section 2.6.1. The parameters of the various models were estimated using the first dataset, in which the ensemble forecasts and observations exhibit temporal correlation. Parameters are estimated from training samples of size $N = 100, 300$ and 1000. The training samples are nested, such that the training samples of size 100 are also the first 100 ensemble forecasts and observations of the training samples of size 300 which, in turn, are the first 300 ensemble forecasts and observations of the training samples of size 1000. In order to suppress the effects of sampling variation on our results, we performed parameter estimation for each model under consideration on 500 training samples of the desired size. The training samples were disjoint, so that each of the 500 parameter estimates for a given model can be assumed independent. We used an ensemble size of $M = 24$, in keeping with the size of the Met Office Global and Regional Ensemble Prediction System (MOGREPS) of the UK Met Office.

Having obtained parameter estimates for the various statistical models of interest, out-of-sample probability forecasts were calculated for observations $y_t, t = 1, 2, \ldots, T$, for each of the 500 parameter estimates, using the second dataset described in section 2.6.1 for which ensemble forecast-observation pairs are temporally uncorrelated.

The size of the test dataset was $T = 190\,000$, and so the probabilistic forecasting skill of each model was assessed using $500 \times 190\,000 = 9.5 \times 10^7$ independent forecasts and observations. This setup enables us to report on the skill of the various probability forecasts for both common and rare events, without being inhibited by a lack of data. In particular, such large test datasets are valuable for assessing the skill of probability forecasts $\Pr(y_t \leq q)$, for extreme thresholds q, as (by definition) occurrences of the binary event $\{y_t \leq q\}$ are rare. Furthermore, the stationary nature of the data-generating process (the Lorenz 1996 system) means that we need not concern ourselves with seasonal effects that can be encountered in 'real-world' scenarios, such as those highlighted later in this thesis (see chapter 5). In this study we calculate probability forecasts for thresholds given by the 1% and 2% quantiles of the climatology of the Lorenz 1996 system, as well as the lower turcile and the median quantiles as considered by Wilks [2006a].

3.4.2 Forecast verification

We assess the skill of the various post-processing methods in issuing forecasts of the binary event $\{y_t \leq q\}$, where y_t is an out-of-sample verifying observation, using the Brier score and its decomposition (see sections 2.5.3.2 and 2.5.5), and reliability diagrams (see section 2.5.2.2), for the thresholds q given above. Care must be taken when comparing the Brier scores of probability forecasts at different thresholds. A climatology forecast, say α, for the probability forecast $\Pr(y_t \leq q_\alpha)$, where q_α is the α-quantile of the climatological distribution, achieves a Brier score, say B_{Clim}, with expectation

$$\begin{aligned} E(B_{Clim}) &= \alpha \cdot (\alpha - 1)^2 + (1 - \alpha) \cdot \alpha^2 \\ &= \alpha(1 - \alpha). \end{aligned} \tag{3.9}$$

Therefore, a climatology forecast of the median $q_{1/2}$ has an expected Brier score of $1/4$, whereas a climatology forecast for the extreme threshold $q_{1/100}$, the 1% quantile of the climatology, has an expected Brier score of $99/10\,000$. Provided that the Brier score has finite variance, therefore, its expectation tends to 0 as the threshold q tends to extreme values. As a consequence, Brier scores of forecasts of different thresholds should not be compared. Rather, the improvements in skill of a forecast over a baseline forecast, such as climatology or DMO (see section 2.4.2.1), should be used as a measure to compare forecast skill at different thresholds.

The size of the test dataset used for forecast verification in this study, given by $T = 190\,000$, combined with the 500 independent instances of each post-processing method, was chosen so as the scores reported in the following section can be trusted

to at least the degree of accuracy shown. In other words, the standard errors of the scores are small enough to allow us to report the scores to at least the number of significant figures shown.

3.5 Results

3.5.1 Brier scores

We first present the Brier scores for the out-of-sample probability forecasts that are issued by the NGR, BMA and BMD ensemble post-processing methods under the CC, LC and LCR ensemble adjustment schemes at various thresholds q, forecast lead times t and training sample sizes N. As described in section 3.3.2, we also show probability forecasts for logistic regression (LR) model, for two linear predictors, under the labels of LC and LCR. Tables 3.2, 3.3 and 3.4 show the Brier scores for the CC, LC and LCR schemes for forecast lead times $1, 3$ and 5, for thresholds q taken as the 50%, 2% and 1% quantiles of the climatology of the Lorenz 1996 data respectively. Model parameters were estimated by minimising the negative log-likelihood (NLL), using the largest training samples of size $N = 1000$. To ease reading, the scores have been scaled by a factor of 10^4 at the 50% threshold (table 3.2), and by a factor of 10^5 for the 2% and 1% thresholds (tables 3.3 and 3.4). The DMO scores are shown so as to provide a measure against which the ensemble post-processing methods can be compared. The DMO forecasts do not change with the CC, LC and LCR ensemble adjustment schemes, as they are a function of the raw, out-of-sample ensemble forecasts only.

The Brier scores indicate that ensemble post-processing is most beneficial at longer forecast lead times and for forecasts of rare, extreme events. The improvements of the ensemble post-processing methods over the DMO forecasts, particularly under the LC and LCR schemes, are significantly larger at longer lead times and, most interestingly, at the more extreme 2% and 1% thresholds. Here the Brier scores indicate that the DMO forecasts are only slightly more skilful than the climatology forecasts or, in other words, the raw ensemble forecasts contain little skill. However, the ensemble post-processing methods considered (especially NGR, BMA and BMD) yield Brier scores that improve significantly over those of the DMO forecasts, particularly under the LC and LCR schemes.

It is also interesting to compare the skill of forecasts under the three ensemble adjustment schemes. Beginning with the 50% threshold, the scores suggest that the constant correction (CC) adjustment scheme is sufficient for bias correction — there is little to be gained with the linear bias correction (LC) scheme, and, indeed,

		DMO	BMA	BMD	NGR	LR
Lead time 1	CC	138	114	114	114	
	LC	138	114	114	114	118
	LCR	138	114	114	114	120
Lead time 3	CC	458	411	412	411	
	LC	458	411	412	411	422
	LCR	458	405	406	403	417
Lead time 5	CC	951	883	885	882	
	LC	951	884	887	882	908
	LCR	951	885	885	876	910

Table 3.2 Brier scores for the CC, LC and LCR schemes at forecast lead times $1, 3$ and 5, evaluated at the 50% threshold. The climatology forecast score is 2500.

		DMO	BMA	BMD	NGR	LR
Lead time 1	CC	288	208	208	208	
	LC	288	215	215	214	243
	LCR	288	214	214	214	279
Lead time 3	CC	894	796	800	796	
	LC	894	596	604	596	619
	LCR	894	595	591	603	644
Lead time 5	CC	1562	1474	1498	1471	
	LC	1562	1191	1204	1199	1237
	LCR	1562	1195	1195	1228	1249

Table 3.3 Brier scores for the CC, LC and LCR schemes at forecast lead times $1, 3$ and 5, evaluated at the 2% threshold. The climatology forecast score is 1960.

		DMO	BMA	BMD	NGR	LR
Lead time 1	CC	198	124	124	124	
	LC	198	127	127	127	166
	LCR	198	127	127	127	214
Lead time 3	CC	566	496	502	497	
	LC	566	343	349	342	369
	LCR	566	339	337	342	403
Lead time 5	CC	987	958	982	959	
	LC	987	700	711	707	742
	LCR	987	700	700	730	758

Table 3.4 Brier scores for the CC, LC and LCR schemes at forecast lead times $1, 3$ and 5, evaluated at the 1% threshold. The climatology forecast score is 990.

in some instances the Brier scores slightly deteriorate. Furthermore, rescaling the ensemble forecasts with the LCR scheme yields only small improvements at lead times 3 and 5. Similar comments hold for forecast lead time 1 at the 2% and 1% thresholds. By contrast, however, at these more extreme thresholds the LC scheme yields significant improvements in the Brier scores for the longer forecast lead times, clearly demonstrating the benefits of bias corrections that depend on the ensemble mean for probability forecasts at such thresholds. Again the Brier scores indicate that there is little to be gained with the LCR scheme, and indeed the forecast skill deteriorates in some instances, most notably for the NGR forecasts at lead time 5.

The BMA, BMD and NGR post-processing methods yield largely similar Brier scores across the three ensemble adjustment schemes, and it is not possible to determine one method as the most skilful. Unlike BMA, the BMD forecasts appear to improve slightly under the LCR scheme, particularly at longer forecast lead times. The Brier scores indicate that both LR models yield probability forecasts that are markedly less skilful than BMA, BMD and NGR, particularly at the more extreme thresholds. This presumably derives from the fact that realisations of the binary event $\{y \leq q\}$ are rare for extreme thresholds, and hence there are a lack of events with which to estimate the model parameters. Inclusion of the ensemble variance in the linear predictor ($\eta = a + b\bar{x} + cs^2$, the LCR scheme) yields forecasts whose Brier scores are slightly worse than those of the simpler alternative, $\eta = a + b\bar{x}$.

Brier scores for post-processing methods whose model parameters are estimated with the CRPS objective function (not shown) are both qualitatively and quantitatively similar. Furthermore, the Brier scores for the smaller training sample sizes are qualitatively similar. In general the improvements in forecast skill with training sample size appear small, with the exception of LR, for which the forecast skill deteriorates markedly for smaller training samples, particularly at rare thresholds. This is illustrated in figure 3.1 above. Brier scores for the NGR^0 scheme described in section 3.3.2 (not shown), corresponding to the modelling constraint $c = 0$, are significantly worse than the standard case in which c is estimated as a model parameter. At common thresholds the Brier scores for the NGR^0 forecasts are worse than those of the LR forecasts, and are comparable at rare thresholds. We do not comment on the skill of NGR^0 probability forecasts hereafter.

Figure 3.1 shows the Brier scores as a function of forecast lead time for the 50%, 2% and 1% thresholds, and for training samples of size 1000, 300 and 100. As we expect, the Brier scores deteriorate with forecast lead time in all cases. The BMA, BMD and NGR post-processing methods are relatively insensitive to the size of training sample, whereas the Brier scores of the LR forecasts deteriorate markedly for small training samples as well as threshold extremity. The figure again illustrates the benefits of ensemble post-processing for the most difficult forecasts of extreme

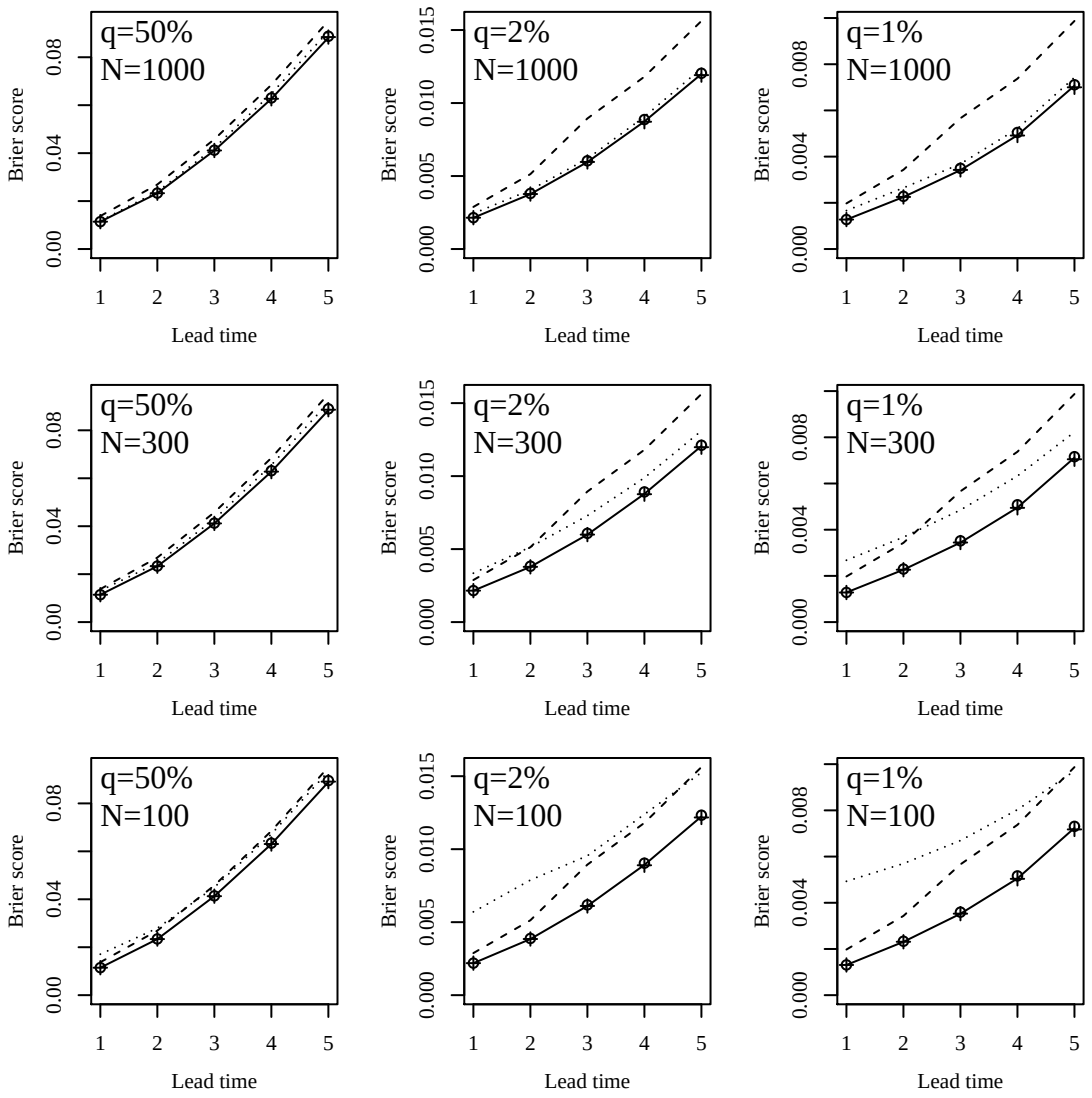

Figure 3.1 Brier scores as a function of lead time for the DMO (dashed), BMA (crosses), BMD (circles), NGR (solid) and LR (dotted) forecasts under the LC scheme at the 50%, 2% and 1% thresholds with parameter estimation performed with training samples of size $N = 1000$, 300 and 100.

thresholds and longer forecast lead times — the improvements over the DMO Brier scores are large relative to the 50% threshold and small forecast lead times.

3.5.2 Forecast reliability and resolution

We now turn to a discussion of the reliability of probability forecasts issued by the various post-processing models under the CC, LC and LCR ensemble adjustment schemes. Tables 3.5, 3.6 and 3.7 show the reliability component of the decomposition of the Brier scores (see equation (2.51) on page 62) at the 50%, 2% and 1% thresholds, respectively, for forecast lead times $1, 3$ and 5. Figures 3.2, 3.3 and 3.4 show reliability diagrams at the 50% and 1% thresholds for forecast lead times $1, 3$ and 5, respectively. It should be kept in mind that the reliability diagrams at ex-

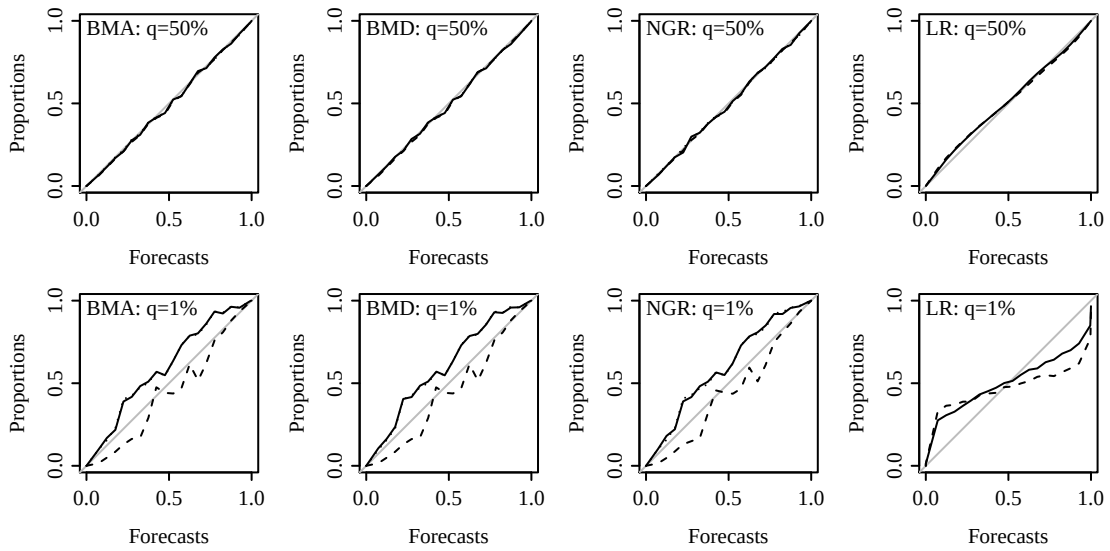

Figure 3.2 Reliability diagrams at lead time 1 for BMA, BMD and NGR with adjustment schemes CC (dashed), LC (dotted) and LCR (solid) for common and extreme thresholds, q. Also LR with $\eta = a + b\overline{x}$ (solid) and $\eta = a + b\overline{x} + ds^2$ (dashed).

treme thresholds are largely comprised of small forecast probabilities, and so the majority of forecasts fall in the lower portion of the reliability curves.

As with the Brier scores, forecast reliability is affected by the threshold of interest, the forecast lead time, and the choice of ensemble adjustment scheme (CC, LC, or LCR). The LR forecasts are consistently less reliable than the BMA, BMD and NGR forecasts, particularly at extreme thresholds, and so we omit them from our discussion hereafter. For forecast lead time 1, there seems little to be gained in using the LC scheme in place of the CC scheme, for either the extreme or common thresholds considered in this study (see figure 3.2, and the three aforementioned tables). Indeed, at the 2% and 1% thresholds, the LC scheme has the effect of overcompensating the bias in probability forecasts that is observed for the lower and mid-range of probabilities under the CC scheme, which explains the deterioration in the reliability scores shown in tables 3.6 and 3.7. Under the LC scheme, the probability forecasts at the 1% threshold are biased throughout the range of probabilities — the probability forecasts consistently underestimate the realised proportion of event occurrences. Similarly, the LCR scheme makes no significant improvement to the forecast reliability at forecast lead time 1 for any of the thresholds considered.

For longer forecast lead times (figures 3.3 and 3.4), however, the parametric form of the expectation of the probability forecast distributions has a notable effect on forecast reliability at the 2% and 1% thresholds, but not at the 50% threshold. At the 50% threshold, the LC scheme again yields probability forecasts whose reliability is no better, and in some cases worse than those of the CC scheme, particularly for forecasts issued by the BMD post-processing method. For the 2% and 1% thresholds, however, the LC scheme significantly improves the forecast reliability, by correcting

		DMO	BMA	BMD	NGR	LR
Lead time 1	CC	1348	20.9	21.5	16.8	
	LC	1348	17.8	18.3	16.0	258.0
	LCR	1348	20.8	19.9	18.4	270.0
Lead time 3	CC	2499	182	234	191	
	LC	2499	196	257	206	538
	LCR	2499	86.9	81.9	92.5	440
Lead time 5	CC	4394	368	391	416	
	LC	4394	384	590	477	1080
	LCR	4394	334	335	524	1060

Table 3.5 The reliability component of the Brier score decomposition for forecast lead times $1, 3$ and 5 at the 50% threshold.

		DMO	BMA	BMD	NGR	LR
Lead time 1	CC	1378	67	67	65	
	LC	1378	140	140	140	213
	LCR	1378	140	140	140	230
Lead time 3	CC	1710	2000	2100	2000	
	LC	1710	81	150	73	237
	LCR	1710	73	51	130	257
Lead time 5	CC	2859	2800	3000	2700	
	LC	2859	110	210	150	198
	LCR	2859	120	110	370	382

Table 3.6 The reliability component of the Brier score decomposition for forecast lead times $1, 3$ and 5 at the 2% threshold.

		DMO	BMA	BMD	NGR	LR
Lead time 1	CC	1353	59	59	60	
	LC	1353	100	100	99	191
	LCR	1353	100	100	96	257
Lead time 3	CC	1070	1600	1700	1600	
	LC	1070	100	160	91	269
	LCR	1070	64	54	73	306
Lead time 5	CC	2071	2600	2800	2500	
	LC	2071	91	170	130	385
	LCR	2071	100	100	320	526

Table 3.7 The reliability component of the Brier score decomposition for forecast lead times $1, 3$ and 5 at the 1% threshold.

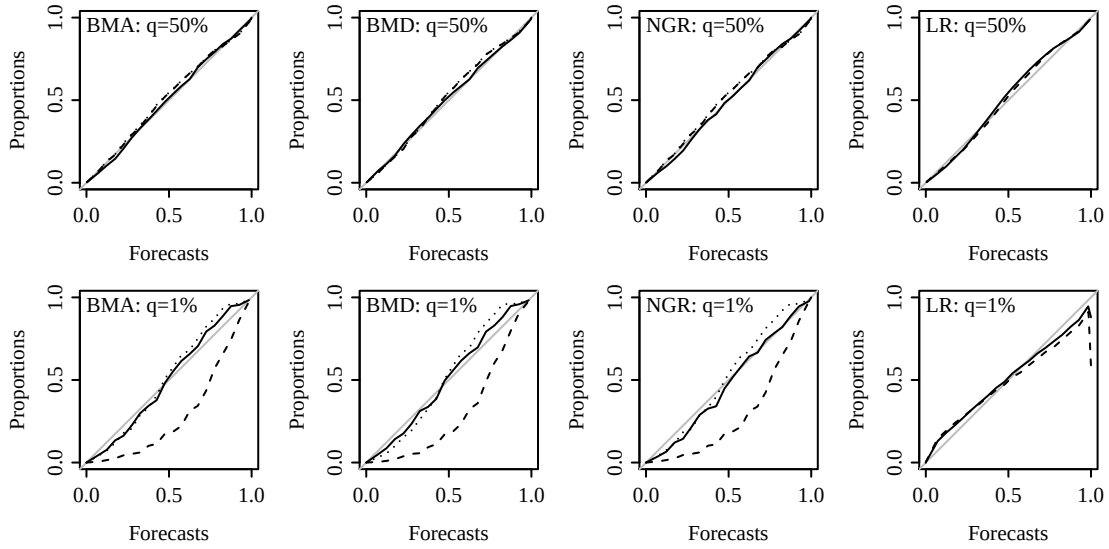

Figure 3.3 As for figure 3.2 but for forecast lead time 3.

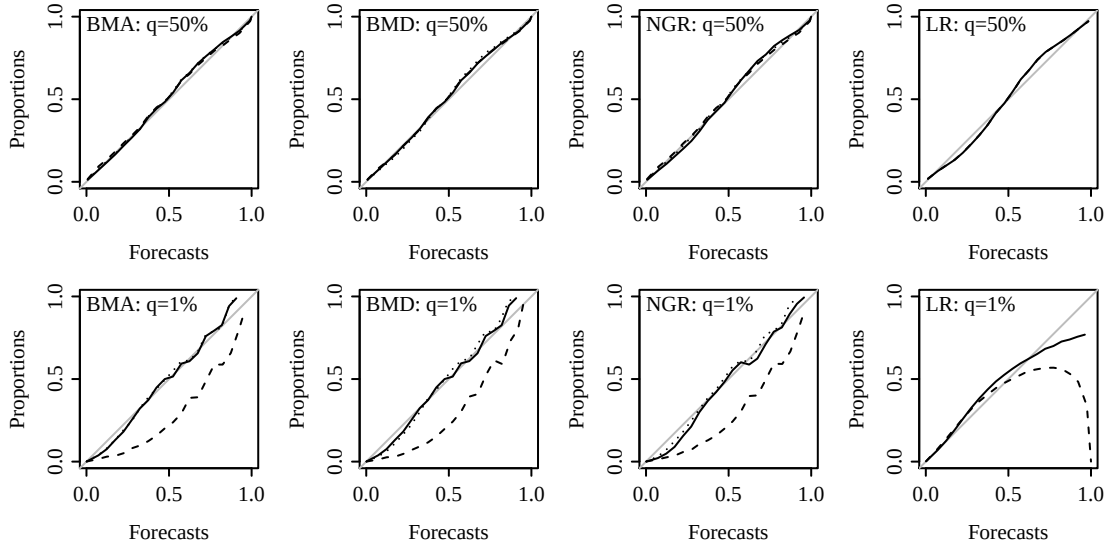

Figure 3.4 As for figure 3.3 but for forecast lead time 5.

for much of the systematic bias in the probability forecasts that is observed for the CC scheme. At these thresholds, the CC probability forecasts are consistently too large — the proportion of realisations of the binary events $\{y_t \leq q\}$ is significantly less than the corresponding forecast probabilities, particularly for the small probabilities that are most common for these rare events. In other words, the NGR, BMA and BMD probability forecast distributions with expectations and variances given by the CC ensemble adjustment scheme on average place too much probability density below the 1% threshold, a feature that is (to a large extent) corrected by the alternative specification of the expectations of the forecast distributions under the LC scheme.

For longer lead times, the effects of specifying the forecast variance as a linear function of the ensemble variance (by using the LCR adjustment scheme), in addition to the linear correction to the bias in ensemble location (the LC scheme) for forecasts of rare events depend on both the lead time and post-processing method. At lead time 3, the LCR adjustment scheme improves the reliability of the BMA, BMD and NGR forecasts compared to the LC scheme, although the tendancy of the forecasts to be underconfident (corresponding to overdispersed forecast distributions) remains for LCR. The underconfidence of both the LC and LCR forecasts is indicated by the S-shaped reliability diagrams in the lower panel of figure 3.3. By contrast, however, at lead time 5 the LCR adjustment scheme improves only marginally on the LC scheme and, in the case of the NGR post-processing method, results in a deterioration in the forecast skill. The LCR scheme is, however, beneficial for forecasts given by the BMD post-processing method, which appear similar in skill to the BMA forecasts at the 2% and 1% thresholds for forecast lead time 5. This feature is explained by comparing the parameter estimates of the two competing post-processing models (not shown). Under the LC scheme, the parameter estimates $\hat{a}$ and $\hat{b}$, that determine the ensemble bias corrections, are very similar. However, the estimated dressing kernel variance, $\hat{c}^{\text{kernel}}$, is on average somewhat larger for the BMD method. Under the LCR scheme, however, the four parameter estimates, $\hat{a}, \hat{b}, \hat{c}^{\text{kernel}}$ and $\hat{d}$ are very similar, and so the BMA and BMD probability forecast distributions are also very similar.

Under the LCR scheme, the NGR forecasts appear less reliable than the BMA and BMD forecasts at extreme thresholds and for longer forecast lead times. The parameter estimates of d (not shown), which correspond to the rescaling of the ensemble forecasts, are somewhat larger than their BMA and BMD counterparts, and on average the variance of the NGR forecast distributions is larger. In cases where some of the ensemble members are below the extreme threshold, the restriction of the NGR method to Gaussian forecast distributions may necessitate larger forecast variances to achieve a non-negligible probability forecast, whereas the more flexible

mixture distributions of the BMA and BMD methods achieve this as a matter of course, without needing to artificially inflate the forecast variance.

Perhaps surprisingly, with the exception of the BMD method, the forecasts at extreme thresholds under the LC adjustment scheme appear slightly more reliable at the theoretically more difficult forecast lead time 5 (figure 3.4) than at lead time 3 (figure 3.3). This is also in keeping with the reliability scores given in table 3.7. While the reliability diagrams at lead time 5 (figure 3.4) again indicate underconfidence in the BMA, BMD and NGR forecasts, particularly for small forecast probabilities, the effect is slightly less pronounced than at the shorter lead time. However, the improvements in forecast reliability when using the LCR scheme at lead time 3 (as discussed above) results in forecasts that are slightly more reliable than the most reliable forecasts at lead time 5, as we might reasonably expect.

In contrast to the forecast reliability, there is little to be learned from the resolution components of the Brier score decomposition (see equation (2.52) on page 62, not shown). As we expect, the forecast resolution decreases with forecast lead time. The DMO forecasts, which are based on the verification data only, exhibit considerably higher resolution than those of the post-processing methods. However, this increased resolution is at the expense of forecast reliability — plainly the calibration of the DMO probability forecasts is considerably worse than the corresponding post-processed forecasts. Unlike the reliability scores discussed above, the resolution scores for the NGR, BMA and BMD probability forecasts are insensitive to the choice of ensemble adjustment scheme. The scores are also similar across the various ensemble post-processing methods, with the exception of LR. The LR forecasts generally exhibit similar resolution to the BMA, BMD and NGR forecasts at the 50% threshold, but considerably worse resolution for forecasts at the extreme thresholds considered here. At lead time 5, the resolution scores at the extreme thresholds are less than half those for lead time 3, which provides an explanation for the comparable forecast reliability between the two lead times as discussed above. The forecasts improve only slightly in resolution compared with the climatalogical forecast at lead time 5 and, as discussed in section 2.4.2.1, the climatology is a reliable forecast.

The forecast reliability and resolution for post-processing methods with parameters estimated by CRPS optimisation (not shown) are qualitatively and quantitatively similar. The reliability scores suggest that CRPS parameter estimation yields forecasts that are slightly more reliable at longer lead times ($t > 3$), while likelihood-based parameter estimates might be preferable at shorter lead times. In general, though, it is not clear from our results that one objective function is preferable over the other.

The qualitative dependence of the threshold and forecast lead time on forecast re-

liability is thought to be a result of the difficulty inherent in predicting extreme events, and also possible deficiencies in the assumed relationships governing the forecast mean and variance. In this study all training data are equally weighted, meaning that the majority of the forecasts and observations contained in the training samples used for parameter estimation pertain to common, rather than extreme values. Furthermore, it may well be the case that the assumed linear relationships that link the ensemble means and variances with the expectations and variances of the NGR, BMA and BMD probability forecast distributions are inadequate, either in general (for certain lead times) or for extreme events. This point is discussed further in chapter 6 of this thesis.

3.6 Discussion and conclusions

In this chapter we have investigated how certain 'features' of ensemble post-processing methods affect the skill of probability forecasts of the form $\Pr(y \leq q)$, in particular the parametric functions that are used to model the expectations and variances of the probability forecast distributions and the objective function that is used for parameter estimation. Using data from the Lorenz 1996 system, our study has demonstrated that allowing flexible parametric forms for the expectation and variance of probability forecast distributions significantly improves the skill of probability forecasts for rare, or extreme events. In particular, correcting biases in ensemble location with a linear function of the ensemble mean was shown to be the most influential factor on the skill of such probability forecasts. Our study has also facilitated a fair comparison of several popular post-processing methods, in particular Bayesian model averaging, best member dressing and nonhomogeneous Gaussian regression. We have found that these ensemble post-processing methods yield probability forecast distributions that are similar in skill, when their expectations and variances are of the same parametric form.

Our results indicate that the choice of ensemble adjustment scheme is the most important feature of ensemble post-processing for probability forecasts of rare events. While the constant correction (CC) scheme is sufficient for the shortest forecast lead times, we have found that the use of the ensemble mean in the bias correction (the LC scheme) results in significant improvements to forecast skill for longer forecast lead times. In general, the LCR scheme did not improve the skill of forecasts of extreme events, with the exception of the BMD method, which performed similarly to BMA under LCR, but worse under LC. Indeed, in several instances the forecasts under the LCR scheme were less reliable than those under the LC scheme. In contrast, for probability forecasts at common thresholds, the constant correction (CC) scheme was sufficient. Indeed, forecasts under the LC and LCR schemes were often less

skilful. It may well be the case, however, that the LC scheme would prove useful for forecasts of common events with lead times beyond those considered in this study.

The logistic regression (LR) forecasts were significantly less skilful than those of BMA, BMD and NGR, particularly at extreme thresholds. Furthermore, the skill of the LR forecasts degrades notably for small training samples, and therefore the LR method is not at all competitive with the small training samples that are likely to be encountered in real world scenarios. King and Zeng [2001] proposed an adapted logistic regression model for the probabilistic forecasting of rare events. We did not consider this method, as such an investigation was not in keeping with the primary purpose of the work, namely to extend pre-existing ensemble post-processing methods and to investigate the effect of the choice of ensemble adjustment scheme on the skill of forecasts derived from the continuous probability forecast distributions. Nonetheless, the work of King and Zeng [2001] may prove useful for scenarios in which only binary observations are available.

There is some evidence that the mixture forecast distributions of the BMA and BMD post-processing methods yield more reliable probability forecasts of rare events, compared with the more restrictive Gaussian distributions issued by the NGR method. Intuitively, it seems plausible that ensemble forecasts located near to extreme values may exhibit larger spread and skewness than ensemble forecasts located at more common values, due to the unpredictable nature of extreme atmospheric conditions. If this is indeed the case, the more flexible form of the mixture distributions over the NGR distributions would seem to be advantageous. Further comparisons of the relative skill of these three post-processing methods for probabilistic forecasting of rare events are encouraged, particularly for 'real-world' forecasting scenarios.

We found that the choice of objective function (NLL or CRPS) was not influential on the skill of probability forecasts, for either common or rare events. It may be worth verifying this finding with meteorological data, and alternative suggestions for objective functions would be welcome. If such findings persist, however, we suggest using likelihood-based parameter estimates, due to their appealing properties (see section 2.4.7) and relative lack of computational cost. We found CRPS parameter estimation to be approximately 100 times slower than NLL estimation for the BMA and BMD post-processing methods.

4 A distribution-free ensemble post-processing method

4.1 Introduction

The majority of ensemble post-processing methods reviewed in chapter 2 require the forecaster to specify a family of parametric distributions with which to model the distribution of the verifying observations, conditional on the corresponding ensemble forecasts. The parameter values are typically chosen as those that minimise an objective function, such as the negative log-likelihood (NLL) or continuous rank probability score (CRPS). Such objective functions depend on the form of the chosen family of distributions through its distribution function. The results presented in chapter 3 and Williams et al. [2014] suggest that seemingly significant differences in the choice of the family of distributions do not result in significant differences in forecast skill, at least in the Lorenz 1996 system. For example, the mixture distributions given by the Bayesian model averaging (BMA) and Best member dressing (BMD) post-processing methods can differ significantly from the Gaussian NGR distributions in terms of skewness and possible multimodality, but display similar forecast skill. Indeed, the most significant differences in forecast skill were attributed to the choice of ensemble adjustment scheme (see section 3.3.2), which enable the forecaster to specify the form of the first and second moments, or expectation and variance, of the probability forecast distributions.

With the above comments and the results presented in chapter 3 in mind, one might think that the only problems worth addressing are the recalibration of the moments of the probability forecast distributions, such as their expectation and variance. Indeed, these were the subject of interest in section 3.3, in which we introduced three 'ensemble adjustment schemes' that facilitated more flexible specifications for the expectation and variance of the probability forecast distributions than had been previously considered. An interesting question, therefore, is to what extent does specifying a parametric family of distributions improve forecast skill, compared with forecasts that do not make distributional assumptions? To answer this question, it is necessary to formulate an ensemble post-processing method that allows for modelling

of the moments in an analogous manner to that of the aforementioned ensemble adjustment schemes, but that allows for estimation of the model parameters, that are used for correcting the bias in ensemble location and dispersion, in a distribution-free setting.

A forecast user may also wish to avoid distributional assumptions if they require post-processed ensemble forecasts, rather than probability forecasts. As mentioned in section 2.4.6, this may be necessary for forecasts over high-dimensional multivariate domains, such as spatial fields. In order to issue probability forecasts of such multivariate quantities, it is necessary to apply a post-processing method to each of the marginal variables of interest — for example, at each gridpoint in the case of gridded forecasts. It is further necessary to issue the high-dimensional and often complicated dependence structure of the marginal variables, which is likely to prove difficult to work with for forecast users. On the other hand, ensemble forecasts of multivariate variables simply require the forecaster to issue (post-processed) ensemble forecasts at each location. As is the case with the ECC methodology [Schefzik et al., 2013], incorporating the dependence structure of the marginal variables forms part of the post-processing method, and there is no need to issue any information about this structure as part of the forecast. As described in section 2.4.6 and in Schefzik et al. [2013], ensemble forecasts for each marginal variable are typically sampled from probability distributions that result from the application of an ensemble post-processing method. Schefzik et al. [2013] suggest that sampling equidistant quantiles $q_{m/(M+1)}, m = 1, 2, \ldots, M$, from the probability distributions is preferable, although the forecaster may also wish to sample ensemble members as independent and identically distributed draws from the distributions. However, the forecaster may prefer to avoid the necessary steps of specifying a family of distributions and choosing an ensemble sampling scheme, and to recalibrate ensemble forecasts directly. This is an appealing benefit of the distribution-free post-processing method introduced in this chapter — the dependence structure of the multivariate ensemble forecasts is inherently preserved, and we simply post-process the ensemble forecasts for each of the marginal variables.

As mentioned previously, in this chapter we introduce an ensemble post-processing method that obviates the need for distributional assumptions. The method leads immediately to recalibrated ensemble forecasts, and thus also circumvents the need for a choice of ensemble sampling scheme, such as those discussed in section 2.4.6. Our distribution-free method allows for the correction of biases in the ensemble location and dispersion using linear functions of the ensemble mean and variance, in an analogous manner to the NGR method (see equations (2.19) and (2.20) on page 38) and the LCR ensemble adjustment scheme introduced in chapter 3. The post-processing method therefore requires the estimation of four parameters.

A key challenge is in the estimation of the parameters used in correcting the bias of the ensemble location and dispersion. We use the method of moments (see section 2.4.7.2) which, unlike the NLL and CRPS objective functions, depends only on the moments of summary statistics of the data, and does not require the specification of a family of distributions. The method of moments merely requires that the moments exist, an assumption that is satisfied in the examples used here. Comparisons of forecast skill between our moment-based, distribution-free methods and standard post-processing methods, such as NGR, will therefore shed light on the extent to which distributional assumptions are beneficial, or otherwise.

Our distribution-free ensemble post-processing method is introduced by using the idea of a latent variable model, in which the observations and ensemble forecasts are linked through an underlying *ensemble distribution*, previously motivated in section 2.2.4. The latent variables are the expectation, ξ, and variance, θ^2, of the ensemble distributions. We propose two slightly different approaches to parameter estimation, that depend on the interpretation of the ensemble forecasts. Firstly, the ensemble members are viewed as known constants, and therefore the ensemble means and variances are also viewed as being known exactly. This approach is in keeping with the majority of the post-processing literature, and indeed regression problems in general. Secondly, we view the ensemble forecasts as independent and identically distributed (IID) realisations from the ensemble distributions, and thus the ensemble means and variances are estimates of the unobserved latent variables ξ and θ^2. This approach gives rise to the use of a measurement error model, the application of which is novel in this field.

The remainder of this chapter is organised as follows. In section 4.2 we explain our novel statistical approach to distribution-free ensemble post-processing. In this section we detail our underlying latent variable model, explain our use of measurement error models in the context of IID ensemble members, derive parameter estimates under both interpretations and discuss the issue of out-of-sample forecasting. Results are presented in section 4.3 with a simulation experiment, the Lorenz 1996 system and gridded forecasts of 2 metre temperature. Concluding remarks and a discussion are given in section 4.4.

4.2 Distribution-free ensemble post-processing: methodology

4.2.1 The model

As mentioned in the previous section, our distribution-free ensemble post-processing method is based on a statistical model that relates the ensemble forecasts and verifying observations through latent variables that are either assumed to be observed exactly, or with additional noise. We make clear the distinction between the two cases below. For now, we denote by ξ_i and θ_i^2 the expectation and variance of the ith 'ensemble distribution' (see the discussion in section 2.2.4), from which the members of the ensemble forecast $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$ are assumed to be sampled. The ensemble members could be either known constants, for example if sampled as quantiles of the ensemble distribution, or random variables, if the members are sampled at random — we return to this point later in this section. We again denote the ith verifying observation by y_i. We assume that y_i depends on the latent variables ξ_i and θ_i^2 through the relations

$$E(y_i \mid \xi_i, \theta_i) = a + b\xi_i \tag{4.1}$$

$$\text{var}(y_i \mid \xi_i, \theta_i) = c + d\theta_i^2. \tag{4.2}$$

The constants a, b, c and d are parameters that are to be estimated from a training sample of historical ensemble forecasts $\boldsymbol{x}_i$ and observations y_i, where $i = 1, 2, \ldots, N$ and N is the training sample size.

Equivalently, equations (4.1) and (4.2) can be thought of as a regression model of the form

$$y_i = a + b\xi_i + \epsilon_i, \tag{4.3}$$

where ϵ_i is a realisation of a random variable whose distribution is a member of the family of ensemble distributions, with conditional expectation $E(\epsilon_i \mid \xi_i, \theta_i) = 0$ and conditional variance $\text{var}(\epsilon_i \mid \xi_i, \theta_i) = \sigma_{\epsilon,i}^2 = c + d\theta_i^2$. In the regression literature ϵ_i is commonly referred to as the 'error in the equation'. Unlike most regression problems, in which the variance $\sigma_{\epsilon,i}^2$ is assumed constant (and so independent of ξ_i and θ_i^2) for all i, we stress the distinction that our model is nonhomogeneous with variance that depends linearly on the variance of the ensemble distribution, θ_i^2.

The statistical model described above allows for biases in ensemble location and dispersion that are linear functions of the expectation, ξ, and variance, θ^2, of the ensemble distributions, respectively. These assumptions are analogous to those of the NGR post-processing model (see equations (2.19) and (2.20)), and the LCR en-

semble adjustment scheme (see section 3.3) introduced in chapter 3. Using training samples of observations and ensemble forecasts, our task is to estimate the parameters a, b, c and d without making distributional assumptions, and subsequently to use these estimates to post-process out-of-sample ensemble forecasts.

4.2.2 The effect of noisy covariates

In section 2.2.4 we discussed two common interpretations of ensemble members, with a focus on the context of forecast verification. However, the possible interpretations of ensemble members has received almost no attention in the context of the statistical models used for ensemble post-processing. Typically, as is usually the case with regression models, the ensemble statistics, such as the mean and variance, which are used as covariates in the statistical post-processing models, are viewed as known, with the observations viewed as random variables. 'Known' model covariates are justified if the members of ensemble forecasts are interpreted as known constants, which (as described in section 2.3.2) is appropriate when interpreting the empirical distribution function (EDF) of the ensemble forecasts as probability distributions for the verifying observations, y. For example, the ensemble members might be the equidistant quantiles of the ensemble distribution. On the other hand, if the members of ensemble forecasts are interpreted as realisations of independent and identically distributed (IID) random variables, then the aforementioned sample statistics that are used as model covariates are also random variables. It therefore seems natural that, under this interpretation, the effects of random, rather than known model covariates are accommodated in the statistical models specified by ensemble post-processing methods. To our knowledge, the only instance in which an assumption of IID ensemble members has been included in the statistical model used for ensemble post-processing is in the derivation of the analytic expression used for the dressing kernel variance of the 'best member dressing' (BMD) method with Gaussian-distributed dressing kernels (Wang and Bishop [2005], see equation (2.9)). In this chapter, therefore, we consider both 'known' and 'random' model covariates.

Both interpretations of ensemble forecasts can be accommodated by the latent variable model introduced in the previous subsection. In the case of known covariates, we simply assume that the sample mean and variance of the ensemble forecast $\boldsymbol{x}_i$, given by $\bar{x}_i = M^{-1}\sum_{m=1}^{M} x_{im}$ and $s_i^2 = (M-1)^{-1}\sum_{m=1}^{M}(x_{im} - \bar{x}_i)^2$, respectively, are exactly equal to the expectation and variance of the corresponding ensemble distribution, so that

$$\bar{x}_i = \xi_i \tag{4.4}$$

$$s_i^2 = \theta_i^2, \tag{4.5}$$

for all i. On the other hand, interpreting the members of the ensemble forecast $\boldsymbol{x}_i$ as a collection of M realisations of IID random variables distributed according to the ensemble distribution results in covariates $\bar{x}_i$ and s_i^2 that are 'noisy', or imprecise estimates of the 'true', unobserved latent variables ξ_i and θ_i^2. Measurement error models are commonly employed to tackle the problems of mismeasured covariates in statistical modelling. In the following two subsections we derive parameter estimates for our distribution-free ensemble post-processing method, using the method of moments, both with and without the assumption of mismeasured covariates. Before doing so, however, it helps to introduce some notation that we use frequently throughout the remainder of this section.

In what follows, we make frequent use of the properties of conditional expectation — in particular, the result $E(A) = E\{E(A \mid B)\}$ for random variables A and B. We also make frequent use of the sample variance and covariance statistics between vectors $r = (r_1, r_2, \ldots, r_N)$ and $t = (t_1, t_2, \ldots, t_N)$, denoted S_r^2 and $S_{r,t}$, respectively. Specifically,

$$S_r^2 = \frac{1}{N-1} \sum_{i=1}^{N} (r_i - \bar{r})^2$$
$$S_{r,t} = \frac{1}{N-1} \sum_{i=1}^{N} (r_i - \bar{r})(t_i - \bar{t}),$$

where $\bar{r} = N^{-1} \sum_{i=1}^{N} r_i$ and $\bar{t} = N^{-1} \sum_{i=1}^{N} t_i$ are the sample means of the vectors r and t.

4.2.3 Parameter estimation for known covariates

In what follows, we assume that the 'error in the equation', ϵ_i, is independent of the latent variable ξ_i for all i. Then

$$\operatorname{cov}(\xi_i, \epsilon_i) = 0.$$

We now turn to a derivation of our parameter estimates. We first consider the situation in which we assume complete knowledge of ξ_i and θ_i^2, for all i. Denote by $\xi = (\xi_1, \xi_2, \ldots, \xi_N)$ and $\theta = (\theta_1, \theta_2, \ldots, \theta_N)$ the vector of realised values of the latent variables ξ and θ, and $\mathbf{y} = (y_1, y_2, \ldots, y_N)$, the vector of verifying observations y_i. To derive estimates for the model parameters a, b, c and d using the method of moments, we require at least four equations, referred to as 'moment equations', that are written in terms of the parameters a, b, c and d, and can be solved to give a unique solution for the estimators $\hat{a}, \hat{b}, \hat{c}$ and $\hat{d}$.

We consider the following sample statistics calculated over the training data:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i,$$

the sample mean of the observations, $\mathbf{y}$,

$$\overline{\xi} = \frac{1}{N}\sum_{i=1}^{N} \xi_i,$$

the sample mean of the realised values of the latent variable ξ,

$$S_\xi^2 = \frac{1}{N-1}\sum_{i=1}^{N} \left(\xi_i - \overline{\xi}\right)^2,$$

the sample variance of ξ and

$$S_{\xi,y} = \frac{1}{N-1}\sum_{i=1}^{N} \left(\xi_i - \overline{\xi}\right)(y_i - \overline{y}),$$

the sample covariance between the vectors ξ and $\mathbf{y}$. Referring to equation (4.1) for the expectation of observation y_i, conditional on ξ_i and θ_i, and conditioning on the realised values ξ and θ, we obtain the following relations:

$$\begin{aligned} E(\bar{y} \mid \xi, \theta) &= E(\overline{a + b\xi + \epsilon} \mid \xi, \theta) \\ &= a + b\overline{\xi} \qquad (4.6) \\ E(S_{\xi,y} \mid \xi, \theta) &= E(S_{\xi, a+b\xi+\epsilon} \mid \xi, \theta) \\ &= bS_\xi^2, \qquad (4.7) \end{aligned}$$

where here we have used the notation $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_N)$ to denote the vector of random variables ϵ_i. Equation (4.6) follows from the modelling assumption $E(\epsilon_i \mid \xi_i, \theta_i) = 0$, for all i. Therefore, rearranging equations (4.6) and (4.7), and replacing the expectations of the summary statistics with their realised values yields the following parameter estimates for a and b:

$$\hat{a} = \bar{y} - \hat{b}\overline{\xi} \qquad (4.8)$$

$$\hat{b} = \frac{S_{\xi,y}}{S_\xi^2}. \qquad (4.9)$$

Observe that these parameter estimates are equivalent to those that are obtained with the method of ordinary least squares when regressing the vector of observations $\mathbf{y}$ on the covariates ξ. Furthermore, if the observations y_i are (conditionally on ξ_i and θ_i) assumed to be normally distributed, then the parameter estimates are also

those that are obtained from minimisation of the negative log-likelihood (NLL).

In order to estimate the parameters c and d we use the following summary statistics:

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2,$$

the sample variance of the realised observations $\mathbf{y}$,

$$\overline{\theta^2} = \frac{1}{N}\sum_{i=1}^{N}\theta_i^2,$$

the sample mean of the realised values θ^2,

$$S_{\theta^2}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\theta_i^2 - \overline{\theta^2}\right)^2,$$

the sample variance of θ^2,

$$\overline{y^2} = \frac{1}{N}\sum_{i=1}^{N}y_i^2,$$

the sample mean of the squared observations, $\mathbf{y}^2$,

$$S_{\theta^2,y^2} = \frac{1}{N-1}\sum_{i=1}^{N}(\theta_i^2 - \overline{\theta^2})(y_i^2 - \overline{y^2}),$$

the sample covariance between the vectors θ^2 and $\mathbf{y}^2$,

$$S_{\xi,\theta^2} = \frac{1}{N-1}\sum_{i=1}^{N}(\xi_i - \overline{\xi})(\theta_i^2 - \overline{\theta^2}),$$

the sample covariance between the vectors ξ and θ^2,

$$\overline{\xi^2} = \frac{1}{N}\sum_{i=1}^{N}\xi_i^2,$$

the sample mean of the squared realisations of ξ, and

$$S_{\xi^2,\theta^2} = \frac{1}{N-1}\sum_{i=1}^{N}(\xi_i^2 - \overline{\xi^2})(\theta_i^2 - \overline{\theta^2}),$$

the sample covariance between the vectors ξ^2 and θ^2. Then, conditioning on ξ and θ and using the regression model $y_i = a + b\xi_i + \epsilon_i$ (see equation (4.3)), we obtain

the following relations for the expectations of the summary statistics.

$$\begin{aligned} E(S_y^2 \mid \xi, \theta) &= E(S_{a+b\xi+\epsilon}^2 \mid \xi, \theta) \\ &= E(S_{b\xi}^2 \mid \xi, \theta) + E(S_\epsilon^2 \mid \xi, \theta) \\ &= b^2 S_\xi^2 + c + d\overline{\theta^2}, \end{aligned} \tag{4.10}$$

where the result $E(S_\epsilon^2 \mid \xi, \theta) = c + d\overline{\theta^2}$ follows immediately from the equation for the (conditional) variance of ϵ_i given in equation (4.2), $\text{var}(\epsilon_i \mid \xi_i, \theta_i) = c + d\theta_i^2$. We also have

$$\begin{aligned} E(S_{\theta^2,y^2} \mid \xi, \theta) &= E(S_{\theta^2,(a+b\xi+\epsilon)^2} \mid \xi, \theta) \\ &= E(S_{\theta^2,(2ab\xi+2a\epsilon+2b\xi\epsilon+b^2\xi^2+\epsilon^2)} \mid \xi, \theta). \end{aligned} \tag{4.11}$$

Now observe that

$$\begin{aligned} E(S_{\theta^2,2a\epsilon} \mid \xi, \theta) &= E\left\{ \frac{1}{N-1} \sum_{i=1}^{N} (\theta_i^2 - \overline{\theta^2})(2a\epsilon_i - \overline{2a\epsilon}) \;\middle|\; \xi, \theta \right\} \\ &= \frac{2a}{N-1} \sum_{i=1}^{N} \left(\theta_i^2 - \overline{\theta^2}\right) \times E\left(\epsilon_i - \overline{\epsilon} \mid \xi, \theta\right) \\ &= 0, \end{aligned}$$

where the final line follows from the (conditional) expectation of the ϵ_i, $E(\epsilon_i \mid \xi, \theta) = 0$. Similar calculations yield $E(S_{\theta^2,2b\xi\epsilon} \mid \xi, \theta) = 0$. Thirdly, we have

$$\begin{aligned} E(S_{\theta^2,\epsilon^2} \mid \xi, \theta) &= \frac{1}{N-1} \sum_{i=1}^{N} \left(\theta_i^2 - \overline{\theta^2}\right) \times E\left(\epsilon_i^2 - \overline{\epsilon^2} \;\middle|\; \xi, \theta\right) \\ &= \frac{1}{N-1} \sum_{i=1}^{N} \left(\theta_i^2 - \overline{\theta^2}\right) \times \left\{(c + d\theta_i^2) - (c + d\overline{\theta^2})\right\} \\ &= \frac{d}{N-1} \sum_{i=1}^{N} \left(\theta_i^2 - \overline{\theta^2}\right)^2 \\ &= dS_{\theta^2}^2. \end{aligned}$$

Substituting these three results in to equation (4.11) yields

$$E(S_{\theta^2,y^2} \mid \xi, \theta) = 2abS_{\xi,\theta^2} + b^2 S_{\xi^2,\theta^2} + dS_{\theta^2}^2. \tag{4.12}$$

Rearranging equations (4.10) and (4.12), and replacing the expectations of the summary statistics with their observed values, we obtain method of moments estimates

for c and d:

$$\hat{c} = S_y^2 - \hat{d}\overline{\theta^2} - \hat{b}^2 S_\xi^2 \tag{4.13}$$

$$\hat{d} = \frac{S_{\theta^2,y^2} - 2\hat{a}\hat{b}S_{\xi,\theta^2} - \hat{b}^2 S_{\xi^2,\theta^2}}{S_{\theta^2}^2}. \tag{4.14}$$

Equations (4.8), (4.9), (4.13) and (4.14) give parameter estimates for the hypothetical case in which ξ_i and θ_i are known exactly, for all $i = 1, 2, \ldots, N$. In practice, ξ_i and θ_i^2 are estimated with the sample mean and variance of the ensemble forecast $\boldsymbol{x}_i$, namely $\bar{x}_i$ and s_i^2. Therefore, in the case of 'known' covariates, when the ensemble members are interpreted as constants and so we assume $\bar{x}_i = \xi_i$ and $s_i^2 = \theta_i^2$, for all i, estimates for the model parameters are obtained by substituting $\bar{x}_i$ and s_i^2 for ξ_i and θ_i^2:

$$\hat{a}_{\text{Known}} = \bar{y} - \hat{b}_{\text{Known}}\overline{\overline{x}} \tag{4.15}$$

$$\hat{b}_{\text{Known}} = \frac{S_{\overline{x},y}}{S_{\overline{x}}^2} \tag{4.16}$$

$$\hat{c}_{\text{Known}} = S_y^2 - \hat{d}_{\text{Known}}\overline{s^2} - \hat{b}_{\text{Known}}^2 S_{\overline{x}}^2 \tag{4.17}$$

$$\hat{d}_{\text{Known}} = \frac{S_{s^2,y^2} - 2\hat{a}_{\text{Known}}\hat{b}_{\text{Known}}S_{\overline{x},s^2} - \hat{b}_{\text{Known}}^2 S_{\overline{x}^2,s^2}}{S_{s^2}^2}, \tag{4.18}$$

where in the above equations the summary statistics are defined in the obvious way, by substituting $\bar{x}$ and s^2 in place of ξ and θ^2 (see the earlier overview of this notation).

4.2.4 Parameter estimation for mismeasured covariates

In this subsection we explain how so-called measurement error models have an intuitive interpretation in the context of ensemble post-processing, when the members of ensemble forecasts are interpreted as IID draws from underlying ensemble distributions. We refer the reader to Buonaccorsi [2010, chapters 4–6] for a detailed exposition of the material used to motivate this work. In particular, we make use of the idea of replication. It is intuitive to view the M members of an ensemble forecast, $\boldsymbol{x}_i$, as each representing a replicate, mismeasured estimate of the latent variable ξ_i. In this case, the ensemble mean $\bar{x}_i$ is an unbiased estimator of ξ_i. Furthermore, the sample variance, s_i^2, is an unbiased estimator for the latent variable θ_i^2. The idea of replication is used for estimating the parameters a, b, c and d of the statistical model given by equations (4.1) and (4.2) for model covariates that are subject to measurement error. In section 4.2.4.1 we describe our measurement error model, and note the assumptions that are used in the derivations of parameter

estimates that follow in section 4.2.4.2.

4.2.4.1 A measurement error model for ensemble post-processing

We denote by $u_{\xi,i}$ and $u_{\theta^2,i}$ the error in estimating the latent variables ξ_i and θ_i^2 with the ensemble mean $\bar{x}_i$ and ensemble variance s_i^2, respectively. We assume a so-called 'additive measurement error model', such that the expectations of $u_{\xi,i}$ and $u_{\theta^2,i}$, conditional on ξ_i and θ_i, are $E(u_{\xi,i} \mid \xi_i, \theta_i) = E(u_{\theta^2,i} \mid \xi_i, \theta_i) = 0$ for all i. This additive model satisfies the unbiasedness of the sample statistics $\bar{x}_i$ and s_i^2 as estimators of the expectation ξ_i and variance θ_i^2 of the ith ensemble distribution, under the interpretation that the members of the ensemble forecast $\boldsymbol{x}_i$ are IID draws from that distribution. The (mismeasured) model covariates, $\bar{x}_i$ and s_i^2, can thus be written as

$$\bar{x}_i = \xi_i + u_{\xi,i} \tag{4.19}$$

$$s_i^2 = \theta_i^2 + u_{\theta^2,i}, \tag{4.20}$$

for all i. We further denote the variances of the measurement error variables by $\sigma^2_{u_{\xi,i}}$ and $\sigma^2_{u_{\theta^2,i}}$. An important topic in the field of measurement error models, as we shall see below, is the estimation of these variances.

We assume mutual, unconditional independence between the 'error in the equation', ϵ_i, and the measurement error variables, $u_{\xi,i}$ and $u_{\theta^2,i}$, for all i. We further assume unconditional independence of the measurement error variables $u_{\xi,i}$ and $u_{\theta^2,i}$. A corollary of these two assumptions is that the measurement error variables $u_{\xi,i}$ and $u_{\theta^2,i}$ are (conditional on ξ_i and θ_i^2) independent of the verifying observations y_i, and the squared observations y_i^2. Finally, we assume unconditional temporal independence between the measurement error variables. That is, we assume $f(u_{\xi,i}, u_{\xi,j}) = f(u_{\xi,i}) \times f(u_{\xi,j})$ for all $i \neq j$, where $f(\cdot)$ denotes the probability density function (PDF) of its argument, and similarly for $u_{\theta^2,i}$ and $u_{\theta^2,j}$. We refer to these assumptions in the derivations of the parameter estimates in the calculations that follow in the next section.

The interpretation of ensemble forecasts $\boldsymbol{x}_i$ whose members are IID realisations of a random variable distributed according to the ith ensemble distribution, with expectation ξ_i and variance θ_i^2, is appealing, as it enables us to estimate the measurement error variances $\sigma^2_{u_{\xi,i}}$ and $\sigma^2_{u_{\theta^2,i}}$. Under the aforementioned interpretation of the ensemble members as IID samples from the ith ensemble distribution, with variance $\text{var}(x_{im}) = \theta_i^2$ for all $m = 1, 2, \ldots, M$, it follows immediately from equation (4.19)

that the (conditional) variance of $\bar{x}_i$ is

$$\text{var}(\bar{x}_i \mid \xi_i, \theta_i) = \sigma^2_{u_{\xi,i}} = \theta_i^2/M.$$

In our measurement error model, the latent variable θ_i^2 is unobserved. We therefore estimate $\sigma^2_{u_{\xi,i}}$ by replacing θ_i^2 with its sample estimate, s_i^2, to give the estimate

$$\hat{\sigma}^2_{u_{\xi,i}} = s_i^2/M. \tag{4.21}$$

Estimating the variances $\sigma^2_{u_{\theta^2,i}}$ of the measurement error variables $u_{\theta^2,i}$, is less straightforward. We turn to the method of jackknife resampling [Efron and Gong, 1983], which again exploits the interpretation of ensemble members as IID random variables. The jackknife method can be used to estimate the variance of any estimator in the following way. For a M-member ensemble forecast $\boldsymbol{x}_i$, with expectation $E(x_{im}) = \xi_i$ and $\text{var}(x_{im}) = \theta_i^2$, for $m = 1, 2, \ldots, M$, let $\boldsymbol{x}_{i,-m}$ denote the $(M-1)$-member ensemble that excludes member x_{im}. The jackknife estimate of $\sigma^2_{u_{\theta^2,i}} = \text{var}(s_i^2)$, where s_i^2 denotes the sample variance of ensemble forecast $\boldsymbol{x}_i$, is given by

$$\hat{\sigma}^2_{u_{\theta^2,i}} = \frac{1}{M}\sum_{m=1}^{M}\left(s^2_{i,-m} - s_i^2\right)^2, \tag{4.22}$$

where

$$s^2_{i,-m} = \frac{1}{M-2}\sum_{l=1;l\neq m}^{M}\left(x_{il} - \bar{x}_{i,-m}\right)^2$$

and

$$\bar{x}_{i,-m} = \frac{1}{M-1}\sum_{k=1;k\neq m}^{M} x_{ik}$$

are the sample variance and sample mean of $\boldsymbol{x}_{i,-m}$.

Under the interpretation of IID ensemble members, the estimator $\hat{\sigma}^2_{u_{\xi,i}} = s_i^2/M$ (equation (4.21)) is unbiased for $\sigma^2_{u_{\xi,i}}$. This follows from the fact that s_i^2 is an unbiased estimator of θ_i^2. Efron [1981] stated that the jackknife estimates of variance are biased upwards. We found this result to be in keeping with a small simulation experiment that we conducted with randomly generated data (not shown). We also experimented with estimating measurement error variances $\hat{\sigma}^2_{u_{\theta^2,i}}$ by bootstrap resampling. We found there to be little difference in the jackknife and bootstrap estimates, although the bootstrap comes at additional computational cost.

4.2.4.2 Parameter estimation with mismeasured covariates

Referring to the assumptions given in the previous section, and the estimates for the variances of the measurement error variables $u_{\xi,i}$ and $u_{\theta^2,i}$ (see equations (4.21) and (4.22)), we can now use the method of moments to find estimates for the model parameters a, b, c and d that account for mismeasured estimates of the latent variables ξ_i (by $\bar{x}_i$) and θ_i^2 (by s_i^2). Using conditional expectations and referring to equation (4.1) for the expectation of observation y_i, conditional on ξ_i and θ_i, we have

$$\begin{aligned} E(\bar{y}) &= E\{E(\bar{y} \mid \xi, \theta)\} \\ &= E\{E(\overline{a + b\xi + \epsilon} \mid \xi, \theta)\} \\ &= E(a + b\bar{\xi}) \\ &= a + b \cdot \frac{1}{N} \sum_{i=1}^{N} E(\xi_i). \end{aligned} \tag{4.23}$$

To derive a moment equation, therefore, we need an estimate of $E(\bar{\xi})$. Unlike the so-called 'known' estimates derived in the previous subsection, we do not observe the ξ_i, and so must use the estimate $\bar{x}_i$. This gives

$$\begin{aligned} E(\overline{\bar{x}}) &= E\{E(\overline{\bar{x}} \mid \xi, \theta)\} \\ &= \frac{1}{N} \sum_{i=1}^{N} E\{E(\bar{x}_i \mid \xi, \theta)\} \\ &= \frac{1}{N} \sum_{i=1}^{N} E(\xi_i), \end{aligned} \tag{4.24}$$

where equation (4.24) follows from the additive measurement error model ($E(\bar{x}_i \mid \xi_i, \theta_i) = E(\xi_i + u_{\xi,i} \mid \xi_i, \theta_i) = \xi_i$), for all i. Substituting equation (4.24) in to equation (4.23) yields the first moment equation,

$$E(\bar{y}) = a + bE(\overline{\bar{x}}). \tag{4.25}$$

To derive a second moment equation we have

$$\begin{aligned} E(S_{\bar{x},y}) &= E\{E(S_{\bar{x},y} \mid \xi, \theta)\} \\ &= E\{E(S_{\xi+u_\xi, a+b\xi+\epsilon} \mid \xi, \theta)\} \\ &= bE(S_\xi^2), \end{aligned} \tag{4.26}$$

where equation (4.26) follows from the (conditional on ξ_i and θ_i) independence of the ϵ_i with ξ_i and $u_{\xi,i}$, and $u_{\xi,i}$ with ξ_i, for all i. It remains to find an expression for

$E(S_\xi^2)$, where now the ξ_i are unknown and are estimated with $\bar{x}_i$. We have

$$\begin{aligned} E(S_{\bar{x}}^2) &= E\{E(S_{\xi+u_\xi}^2 \mid \xi, \theta)\} \\ &= E\{E(S_\xi^2 + S_{u_\xi}^2 \mid \xi, \theta)\} \\ &= E\left(S_\xi^2\right) + \frac{1}{N}\sum_{i=1}^{N} E\left(\frac{\theta_i^2}{M}\right), \end{aligned} \tag{4.27}$$

where the second term in equation (4.27) follows since the expectation $E(S_{u_\xi}^2 \mid \xi, \theta)$ is an unbiased estimate of the expectation of measurement error variances, $\sigma_{u_\xi}^2 = \theta_i^2/M$. Finally, we need an expression for $E(\overline{\theta^2})$. We have

$$\begin{aligned} E(\overline{s^2}) &= E\{E(\overline{s^2} \mid \xi, \theta)\} \\ &= E\{E(\overline{\theta^2 + u_{\theta^2}} \mid \xi, \theta)\} \\ &= E(\overline{\theta^2}) \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left(\theta_i^2\right), \end{aligned} \tag{4.28}$$

where the third line in the above working follows from the conditional expectation $E(u_{\theta^2,i} \mid \xi, \theta) = 0$. Substituting equations (4.27) and (4.28) in to equation (4.26) yields a second moment equation:

$$E(S_{\bar{x},y}) = b\left\{E(S_{\bar{x}}^2) - E(\overline{s^2})/M\right\}. \tag{4.29}$$

The above working leads to estimates for the model parameters a and b that account for measurement error. Replacing the expectations of the summary statistics given in equations (4.25) and (4.29) with their observed values, we arrive at

$$\hat{a}_{ME} = \bar{y} - \hat{b}_{ME}\overline{\overline{x}} \tag{4.30}$$

$$\hat{b}_{ME} = \frac{S_{\bar{x},y}}{S_{\bar{x}}^2 - \overline{s^2}/M}. \tag{4.31}$$

Observe that the estimates for the parameters a and b derived using the measurement error model (equations (4.30) and (4.31)) differ from the analogous parameter estimates under the interpretation of known covariates (equations (4.15) and (4.16)) in the denominator of the expression for $\hat{b}$. Using the measurement error model, we correct for the bias in estimating the variance of the latent variable ξ by $S_{\bar{x}}^2$, and instead use the unbiased estimate $S_{\bar{x}}^2 - \overline{s^2}/M$. Note, however, that this does not imply unbiasedness in the parameter estimate $\hat{b}_{ME}$ (equation (4.31)). This point is discussed further in section 4.2.5.

To derive estimates for the parameters c and d we need two further moment equa-

tions. We have

$$\begin{aligned} E(S_y^2) &= E\{E(S_{a+b\xi+\epsilon}^2 \mid \xi, \theta)\} \\ &= b^2 E\left(S_\xi^2\right) + c + d \cdot \frac{1}{N} \sum_{i=1}^{N} E\left(\theta_i^2\right), \end{aligned} \tag{4.32}$$

where equation (4.32) follows immediately from the earlier result for $E(S_y^2 \mid \xi, \theta)$ (when deriving the 'known' parameter estimates), given in equation (4.10). Rearranging equation (4.26) in terms of $E(S_\xi^2)$, and substituting equation (4.28) (for $E(\overline{s^2}) = E(\overline{\theta^2})$) in to equation (4.32) gives a third moment equation:

$$E(S_y^2) = bE(S_{\bar{x},y}) + c + dE(\overline{s^2}). \tag{4.33}$$

To derive a fourth moment equation we proceed as follows. Firstly, we have

$$\begin{aligned} E(S_{s^2,y^2}) &= E\{E(S_{s^2,y^2} \mid \xi, \theta)\} \\ &= 2abE(S_{\xi,\theta^2}) + b^2 E(S_{\xi^2,\theta^2}) + dE(S_{\theta^2}^2), \end{aligned} \tag{4.34}$$

where equation (4.34) follows immediately from equation (4.12) for $E(S_{\theta^2,y^2} \mid \xi, \theta)$, which was used in the derivation of the 'known' estimates, and the (conditional) independence of the measurement errors $u_{\theta^2,i}$ with the squared observations, y_i^2, as noted at the beginning of this section. It remains to find estimates for the terms on the right hand side of equation (4.34). We have

$$\begin{aligned} E(S_{\bar{x}^2,s^2}) &= E\{E(S_{(\xi+u_\xi)^2,\theta^2+u_{\theta^2}} \mid \xi, \theta)\} \\ &= E\{E(S_{\xi^2+2\xi u_\xi+u_\xi^2,\theta^2+u_{\theta^2}} \mid \xi, \theta)\} \\ &= E(S_{\xi^2,\theta^2}) + E(S_{\theta^2}^2)/M. \end{aligned} \tag{4.35}$$

The first term in equation (4.35) follows from the assumed (conditional) independence of $u_{\xi,i}$ with $u_{\theta^2,i}$ for all i. The second term follows from the result

$$\begin{aligned} E(S_{u_\xi^2,\theta^2} \mid \xi, \theta) &= \frac{1}{N-1} \sum_{i=1}^{N} \left(\theta_i^2 - \overline{\theta^2}\right) \times E\left(u_{\xi,i}^2 - \overline{u_\xi^2}\right) \\ &= \frac{1}{N-1} \sum_{i=1}^{N} \left(\theta_i^2 - \overline{\theta^2}\right)^2 / M \\ &= S_{\theta^2}^2/M, \end{aligned}$$

since $E(u_{\xi,i}^2 \mid \xi, \theta) = \sigma_{u_{\xi,i}}^2 = \theta_i^2/M$. Equation (4.35) can therefore be rearranged for $E(S_{\xi^2,\theta^2})$, the first of the terms needed on the right hand side of equation (4.34).

Secondly we have

$$\begin{aligned} E(S_{\bar{x},s^2}) &= E\{E(S_{\xi+u_\xi,\theta^2+u_{\theta^2}} \mid \xi,\theta)\} \\ &= E(S_{\xi,\theta^2}), \end{aligned} \tag{4.36}$$

where equation (4.36) follows from the assumed conditional independence of the measurement error variables, and the assumption that their conditional expectations are 0. Finally we have

$$\begin{aligned} E(S^2_{s^2}) &= E\{E(S^2_{\theta^2+u_{\theta^2}} \mid \xi,\theta)\} \\ &= E(E\{S^2_{\theta^2} + S^2_{u_{\theta^2}} \mid \xi,\theta\}) \\ &= E\left(S^2_{\theta^2}\right) + \frac{1}{N}\sum_{i=1}^{N} E(\sigma^2_{u_{\theta^2,i}}). \end{aligned} \tag{4.37}$$

The second term in equation (4.37) follows from expanding $S^2_{u_{\theta^2}}$, whereupon a short calculation shows that the expectation of this summary statistic, conditional on ξ and θ, is equal to the mean of the measurement error variances $\sigma^2_{u_{\theta^2,i}}$. Rearranging equations (4.35), (4.36) and (4.37), and substituting in to equation (4.34), we arrive at

$$E(S_{s^2,y^2}) = b^2 E(S_{\bar{x}^2,s^2}) + 2abE(S_{\bar{x},s^2}) + (d - b^2/M)\left[E(S^2_{s^2}) - N^{-1}\sum_{i=1}^{N} E(\sigma^2_{u_{\theta^2,i}})\right]. \tag{4.38}$$

We use the jackknife estimates $\hat{\sigma}_{u_{\theta^2,i}}$ (equation (4.22)) in place of $E(\sigma^2_{u_{\theta^2,i}})$ to give the fourth moment equation:

$$E(S_{s^2,y^2}) \approx b^2 E(S_{\bar{x}^2,s^2}) + 2abE(S_{\bar{x},s^2}) + (d - b^2/M)\left[E(S^2_{s^2}) - N^{-1}\sum_{i=1}^{N} \hat{\sigma}^2_{u_{\theta^2,i}}\right]. \tag{4.39}$$

The expectation on the left hand side of equation (4.39) is an approximation due to the bias of the jackknife estimates $\hat{\sigma}^2_{u_{\theta^2,i}}$.

Equations (4.33) and (4.39) can be solved to give estimates for the parameters c and d that account for measurement error as follows:

$$\hat{c}_{ME} = S^2_y - \hat{d}_{ME}\overline{s^2} - \hat{b}_{ME}S_{\bar{x},y} \tag{4.40}$$

$$\hat{d}_{ME} = \frac{\hat{b}^2_{ME} + M\left(S_{s^2,y^2} - \hat{b}^2_{ME}S_{\bar{x}^2,s^2} - 2\hat{a}_{ME}\hat{b}_{ME}S_{\bar{x},s^2}\right)}{M\left(S^2_{s^2} - N^{-1}\sum_{i=1}^{N}\hat{\sigma}^2_{u_{\theta^2,i}}\right)}. \tag{4.41}$$

4.2.4.3 Parameter estimate constraints

Due to the negative terms in the expressions for the 'known' and 'measurement error' parameter estimates for c and d, the parameter estimates themselves are not restricted to strictly positive values. Negative values of the estimates $\hat{c}$ or $\hat{d}$ have the undesirable implication that the forecast variance given by equation (4.2) may itself be negative. We therefore constrain the parameter estimates to be bounded below by 0 by imposing the following constraints. Firstly, for the 'known' parameter estimates, if $\hat{d}_{\text{Known}} < 0$ we set

$$\hat{d}_{\text{Known}} = 0 \tag{4.42}$$

$$\hat{c}_{\text{Known}} = S_y^2 - \hat{b}_{\text{Known}}^2 S_{\bar{x}}^2, \tag{4.43}$$

where equation (4.43) is simply equation (4.17) with the substitution $\hat{d}_{\text{Known}} = 0$. The estimate $\hat{c}_{\text{Known}}$ given by equation (4.43) is positive provided that $S_y^2 > (S_{\bar{x},y})^2/S_{\bar{x}}^2$ — this follows from the expression for $\hat{b}_{\text{Known}}$. We have not obtained any negative estimates $\hat{c}_{\text{Known}}$ from equation (4.43) in the three studies presented in section 4.3. However, if that occurs, we suggest that the user reverts to the more simplistic case of constant (rather than nonhomogeneous) variances in the forecast errors, and uses the unbiased estimator of the variance of the squared residuals for $\hat{c}_{\text{Known}}^2$:

$$\hat{c}_{\text{Known}} = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - \hat{a}_{\text{Known}} - \hat{b}_{\text{Known}}\bar{x}_i\right)^2}. \tag{4.44}$$

Equation (4.44) is the parameter estimate that would be obtained if we fitted a distribution-free post-processing method with the assumption of constant forecast variance, such as in the model output statistics (MOS) method described in section 2.4.4.1.

Similarly, if $\hat{c}_{\text{Known}} < 0$ we set

$$\hat{c}_{\text{Known}} = 0 \tag{4.45}$$

$$\hat{d}_{\text{Known}} = \frac{S_y^2 - \hat{b}_{\text{Known}}^2 S_{\bar{x}}^2}{\overline{s^2}}, \tag{4.46}$$

where equation (4.46) follows from rearranging equation (4.17) with the constraint $\hat{c}_{\text{Known}} = 0$. Again, we have not obtained negative parameter estimates for $\hat{d}_{\text{Known}}$ using equation (4.46). We suggest that negative estimates using equation (4.46) indicate misspecification of the forecast variance, and users should revert to the constant variance case using equation (4.44) and setting $\hat{d}_{\text{Known}} = 0$.

Similarly for the measurement error parameters, if $\hat{d}_{ME} < 0$ we set

$$\hat{d}_{ME} = 0 \tag{4.47}$$

$$\hat{c}_{ME} = S_y^2 - \hat{b}_{ME} S_{\bar{x},y}, \tag{4.48}$$

where equation (4.48) follows directly from equation (4.40) with the constraint $\hat{d}_{ME} = 0$. In an analogous manner to the 'known' parameter estimates, if equation (4.48) gives rise to negative estimates, the alternative estimate for $\hat{c}_{ME}$ under the assumption of constant error variance should be used

$$\hat{c}_{ME} = \sqrt{\frac{1}{N-2} \sum_{i=1}^{N} \left(y_i - \hat{a}_{ME} - \hat{b}_{ME} \bar{x}_i \right)^2 - \hat{b}_{ME}^2 \overline{s^2} / M}, \tag{4.49}$$

where the final term in the above equation removes the bias in estimating c^2 by the mean of the squared residuals that is induced by the measurement error in the covariates $\bar{x}_i$. Buonaccorsi [2010, page 86] suggests that, if the term under the square root is negative, the estimate $\hat{c}_{ME}$ should be set to 0, in which case the problem becomes one of deterministic prediction. This seems a somewhat unsatisfactory solution in the context of ensemble post-processing, since our post-processing method would yield post-processed deterministic, rather than ensemble forecasts. We have not witnessed any negative estimates from equation (4.48) in practice, and suggest that, if they do occur, more thought should be given to the specification of the post-processing model.

Finally, if $\hat{c}_{ME} < 0$ we set

$$\hat{c}_{ME} = 0 \tag{4.50}$$

$$\hat{d}_{ME} = \frac{S_y^2 - \hat{b}_{ME} S_{\bar{x},y}}{\overline{s^2}}, \tag{4.51}$$

where equation (4.51) follows from rearranging equation (4.40) with the constraint $\hat{c}_{ME} = 0$. As with the 'known' estimates, if equation (4.51) gives rise to negative values, we suggest setting $\hat{d}_{ME}$ to 0 and using equation (4.49) to estimate $\hat{c}_{ME}$. We stress, however, that we have not witnessed any negative estimates in the three case studies provided in this chapter, and suggest that, if they do occur, further thought should be given to the model specification.

4.2.5 The sampling properties of parameter estimates

The importance of accounting for measurement error is best motivated by considering the effect of treating error-prone variables as known constants, as is typically the

case in ensemble post-processing and regression problems in general. For example, consider the estimates $\hat{a}_{\text{Known}}$ and $\hat{b}_{\text{Known}}$ for a and b, given in equations (4.15) and (4.16). In the presence of measurement error, we see immediately that the statistic $S^2_{\overline{x}} = S^2_{\xi+u_\xi} > S^2_\xi$ is a biased estimate of $\text{var}(\xi)$, the (unconditional) variance of the latent variable ξ. Consequently, substituting $S^2_{\overline{x}}$ for S^2_ξ in equation (4.9) in the presence of measurement error will bias the estimate of b towards 0. Similar comments apply to the effect of ignoring measurement error in the estimation of θ^2 by s^2. Clearly, therefore, neglecting measurement error by using equations (4.15)–(4.18) may lead to biased parameter estimates.

Measurement error models are typically applied to more standard regression problems, in which the variance of the 'error in the equation', $\sigma^2_{\epsilon,i}$, is constant for all i. In this case, accounting for measurement error by using the estimates given in equations (4.30) and (4.31) reduces the bias of the 'known' parameter estimates $\hat{a}_{\text{Known}}$ and $\hat{b}_{\text{Known}}$ (see equations (4.15) and (4.16)). However, the 'random', or 'measurement error' parameter estimates $\hat{a}_{ME}$ and $\hat{b}_{ME}$ are not unbiased, although they are consistent — that is, they tend to the true values of a and b as the training sample size N tends to ∞.

To our knowledge, parameter estimation by the method of moments has not been applied to our nonhomogeneous regression model, either with or without accounting for measurement error. The sampling properties of the estimates $\hat{c}_{\text{Known}}, \hat{d}_{\text{Known}}, \hat{c}_{ME}$ and $\hat{d}_{ME}$ are unknown. From their derivations (see equations (4.17), (4.18), (4.40) and (4.41)), it is clear that the direction of bias is complex. For example, considering the 'known' parameter estimates, the estimate $\hat{d}_{\text{Known}}$ is a function of both $\hat{a}_{\text{Known}}$ and $\hat{b}_{\text{Known}}$, $\hat{b}^2_{\text{Known}}$, as well as the product $\hat{a}_{\text{Known}}\hat{b}_{\text{Known}}$. Similar comments apply to the 'random' parameters, for which the sampling properties also depend on the properties of the jackknife estimates $\hat{\sigma}_{u_{\theta^2,i}}$ given in equation (4.22). The complicated (and possibly intractable) analyses of the sampling properties of parameter estimates is beyond the scope of this chapter, the focus of which is to post-process ensemble weather forecasts in the absence of distributional assumptions for the verifying observations. We do, however, consider the sampling properties of the parameter estimates in a simulation study, presented in section 4.3.1.

4.2.6 Ensemble post-processing and related issues

4.2.6.1 Distribution-free ensemble post-processing

Having obtained parameter estimates $\hat{a}, \hat{b}, \hat{c}$ and $\hat{d}$ using either equations (4.15)–(4.18) or (4.30)–(4.41), the ensemble forecast $\boldsymbol{x}_t = (x_{t,1}, x_{t,2}, \ldots, x_{t,M})$, with ensem-

ble mean $\bar{x}_t$ and variance s_t^2, where the subscript t indexes the ensemble forecast in the out-of-sample test dataset, is post-processed by

$$\hat{x}_{t,m} = \hat{a} + \hat{b}\bar{x}_t + \sqrt{\hat{c} + \hat{d}s_t^2} \cdot \frac{x_{t,m} - \bar{x}_t}{s_t}, \text{ for } m = 1, 2, \ldots, M. \tag{4.52}$$

The post-processed ensemble forecast, $\hat{\boldsymbol{x}}_t$, has ensemble mean $\bar{\hat{x}}_t = \hat{a} + \hat{b}\bar{x}_t$ and ensemble variance $\hat{s}_t^2 = \hat{c} + \hat{d}s_t^2$. Standardising the ensemble forecasts such that the members have 0 expectation and unit variance in equation (4.52) allows the ensemble forecast to be rescaled (using the term $\sqrt{\hat{c} + \hat{d}s_t^2}$), and shifted in order to account for bias, using the term $\hat{a} + \hat{b}\bar{x}_t$. Observe, therefore, that equation (4.52) is in keeping with the LCR scheme introduced in section 3.3.2, where here the parameter d in the LCR scheme is replaced with the linear function $c + ds_t^2$. If the assumed model (see equations (4.1) and (4.2)) is correct, therefore, we should expect the post-processed ensemble forecasts to be more skilful than the initial, 'raw' ensemble forecasts.

4.2.6.2 A note on out-of-sample forecasting

If the forecast user only wishes to learn about the 'true' values of the model parameters, then the measurement error estimates are theoretically preferable due to their reduced bias. Perhaps surprisingly, however, the issue of which parameter estimates should be used in equation (4.52) to post-process ensemble forecasts is much less straightforward. At first sight, it might appear that the measurement error estimates should be used, given that we expect their bias to be less than that of the known parameter estimates. On the other hand, the known estimates are those that are optimal for the 'mismeasured', or 'noisy' model covariates, in the sense that they solve the moment equations for the noisy covariates. Given that the covariates $\bar{x}_t$ and s_t^2 in the test dataset are also subject to measurement error, the known parameter estimates may actually yield better predictions.

This problem is addressed in Buonaccorsi [2010, section 4.8], although only in the context of point predictions or, in the parlance of weather forecasting, deterministic forecasts, using standard linear regression models. For deterministic forecasts, it is concluded that the known parameter estimates should be used, provided that the out-of-sample predictor variables, here $\bar{x}_t$, are of the same structural form as the covariates in the training sample that is used for parameter estimation. However, the measurement error estimates are preferable, at least for deterministic forecasts, if the out-of-sample covariates are structurally different to those in the training sample [Buonaccorsi, 2010, section 4.8]. An example of such a change could be an increase in the number of ensemble members after parameter estimation, which would result in a change to the variance of the measurement error associated with $\bar{x}_t$. In this

case, the variances of the measurement error variables, $\sigma^2_{u_{\xi,i}}$ reduce, and tend to 0 in the limit as the ensemble size, M, tends to ∞. In the limiting case, therefore, the ensemble mean $\bar{x}_t$ tends to ξ_t, where the subscript t denotes an out-of-sample ensemble forecast, and so the measurement error parameter estimates, which are less biased than the 'known' parameter estimates, should be used for prediction. Perhaps a more likely scenario is that of changing atmospheric conditions, for which the distribution of the latent variable ξ may differ for the out-of-sample forecast, compared with the training sample used for parameter estimation.

In the same vein as the discussion given in section 4.2.5 for the sampling properties of the estimates for the parameters c and d, the effect of the choice of the 'known' or 'random' parameter estimates on the post-processing of ensemble forecasts, using equation (4.52), is unclear. It seems plausible to suggest that the 'known' parameter estimates may in general be preferable, since we should expect the measurement error properties of both the latent variables ξ and θ^2 in the out-of-sample forecasts to be of the same structural form as those in the training sample. Once again, however, changes in the properties of the out-of-sample forecasts compared with those of the training sample may result in a preference for the 'random' parameter estimates. In this study we compare the skill of ensemble forecasts post-processed using equation (4.52), using both the 'known' and 'random' parameter estimates.

4.2.6.3 Preserving the ensemble rank structure for multivariate forecasts

In section 2.4.6 we described the method of ensemble copula coupling (ECC, Schefzik et al. [2013]), which utilises information contained in the rank dependence structure of the 'raw' ensemble forecasts to improve forecast skill over multivariate domains, such as spatial fields. The method requires post-processed ensemble forecasts to be sampled from probability forecast distributions, and the members reordered so as to preserve the rank structure of the initial, 'raw' ensemble forecasts. An appealing property of our distribution-free post-processing method (see equation (4.52)) is that the ensemble rank structure is preserved and, therefore, the skill of the method in issuing ensemble forecasts of spatial fields can be directly compared with that of the ECC method.

4.2.7 A note on ensemble member dependence

In section 4.2.6.1 we stated that the mean and variance of the post-processed ensemble forecast $\hat{\boldsymbol{x}}_t$ are $\overline{\hat{x}}_t = M^{-1}\sum_{m=1}^{M}\hat{x}_{m,t} = \hat{a}+\hat{b}\bar{x}_t$ and $\hat{s}_t^2 = (M-1)^{-1}\sum_{m=1}^{M}\left(\hat{x}_{m,t}-\overline{\hat{x}}_t\right)^2 = \hat{c}+\hat{d}s_t^2$, respectively. However, while these sample results are correct, we point out

that our post-processing method given by equation (4.52) induces dependence between members of the recalibrated ensemble $\widehat{\boldsymbol{x}}_t$ through the use of $\bar{x}_t$ and s_t^2. Therefore, post-processing ensemble forecasts using equation (4.52) yields post-processed ensemble forecasts $\widehat{\boldsymbol{x}}_t$ whose members are dependent, even in the case of IID members of the 'raw' ensemble forecast $\boldsymbol{x}_t$.

In chapter 6 we provide a more detailed discussion of the implications of ensemble member dependencies on the commonly employed verification measures. In particular, we show that verification measures such as the continuous ranked probability score are affected by the strength of inter-member dependencies, which are a function of the parameter estimates and covariates used in the ensemble post-processing. As discussed in section 2.2.4, however, when performing forecast verification to assess forecast skill it is necessary to interpret the ensemble members as either IID realisations of an underlying distribution, or as the empirical distribution function of the verifying observation. In this chapter we comment on both interpretations.

4.2.8 Forecast verification

In the following section we present results for three case studies in which the skill of ensemble forecasts post-processed with our distribution-free method (see equation (4.52)) are compared with those sampled from Gaussian NGR probability forecast distributions. We make frequent use of rank histograms, described in section 2.5.2.3, that provide a graphical assessment of the calibration of the post-processed ensemble forecasts. We also assess the skill of the post-processed ensemble forecasts using the empirical version of the continuous ranked probability score (CRPS, see equation (2.48) on page 60), that is appropriate when the EDFs of the ensemble forecasts are interpreted as probability forecast distributions for the verifying observations, and the fair CRPS (FCRPS, see equation (2.50) on page 61), that is appropriate when interpreting ensemble members as IID draws from underlying ensemble distributions. The energy score (see equation (2.47) on page 59) is used in section 4.3.3 for assessing ensemble forecasts over spatial fields. Finally, the skill of post-processed deterministic forecasts is also assessed by the mean squared error (MSE), where the deterministic forecasts are given by the post-processed ensemble mean

$$\overline{\hat{x}}_t = \frac{1}{M} \sum_{m=1}^{M} \hat{x}_{m,t},$$

where $\hat{\boldsymbol{x}}_t$ denotes the post-processed ensemble forecast.

4.3 Case studies

4.3.1 A simulation experiment

In this subsection we present results of a simulation experiment for which the moments of the ensemble forecasts and observations satisfy the relationships given by equations (4.1) and (4.2) exactly. Furthermore, we simulate observations that are normally distributed, conditional on the ensemble mean and variance, such that the statistical model assumed by the NGR post-processing method (see equations (2.19) and (2.20) in page 38) is also satisfied. As well as quantifying the skill of our distribution-free ensemble post-processing methods in out-of-sample forecasts, this experiment allows us to study the sampling properties of the moment-based parameter estimates. We also compare our distribution-free approach to the popular NGR post-processing method, where the NGR parameters minimise the negative log-likelihood (NLL). The experiment proceeds as follows.

1. Fix a training sample size, N, an ensemble size, M, and the 'true' values of the model parameters a, b, c and d.

2. Simulate N random variables ξ_i and θ_i^2 for $i = 1, 2, \ldots, N$. These represent the 'true' expectation and variance of the N underlying 'ensemble distributions'.

3. For each pair (ξ_i, θ_i^2), simulate M ensemble members, x_{im}, as IID random draws with distribution $N(\xi_i, \theta_i^2)$.

4. For all i, calculate the ensemble mean $\bar{x}_i$ and variance s_i^2.

5. For all i, simulate observations y_i as IID random draws with distribution $N(a + b\xi_i, c + d\theta_i^2)$.

6. Calculate the 'known' and 'random' parameter estimates.

7. Simulate random variables ξ_0, θ_0^2, and an ensemble $\boldsymbol{x}_0$, as described above. Calculate the statistics $\bar{x}_0$ and s_0^2.

8. Use the parameter estimates and equation (4.52) to post-process the ensemble $\boldsymbol{x}_0$. Calculate verification measures for the post-processing ensemble forecast $\hat{\boldsymbol{x}}_0$ and verifying observation y_0.

9. Repeat the above steps a sufficient number of times in order to achieve the desired accuracy of results.

In the following results, we use the parameter values $a = -2, b = 5/4, c = 1/2$ and $d = 3/2$, and an ensemble size of $M = 10$. We show results for samples of

size $N = 50$, but also comment on larger training samples. Results are calculated from 10^6 simulations, in order to ensure that the effects of sampling variability are negligible. The values ξ_i were simulated as N IID draws with distribution $N(1/2, 4)$, and the values of θ_i^2 were taken as the absolute values of IID realisations with the same distribution. The qualitative features of the results were found to be consistent, regardless of the marginal distributions of the random variables ξ and θ^2 — we experimented with exponential and uniform distributions for the ξ_i, and Chi-squared distributions for θ_i^2.

We compare the skill of ensemble forecasts that are post-processed with either the 'known' or 'measurement error' parameter estimates using our distribution-free method (see equation (4.52)) with ensemble forecasts sampled from NGR probability forecast distributions. Recall that the NGR forecast distributions for verifying observation y_t, conditional on ensemble forecast $\boldsymbol{x}_t$, is

$$y_t \mid \boldsymbol{x}_t \sim N(a + b\bar{x}_t, c + ds_t^2),$$

where in practice we substitute parameter estimates $\hat{a}, \hat{b}, \hat{c}$ and $\hat{d}$ for a, b, c and d. In this simulation experiment and the study presented in the following subsection, we consider ensemble forecasts that are sampled from the NGR forecast distributions as equidistant quantiles $q_{m/(M+1)}$ for $m = 1, 2, \ldots, M$, and as IID random draws. As well as constructing NGR forecast distributions (and subsequently sampling M-member ensemble forecasts) using likelihood parameter estimates, we also consider NGR forecasts with parameter estimates given by the 'known' and 'measurement error' method of moments approach. We therefore compare the forecasting skill of five ensemble forecasting systems — two ensemble forecasts post-processed with our distribution-free method, and three ensemble forecasts that are sampled from NGR probability distributions, that differ in the parameter estimates for a, b, c and d.

4.3.1.1 Sampling properties of parameter estimates

It is illuminating to first consider the sampling properties of the parameter estimates obtained in this simulation experiment. In figure 4.1 we examine the sampling distributions of the 'known' and 'random' parameter estimates obtained using the method of moments, as well as the NLL parameter estimates for the NGR model, using box and whisker plots. Parameter estimates were obtained from training samples of size $N = 50$. Firstly, considering the estimates $\hat{a}$ and $\hat{b}$, we see that the 'known' and NGR NLL estimates exhibit some bias, which is largely corrected by the measurement error model, in keeping with theoretical expectations (see the discussion in section 4.2.5).

The box and whisker plots indicate that the estimation of c and d using the method of moments is more problematic. For the estimation of c, the 'random' parameter estimates derived using the measurement error model are less biased than the 'known' parameter estimates — the bias of the 'random' parameter estimates is comparable with that of the NLL parameter estimates obtained using the NGR post-processing method, although the 'random' estimates are more variable. Both of the moment-based parameter estimates are more biased and variable than the corresponding NLL estimates for d. Unfortunately, as shown by the box and whisker plots, a significant proportion of the moment-based parameter estimates for c and d have been set to 0, in accordance with the parameter constraints provided at the end of section 4.2.4.2. Indeed, from further investigations (not shown) we see that in most instances one or other of the estimates has been set to 0, meaning that the initial estimate was negative. This is a result of the remaining estimation bias discussed in section 4.2.5.

However, we stress that the problematic estimation of the parameters c and d is not restricted to the method of moments. Both the Nelder-Mead and BFGS algorithms (see section 2.4.7.3) often fail to converge when estimating the equivalent NGR model parameters by NLL minimisation. In order to ensure the numerical convergence of the NLL parameter estimates, it is necessary to set $c = \gamma^2$ and $d = \delta^2$, and to find the estimates $\hat{\gamma}$ and $\hat{\delta}$. In turn, NLL parameter estimates for c and d are then given by $\hat{c}_{NGR} = \hat{\gamma}^2$ and $\hat{d}_{NGR} = \hat{\delta}^2$. Similar comments were made in the founding paper for the NGR post-processing method [Gneiting et al., 2005]. Without this constraint, even in the case of numerical convergence of the BFGS algorithm, we observe some estimates $\hat{c}_{NGR}$ and $\hat{d}_{NGR}$ that are negative, giving rise to the undesirable occurrence of a negative forecast variance. Similar comments apply to the other examples given in sections 4.3.2 and 4.3.3 that follow.

For larger training samples, the equivalent box and whisker plots (not shown) indicate reduced variability in the parameter estimates, as we should expect, and the above comments concerning the distribution-free, moment-based models still apply. We find that the Nelder-Mead and BFGS algorithms converge without applying the constraints $c = \gamma^2$ and $d = \delta^2$. However, such large training samples are unavailable in most practical settings.

4.3.1.2 Out-of-sample forecasting results

Figure 4.2 shows rank histograms for the post-processing methods under consideration. Again, parameter estimation is performed using training samples of size $N = 50$. Perhaps surprisingly, all of the rank histograms are non-uniform, despite the idealised underlying model for the NGR post-processing method. The shape of the rank histogram for the NGR forecasts with NLL parameter estimates (NGR-

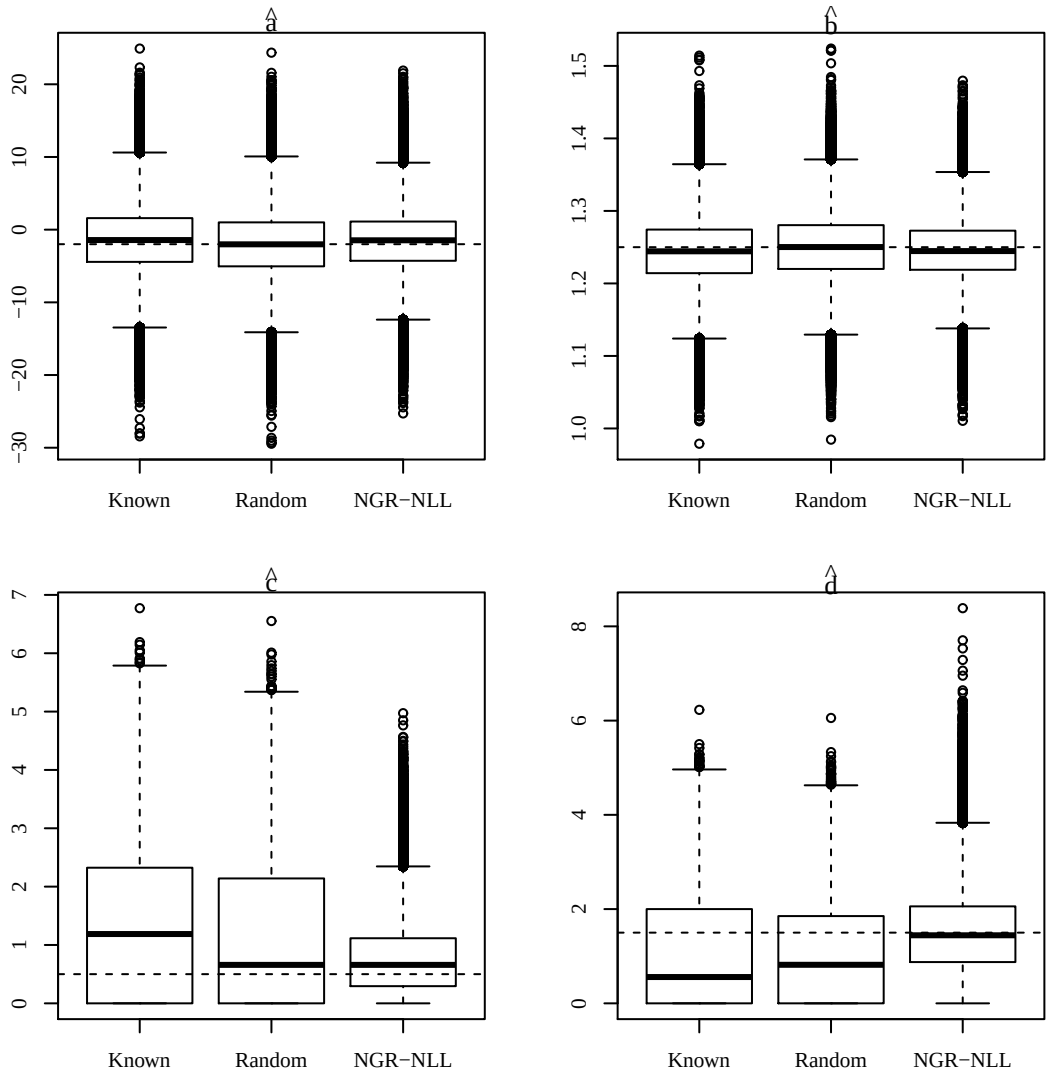

Figure 4.1 Box and whisker plots for the parameter estimates $\hat{a}, \hat{b}, \hat{c}$ and $\hat{d}$ for the 'known' method (left), the measurement error method (middle) and the likelihood-based NGR estimates (right).

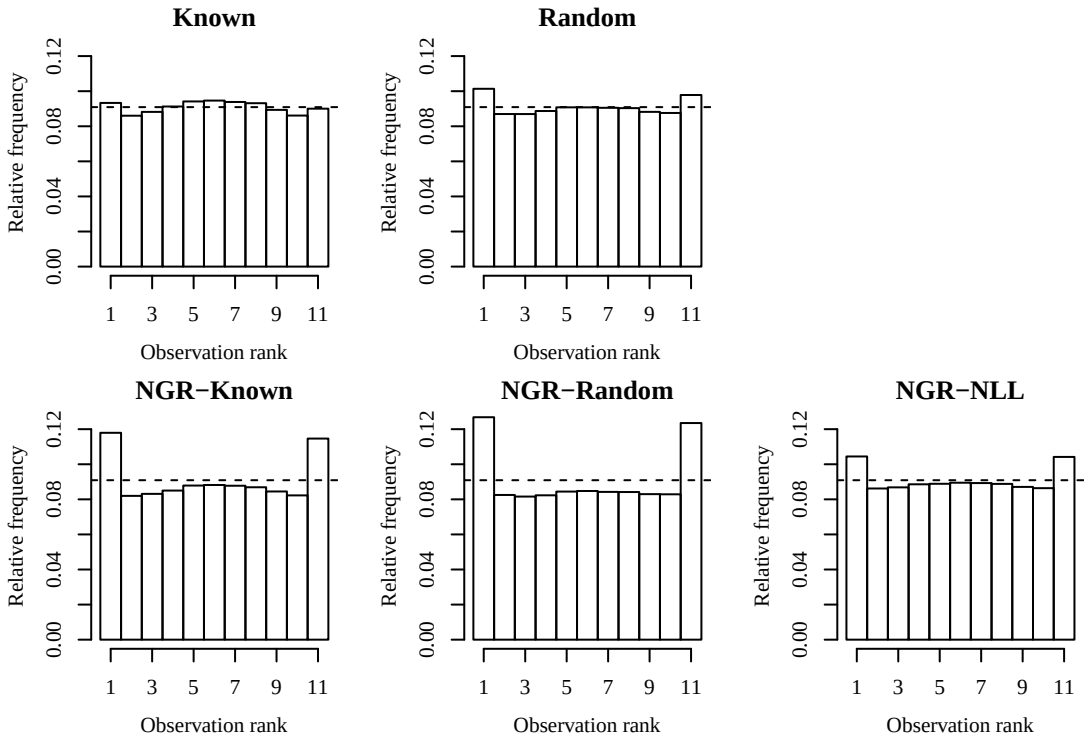

Figure 4.2 Rank histograms for the distribution-free post-processed ensemble forecasts (top row), and ensemble forecasts sampled as equidistant quantiles from NGR distributions using the known, measurement error, and NLL parameter estimates (bottom row). The training sample size is $N = 50$. The horizontal lines indicate the bin heights of uniform histograms.

NLL) is thought to be due to the effects of uncertainty in the parameter estimates, which is studied in detail in chapter 5. In this chapter, we simply focus on determining which of the ensemble forecasts yield the most uniform rank histograms, in addition to the other scores discussed below. To that end, the post-processed ensemble forecasts using equation (4.52), with the 'known' parameter estimates (equations (4.15)–(4.18)) yields the most uniform rank histogram. The rank histogram for the 'random' parameters derived with the measurement error model is qualitatively similar, except that the overpopulation of the outermost bins is exacerbated. For NGR, the likelihood-based parameter estimates (NLL) clearly improve the calibration of the post-processed ensemble forecasts, compared with those of ensemble forecasts sampled from NGR distributions specified with either of the moment-based parameter estimates. The rank histogram for the ensemble forecasts sampled from the NGR-NLL distributions is similar to that of the distribution-free post-processing method with the 'random' parameter estimates. Ensemble forecasts sampled as IID random values from the NGR forecast distributions yield almost identical rank histograms (not shown) to the approach of sampling equidistant quantiles.

Figure 4.3 shows the rank histograms for parameter estimates obtained with training samples of size $N = 500$. In this case, the rank histograms for the NGR method with NLL parameter estimates and the distribution-free method with 'random' parameter

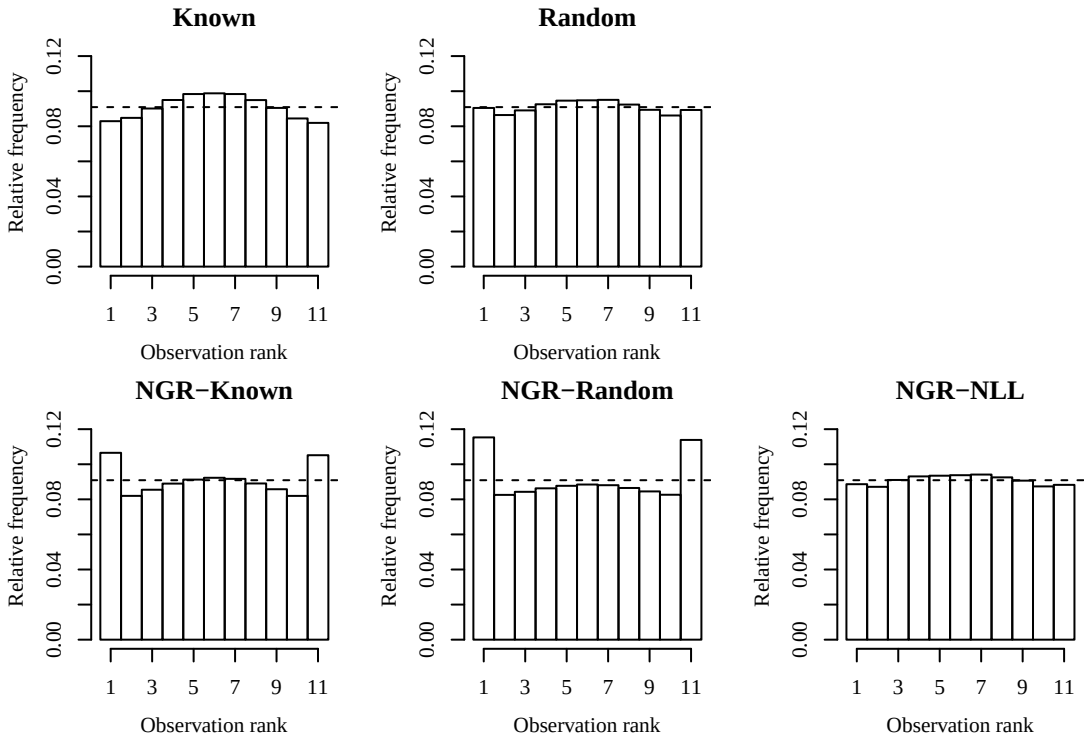

Figure 4.3 As for figure 4.2, but with training samples of size $N = 500$.

estimates are most uniform, while in this case the 'known' parameter estimates yield rank histograms that indicate overdispersed ensemble forecasts. Once again, when using the moment-based parameter estimates, ensembles obtained using the distribution-free post-processing method (equation (4.52)) yield more uniform rank histograms than are obtained by sampling from the Gaussian NGR distributions.

In table 4.1 we show various scores to assess the skill of ensemble forecasts post-processed with the distribution-free method (equation (4.52)) and sampled as equidistant quantiles from Gaussian NGR probability distributions. We show results for both the continuous ranked probability score (CRPS) for ensemble forecasts that we interpret as the empirical distribution functions (EDFs) of the observations (see equation (2.48)), and the fair CRPS (FCRPS) for ensemble forecasts that we interpret as IID draws from ensemble distributions (see equation (2.50)). See the discussion given in section 2.2.4 for further details. The 'Coverage' scores denote the proportion of observations that are bounded by the lower and upper ensemble members. For 10-member ensembles, the range of the ensemble members forms a nominal $100 \times 9/11 \approx 81.8\%$ prediction interval for the corresponding observation. Similar coverage scores to this value are indicative of well-calibrated ensemble forecasts. The scores were calculated as the mean of 10^6 independent simulations. The associated standard errors are sufficiently small (of the order 10^{-4} or 10^{-5}) to allow us to report the scores to at least the degree of accuracy shown.

In order to ease reading, we have only shown the scores for ensemble forecasts sam-

		known	Random	NGR-Known	NGR-Random	NGR-NLL
$N = 50$	MSE	2.533	2.535			2.517
	CRPS	0.917	0.918	0.900	0.904	0.888
	FCRPS	0.832	0.836	0.827	0.834	0.813
	Coverage	0.817	0.801	0.767	0.750	0.791
$N = 500$	MSE	2.432	2.433			2.430
	CRPS	0.894	0.894	0.875	0.878	0.863
	FCRPS	0.808	0.811	0.801	0.806	0.786
	Coverage	0.835	0.820	0.788	0.771	0.823

Table 4.1 Results of the simulation study with parameters estimated from training samples of size 50 and 500. Measures of the skill of deterministic forecasts (MSE) and ensemble calibration (CRPS, FCRPS and Coverage), for the two distribution-free post-processing methods (Known and Random), and ensemble forecasts sampled from NGR distributions using the moment-based parameter estimates (NGR-Known and NGR-Random) and likelihood-based estimates (NGR-NLL).

pled as equidistant quantiles from the NGR probability forecast distributions. We do not show the corresponding scores for ensemble forecasts that are sampled as IID realisations of the probability distributions. However, we stress that such sampling results in significantly worse scores than the approach of sampling equidistant quantiles, as we expect (see section 2.5.3 and Bröcker [2012]). The IID sampling scheme yields ensemble forecasts whose CRPS and FCRPS scores are significantly worse than both the equidistant quantiles approach for NGR and the distribution-free post-processing methods. This highlights the importance of the choice of the scheme used for sampling ensemble forecasts from probability distributions, and the possible implications of this choice on the conclusions that forecasters and users make in regards to determining which of several post-processing methods is best. For example, suppose that a forecaster was told to issue ensemble forecasts whose members truly are IID draws from NGR probability distributions. In this case, under both interpretations of ensemble forecasts (as the EDF of the observation and as random samples), we would conclude from the results presented in this section that the ensemble forecasts post-processed with the distribution-free method using the method of moments parameter estimates are more skilful than ensemble forecasts sampled from NGR probability distributions. However, if (as in this example) the forecaster is free to choose their sampling scheme, they can 'hedge' by sampling the equidistant quantiles of the NGR probability distributions. In this case, under both interpretations, the scores then indicate that ensemble forecasts sampled from NGR distributions are more skilful than the ensemble forecasts post-processed with the distribution-free method.

We now turn to a comparison of the scores of the various post-processed ensemble forecasts presented in table 4.1. The MSE scores indicate that using the NLL parameter estimates from the NGR model yields deterministic forecasts that are

slightly more skilful than those given by the distribution-free, moment-based parameter estimates. Under both interpretations of the ensemble forecasts, the CRPS and FCRPS scores also suggest that the NLL parameters yield more skilful recalibrated ensemble forecasts when sampling equidistant quantiles from the resulting Gaussian probability distributions, as discussed above. It is interesting, however, to note that for the smaller training samples ($N = 50$), the coverage of ensemble forecasts post-processed with the distribution-free method (denoted 'known' and 'Random'), improves on that of the NGR ensemble forecasts, for both the moment-based and NLL parameter estimates. For the larger training samples ($N = 500$), the coverage scores indicate that the prediction intervals given by both the ensemble forecasts post-processed with the distribution-free method, as well as ensemble forecasts sampled from NGR probability distributions with NLL parameter estimates, are slightly too wide. However, the prediction intervals given by ensemble forecasts sampled from NGR distributions with the method of moments parameter estimates remain too narrow. At first sight there is an apparent contradiction in the CRPS and coverage scores of the distribution-free forecasts, and the NGR forecasts using method of moments parameter estimates — the coverage scores are preferable for the distribution-free forecasts, while the CRPS values indicate that the NGR forecasts are preferable. This is explained by the choice of equidistant quantiles when sampling from the NGR Gaussian distributions, which are almost optimal for the CRPS [Bröcker, 2012], and the complicated effect on the CRPS that is observed when inducing dependence between ensemble members with equation (4.52) in our distribution-free model. This latter point was mentioned briefly in section 4.2.7, and is illustrated more fully in section 6.2. These remarks also apply to the results presented in the next subsection.

In conclusion, it is unclear from the results of this simulation study as to which of the post-processing methods is most desirable. On the one hand, the distribution-free, moment-based methods yield the most uniform rank histograms, at least for small training samples that are likely to be encountered in practice. The scores presented in table 4.1, however, suggest that sampling equidistant quantiles from Gaussian NGR distributions yields more skilful ensemble forecasts as measured by the CRPS and FCRPS, and deterministic forecasts as measured by the MSE. In combination with the coverage scores for training samples of size $N = 50$, our results indicate that the NGR ensemble forecasts are sharper than the 'known' and 'random' ensembles given by the distribution-free, moment-based post-processing method, but are less well calibrated. For NGR, the NLL parameter estimates certainly improve the skill of the ensemble forecasts, compared with those sampled from NGR probability distributions specified by the method of moments estimates. The NLL estimates also have more desirable properties, as discussed earlier — in general the NLL parameter estimates are less variable than both the 'known' and 'random'

parameter estimates, are less biased than the 'known' estimates, with similar bias to the 'random' estimates, at least under the constraints $\hat{c}_{NGR} = \gamma^2$ and $\hat{d}_{NGR} = \delta^2$.

4.3.1.3 Other remarks

Earlier in this subsection, we stated that the form of the NGR rank histograms is influenced by uncertainty in the parameter estimates. This effect is particularly evident for smaller training samples, as we discuss in chapter 5. Similarly, the method of moments parameter estimates are also subject to uncertainty. In chapter 5 we demonstrate the effect of uncertainty in the parameter estimates on the shape of rank and PIT histograms. Furthermore, in section 4.2.7, we highlighted that the distribution-free post-processing of ensemble forecasts using equation (4.52) (see page 116) induces dependence between the ensemble members, even if the original, raw ensemble forecasts are truly IID random samples. In chapter 6, we illustrate how the form of ensemble member dependence can affect the shape of the associated rank histograms. These two artefacts, combined with the earlier discussion of the effect of the scheme used to sample ensemble forecasts from NGR probability distributions on the verification scores, highlights the complexity of verifying post-processed ensemble forecasts, and the need to choose an interpretation for the ensemble forecasts as either the EDF of the corresponding observation, or as IID samples from an underlying ensemble distribution.

4.3.2 Distribution-free post-processing in the Lorenz 1996 system

In this subsection we present the results of a study using the Lorenz 1996 system, described in section 2.6.1 and also used in chapter 3. Again using the first dataset described in section 2.6.1, 500 sets of parameter estimates were obtained using training samples of size 100 and 1000, where the training samples were disjoint, and with the samples of size 100 forming the first 100 instances of the larger training samples. Forecast verification was then performed using the test dataset that comprises 190 000 forecasts and observations that are effectively independent.

We first comment briefly on the sampling properties of the parameter estimates (not shown). For all forecast lead times considered, the method of moments and NLL parameter estimates for a and b are similar, although the NLL estimates are slightly less variable. The differences between the 'known' and 'random' estimates, which result from the measurement error model, are small. The qualitative features of the likelihood-based (NLL) parameter estimates for c and d under the NGR model

differ from the moment-based parameter estimates. The NLL parameter estimates are less variable, and suggest a stronger spread-skill relationship (see section 2.4.1 for a discussion) between the ensemble variance and the mean forecast errors — the NLL parameter estimates $\hat{d}$ for d are on average larger than their moment-based ('known' or 'random') counterparts. The 'random' estimates $\hat{d}$ are on average slightly larger than the 'known' estimates, and the opposite is the case for the estimates $\hat{c}$. As with the simulation experiment we find that a significant proportion of the moment-based estimates for c or d have been set to 0, meaning that the initial estimate was negative.

We now turn to an assessment of the skill of out-of-sample forecasts. In table 4.2 we present the MSE, CRPS, FCRPS and Coverage scores for the distribution-free and NGR-based ensemble forecasts, for forecast lead times of $1, 3$ and 5. Parameters were estimated with training samples of size 100. Parameter estimation with larger training samples (not shown) results in small quantitative improvements to the scores, although the qualitative features are unchanged. Again, ensemble forecasts for the three NGR cases are sampled as equidistant quantiles from the Gaussian forecast distributions, using either the 'known' or 'random' parameter estimates calculated with the method of moments, or the likelihood-based parameter estimates (NLL). In order to ease reading, we do not show the corresponding scores for ensemble forecasts that are sampled as IID random draws from the NGR forecast distributions. However, as for the simulation study detailed in the previous subsection, we stress that this sampling scheme results in ensemble forecasts whose CRPS and FCRPS values are significantly worse than both the equidistant quantile approach used here, and ensemble forecasts post-processed with the distribution-free method (see equation (4.52)). Once again, this highlights the importance of the choice of sampling scheme on the conclusions that are drawn from this comparison — forecast users would justifiably conclude that the distribution-free post-processing method yields more skilful ensemble forecasts than those sampled from NGR forecast distributions, if the forecaster was required to issue ensemble forecasts whose members were IID draws from the NGR forecast distributions.

The MSE scores show that, in this example, the two moment-based parameter estimates yield deterministic forecasts that are similar in skill. However, the NLL parameter estimates under the NGR model result in slight improvements to the MSE scores for all lead times considered, in keeping with our findings in the simulation experiment presented in the previous subsection.

At forecast lead times 3 and 5, the CRPS and FCRPS values suggest that the 'known' parameter estimates yield post-processed ensemble forecasts that are slightly more skilful than those given by the 'random' parameter estimates derived with the measurement error model, under both interpretations of ensemble forecasts (see the

		Distribution-free		NGR		
		known	Random	known	Random	NLL
Lead time 1	MSE	0.070	0.070			0.069
	CRPS	0.153	0.153	0.150	0.150	0.148
	FCRPS	0.139	0.139	0.138	0.138	0.135
	Coverage	0.822	0.814	0.773	0.763	0.807
Lead time 3	MSE	0.950	0.950			0.948
	CRPS	0.539	0.539	0.528	0.530	0.513
	FCRPS	0.490	0.492	0.486	0.489	0.466
	Coverage	0.792	0.775	0.746	0.727	0.819
Lead time 5	MSE	4.936	4.936			4.897
	CRPS	1.234	1.234	1.187	1.190	1.156
	FCRPS	1.120	1.125	1.086	1.093	1.047
	Coverage	0.833	0.806	0.799	0.770	0.842

Table 4.2 Measures of the skill of deterministic forecasts (MSE) and ensemble calibration (CRPS, FCRPS and Coverage) for the distribution-free and NGR post-processing methods with moment-based parameter estimates (Known and Random), and likelihood-based estimates (NLL), for post-processed ensemble forecasts in the Lorenz 1996 system.

earlier discussion in the previous subsection and section 2.2.4). As was the case for the simulation study, at all lead times considered, sampling ensemble members as equidistant quantiles of the NGR forecast distributions yields CRPS and FCRPS scores that improve on those of the distribution-free post-processing method, particularly for those forecast distributions specified by the likelihood-based (NLL) parameter estimates. However, the coverage scores indicate that the width of the prediction intervals given by the distribution-free post-processed ensemble forecasts is closer to the nominal value of 0.818 than the NGR-NLL forecasts at lead times 1 and 5, although the opposite is the case at lead time 3. Once again, we note the significant miscalibration in the width of the prediction intervals given by ensemble forecasts sampled from NGR probability distributions with the method of moments parameter estimates, which are significantly too narrow.

In figures 4.4 and 4.5 we show the rank histograms for the raw ensemble forecasts and the post-processing methods displayed in table 4.2, at forecast lead times 3 and 5. All post-processing methods improve on the raw ensemble forecasts, which indicate significant bias and underdispersion. However, the resulting rank histograms remain non-uniform. As noted in the previous subsection, we suspect that the pattern of outer bins that are overpopulated compared with the shape of the inner bins is a result of parameter uncertainty, a topic addressed in chapter 5. Furthermore, there is also evidence of remaining bias in the location of the ensemble forecasts, indicated by comparing the upper and lower tails of the rank histograms. The rank histograms for the 'Random' parameter estimates, obtained using the measurement error model, display increased overpopulation of the outermost bins compared with the 'known' parameter estimates. We suggest this is due to additional sources

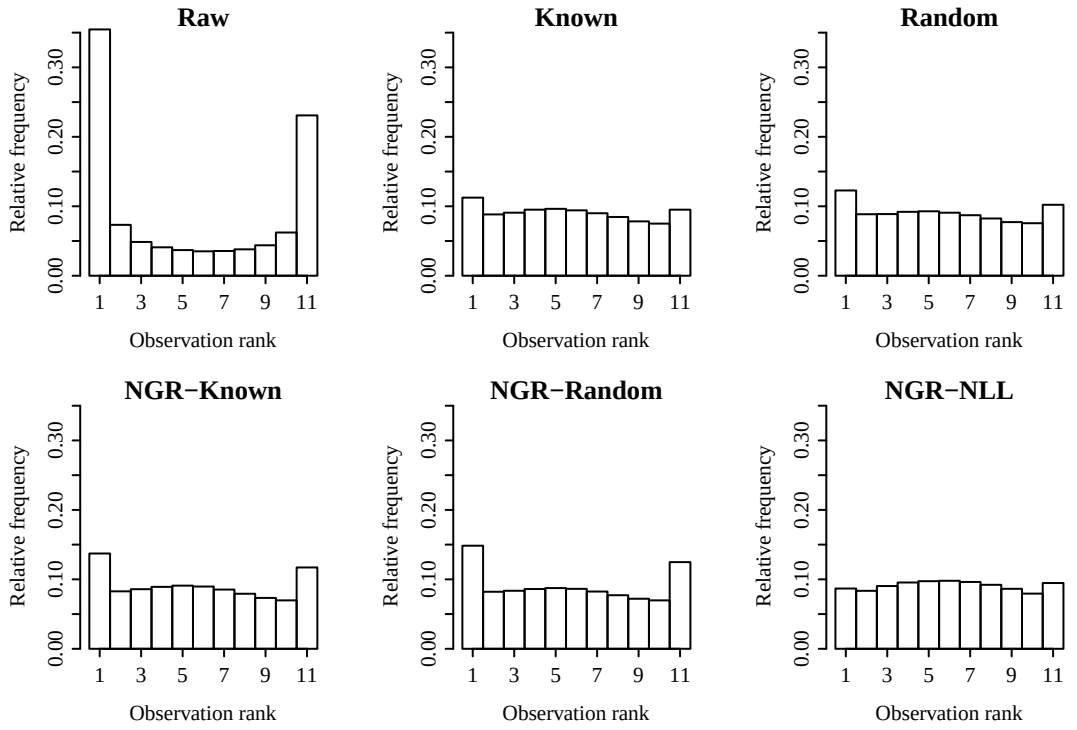

Figure 4.4 Rank histograms for the Raw ensemble forecasts, the two moment-based, distribution-free post-processing methods (Known and Random), and ensemble forecasts sampled from NGR distributions using the moment-based and likelihood (NLL) parameter estimates. The forecast lead time is $t = 3$.

of uncertainty induced by the measurement error model, in particular the use of the jackknife estimates (equation (4.22)). It is not clear from the rank histograms that the NGR-NLL forecasts are the most skilful, particularly for the forecast lead time 5. Interestingly, this is in contrast to the scores presented in table 4.2, which suggest that the use of the NGR method over our distribution-free method is most beneficial at lead time 5. This highlights the importance of using multiple measures for the assessment of forecast skill. We refer the reader to the related discussion at the end of the previous subsection, concerning the (possibly competing) factors of parameter uncertainty, and the induced dependence between ensemble members. Such artefacts aside, however, we suggest that a forecast user may well prefer the NGR-NLL ensemble forecasts to those given by our distribution-free post-processing method. It is worth keeping in mind that unlike the simulation experiment presented in the previous subsection, the statistical model given by the NGR post-processing method is now an approximation to the data-generating process. This should be viewed as another positive reason to select the NGR post-processing method in this case. However, we suggest that the improvements on our baseline distribution-free post-processing method, which concerns only the recalibration of the first and second moments of the ensemble forecasts, might not be as large as one would expect.

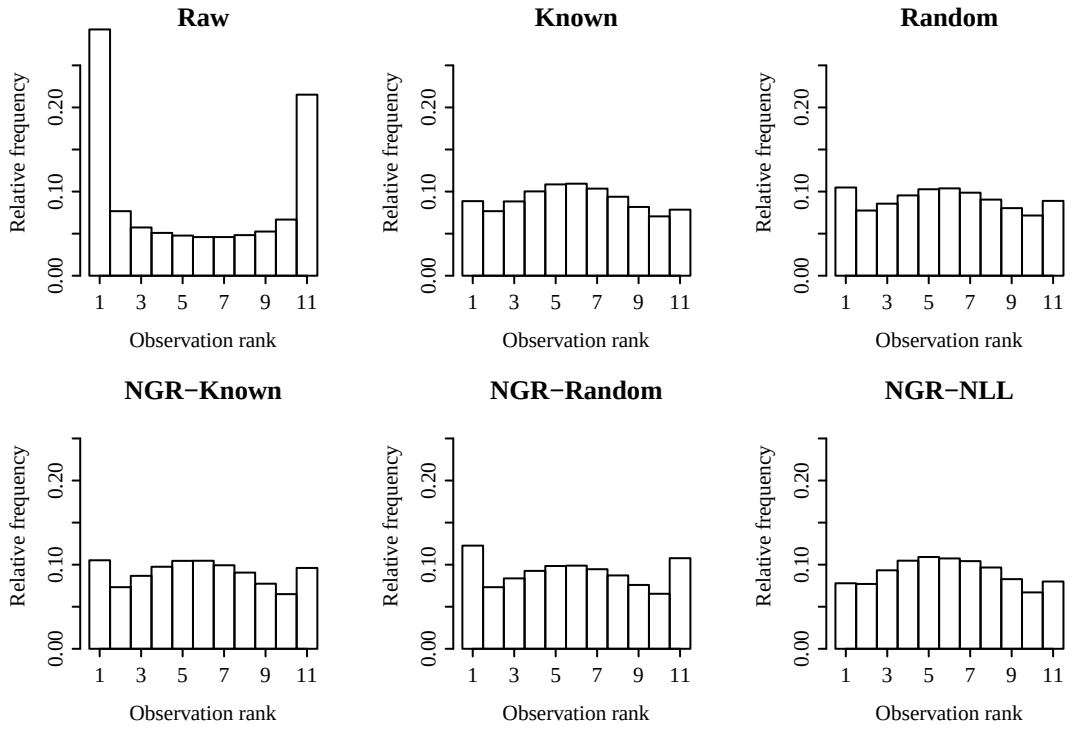

Figure 4.5 As for figure 4.4, for forecast lead time $t = 5$.

4.3.3 Distribution-free post-processing for 2-metre temperature forecasts

Finally, we compare the skill of our distribution-free post-processing method using the method of moments parameter estimates, and ensemble forecasts sampled from NGR probability distributions, in the post-processing of ensemble forecasts for 2-metre temperature. We use the 17×18 grid of 10-member ensemble forecasts and observations that approximately covers the United kingdom, described in section 2.6. We investigate the post-processing of ensemble forecasts at 24 and 72 hour lead times. Parameter estimates were obtained using rolling training samples of the previous 45 forecasts and observations. The qualitative features of the results were found to be similar for other training sample sizes (not shown). The effect of training sample size on the quantitative values of verification scores is illustrated in chapter 5.

In table 4.3 we present univariate scores averaged over the 17×18 grid for the raw ensemble forecasts, ensemble forecasts post-processed using the 'known' and 'random', moment-based estimates using equation (4.52), and the ensemble forecasts sampled as equidistant quantiles of Gaussian NGR forecasts using both the moment-based and NLL parameter estimates. Once again we stress that, for the NGR method, sampling ensemble forecasts as IID random draws from the Gaussian distributions results in significantly worse CRPS and FCRPS scores, although differences in the coverage scores and the associated rank histograms are negligible.

			Distribution-free		NGR		
		Raw	known	Random	known	Random	NLL
24 hour lead time	MSE	0.630	0.561	0.561			0.556
	CRPS	0.443	0.395	0.395	0.384	0.384	0.374
	FCRPS	0.425	0.361	0.361	0.354	0.355	0.343
	Coverage	0.450	0.781	0.774	0.742	0.735	0.787
72 hour lead time	MSE	1.154	1.095	1.097			1.086
	CRPS	0.574	0.557	0.558	0.547	0.549	0.536
	FCRPS	0.535	0.509	0.511	0.505	0.508	0.492
	Coverage	0.654	0.788	0.775	0.746	0.730	0.785

Table 4.3 Univariate scores for the raw ensemble forecasts, and ensemble forecasts recalibrated with the distribution-free and likelihood-based NGR post-processing methods, at forecast lead times of 24 and 72 hours. The scores are averaged over the 17×18 grid that approximately covers the UK.

The qualitative features of the scores are similar to those presented in both the simulation experiment and the study in the Lorenz 1996 system. The likelihood-based parameter estimates for a and b result in improved deterministic forecasts, as measured by the MSE, compared to those of the distribution-free, moment-based methods. Similarly, the CRPS and FCRPS values indicate that the ensemble forecasts sampled as equidistant quantiles from the likelihood-based NGR distributions yield the most skilful ensemble forecasts, under both interpretations of ensemble forecasts (the EDF of the verifying observations, and IID samples from underlying ensemble distributions). For the distribution-free post-processing method (equation (4.52)), the 'known' parameter estimates appear to yield very slight improvements in forecast skill compared to the 'Random' estimates derived with the measurement error model. In contrast to the CRPS and FCRPS scores, the coverage of the ensemble forecasts sampled from the likelihood-based NGR forecasts, and the recalibrated ensemble forecasts obtained from the 'known', distribution-free method is of similar accuracy — recall from section 4.3.1 that the nominal coverage of a 10-member ensemble is 0.818.

We now turn to a comparison of the post-processing methods discussed in this chapter in issuing ensemble forecasts over spatial fields. For this purpose we use a 2×2 subset of the grid described above, that encapsulates the city of London. While assessing forecast skill over a larger spatial field may seem desirable, we are hampered by difficulties in the verification of high-dimensional fields. Indeed, Gneiting et al. [2008] state that the energy score and multivariate rank histogram used here are not suitable for high-dimensional fields. Verification of multivariate forecasts on other subsets of the 17×18 grid (not shown) exhibited similar qualitative features.

In figure 4.6 we display the multivariate rank histograms (see section 2.5.2.3) for the ensemble forecasts recalibrated with the distribution-free post-processing method

using both the 'known' and 'random' parameter estimates, and sampled as equidistant quantiles from Gaussian NGR distributions with NLL parameter estimates. We also show the multivariate rank histogram for the NGR ensemble forecasts that have been reordered to preserve the rank dependence structure of the raw forecasts using the method of ensemble copula coupling (ECC, Schefzik et al. [2013]), described in section 2.4.6. We denote these forecasts by 'NGR-ECC-Q'. As noted in section 4.2.6.3, the distribution-free post-processing method preserves the rank dependence structure of the raw ensemble forecasts, as our transformations (equation (4.52)) of the 'raw' ensemble forecasts are affine. We also show the associated energy score for each multivariate rank histogram.

While interpretations of the multivariate rank histograms are less obvious than in the univariate case (see the discussion given in section 2.5.2.3), it is clear that all post-processing methods considered here result in significant improvements to the calibration of the 'raw' multivariate ensemble forecasts over the 2×2 grid described above. It can be argued that the ensemble forecasts sampled from Gaussian NGR forecast distributions, with the addition of the ECC method, yield the most uniform histograms, although the disparities between these and the histograms for the distribution-free methods are very small relative to the improvement on the raw forecasts. It is striking that the ECC method yields almost no improvement in the energy score for the ensemble forecasts sampled from the NGR method, despite clear improvements in the multivariate rank histogram. However, viewed in the context of the relatively small improvement in the energy scores for the post-processed ensemble forecasts, compared to the transformational improvement in the rank histograms, the qualitative features of these scores are less surprising. Once again, this example highlights the importance of using multiple measures of forecast skill.

Ensemble forecasts that are sampled as IID draws from NGR forecast distributions yield almost identical multivariate rank histograms (not shown) to the approach of sampling equidistant quantiles, both with and without reordering the ensemble members with the ECC methodology. However, as for the earlier discussions in relation to the CRPS and FCRPS, the energy scores for ensemble forecasts whose members are IID are significantly worse than both the equidistant quantiles sampling scheme, and the ensemble forecasts post-processed using our distribution-free method. Again, therefore, if a forecaster was required to issue ensemble forecasts whose members are truly IID, we may be less confident in concluding that sampling from NGR forecast distributions is advantageous compared with our baseline, distribution-free method.

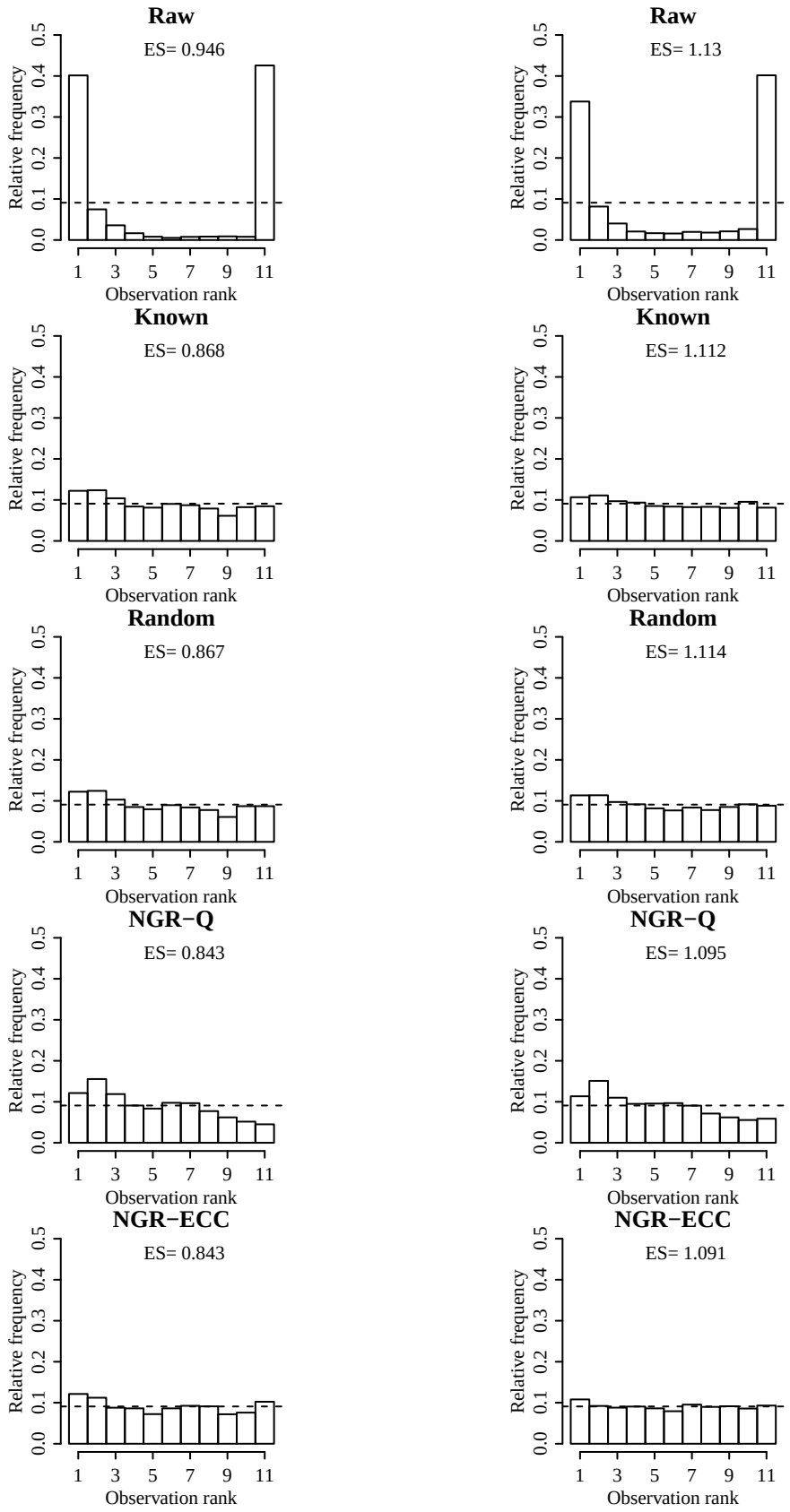

Figure 4.6 Multivariate rank histograms and energy scores (ES) for the raw ensemble forecasts, the 'known' and 'random' forecasts post-processed with the distribution-free post-processing method, and forecasts sampled as equidistant quantiles from NGR forecast distributions with NLL parameter estimates, with and without ECC. Forecast lead times are 24 hours (left) and 72 hours (right).

4.4 Discussion and conclusions

In this chapter we have introduced a novel post-processing method that circumvents the need for the specification of a parametric family of probability distributions to model the verifying observations conditionally on the corresponding ensemble forecasts. Our distribution-free post-processing method is derived from specifying the expectation and variance of the observations as linear functions of the ensemble mean and variance, analogous to the statistical model specified by the popular NGR post-processing method. Our post-processing method leads directly to recalibrated ensemble forecasts, rather than probability forecasts. For situations in which the user requires ensemble forecasts, therefore, our distribution-free method also circumvents the need to specify a means of sampling ensemble forecasts from probability distributions, which is common practice in the literature.

Parameter estimation for our distribution-free post-processing method is achieved with the method of moments, which does not require distributional assumptions. While the method of moments has a long history in the literature, we are not aware of any examples of its use in the estimation of parameters in the modelling of nonhomogeneous forecast errors. Of course, the method of moments could equally be used to estimate the parameters for the more simplistic model output statistics method (MOS, see section 2.4), which would be valid for data that did not exhibit spread-skill relationships between the ensemble forecasts and observations.

As well as treating the ensemble mean and variance as known, known covariates, as is typically the case in the ensemble post-processing literature, we have made the novel introduction to the field of a measurement error model that accommodates the possibility of measurement error in the covariates. Such an approach is motivated by the popular interpretation of ensemble forecasts whose members are realisations of an underlying ensemble distribution.

We compared the skill of ensemble forecasts post-processed with our novel distribution-free method with those sampled from probability distributions specified by the NGR post-processing method. Perhaps surprisingly, we found that the improvement attributable to the specification of the Gaussian probability distributions in the NGR method results in only small improvement to forecast skill and, in some instances, it can be argued that our distribution-free results in forecasts of equal skill. Indeed, if a forecaster was instructed to issue ensemble forecasts whose members truly are IID draws from the underlying NGR distributions, we may even prefer the ensemble forecasts that result from our distribution-free post-processing method.

We suggest that our distribution-free post-processing method should be used as a baseline upon which forecasters should seek to improve. In other words, specifying a

family of probability distributions from which ensemble forecasts are sampled should result in more skilful ensemble forecasts than are obtained from our distribution-free method. Unlike the simplistic, frequency-based approaches that are often used as baseline forecasts, our distribution-free method seeks to issue post-processed ensembles whose mean and variance are well-calibrated with the corresponding moments of the verifying observations. As mentioned below, the parameter estimates for our distribution-free method could be improved. We feel the method is worthy of further research, particularly given its suitability for producing post-processed ensemble forecasts over spatial fields by preserving the rank dependence structure of the ensemble forecasts. Furthermore, we suggest that users may wish to use our parameter estimates obtained with the method of moments as starting values for the optimisation of objective functions, such as the negative log-likelihood, for situations in which the use of probability forecast distributions is required and/or justified.

We found that the parameter estimates obtained with the so-called 'known' parameter estimates, which treat the ensemble mean and variance as known constants, generally yielded post-processed ensemble forecasts that are slightly more skilful than those of the measurement error model, which treats the ensemble mean and variance as random variables. As explained in section 4.2.6, this is in keeping with theoretical expectations, in that the out-of-sample covariates (the ensemble mean and variance) are subject to the same errors as those of the training sample that are used for parameter estimation. However, we believe that the alternative estimation procedure using the measurement error model should not be disregarded. Firstly, if the user is interested in the underlying data-generating process, rather than issuing out-of-sample forecasts, the measurement error parameters are theoretically preferable. Furthermore, theoretical results show that the measurement error model is likely to be preferable in the event of changes to the forecasting system, such as the addition or removal of ensemble members.

This chapter has also highlighted issues for parameter estimation. It is regrettable that the estimation of the parameters c and d, that determine the rescaling of the ensemble forecasts, is problematic when using the method of moments. Perhaps further work to reduce the bias of these estimates, for example by using the Delta method to obtain more accurate estimates of the terms ab and b^2, would assist in reducing the estimation bias for c and d. However, the method of moments is seemingly limited for problems that require the estimation of more parameters, such as the example provided in chapter 6 — the requirement to specify a system of moment equations that provide a unique estimate for each of the model parameters is likely to be problematic in models beyond linear functions of the ensemble mean and variance. In this case, it would seem preferable to optimise an objective function, although it may be possible to easily calculate a subset of parameter estimates using

the method of moments that can be used as starting values.

Our distribution-free post-processing method is appropriate for the post-processing of ensemble forecasts for weather variables that are unbounded. Analogous distribution-free methods for bounded variables, such as precipitation and windspeed, are left for future research. The ensemble rescaling used in this chapter may result in ensemble members that do not satisfy the natural bounds of the weather variables under consideration. For example, recalibrated ensemble forecasts of precipitation may include negative values.

We suggest that measurement error models may have other uses in the context of probabilistic weather forecasting. For example, such models could be used to quantify the uncertainty in the verifying observations (or analyses) which, as in this work, are typically treated as known values. This could be accomplished by collecting analyses from multiple meteorological centres. In doing so, forecasters could quantify the extent to which their uncertainty is due to the standard forecasting problems (NWP model errors, misspecification of initial conditions, etc.), and to uncertainty in the observations that arises from the data assimilation process. Forecasters could then estimate the extent to which forecast uncertainty would be reduced if the analyses were more precise estimates of the true atmospheric state.

5 Parameter uncertainty in ensemble post-processing

5.1 Introduction and motivation

As detailed in the overview provided in section 2.4, ensemble post-processing methods often take the form of statistical models that enable the user to make probability-based statements of the unknown future atmospheric state. The models are based on ensemble forecasts for the verifying observations, and often use their properties, such as the ensemble mean and ensemble variance, as predictor variables. The statistical models usually require the specification of a parametric family of probability distributions for the observations, which are viewed as random variables. In the majority of cases the parameter estimates are chosen as those that optimise an objective function such as the negative log-likelihood (NLL) or continuous ranked probability score (CRPS), which is calculated over a training sample of historical ensemble forecasts and observations. While the parameter estimates are typically treated as point values, they are a function of the random observations, and are thus also subject to uncertainty. In other words, for a given set of ensemble forecasts, $\boldsymbol{x}_i$, and verifying observations, $y_i, i = 1, 2, \ldots, N$, where N denotes the training sample size, different realisations of the (random) observations would yield different realisations of the parameter estimates. In turn, different realisations of the parameter estimates yield different probability forecasts for the out-of-sample observations. The distribution of the parameter estimates is known as the 'sampling distribution', and adds a further source of uncertainty to the probability forecasts. The fact that parameter uncertainty affects the properties of statistical forecasts is well-known, but has hitherto been largely neglected in the ensemble post-processing literature. In this chapter we address the issue of parameter uncertainty in ensemble post-processing.

Typically in the ensemble post-processing literature, the parameter estimates that are obtained by optimising an objective function are used to represent the 'true' model parameters. Out-of-sample probability forecast distributions for the unknown, verifying observations are obtained by direct substitution of the point parameter estimates in to the chosen parametric family of probability distributions,

along with the corresponding out-of-sample ensemble forecast. For example, the NGR forecast distribution for a future observation, y_t is

$$y_t \sim N(\hat{a}_{NGR} + \hat{b}_{NGR}\bar{x}_t, \hat{c}_{NGR} + \hat{d}_{NGR}s_t^2),$$

where $\bar{x}_t$ and s_t^2 are the mean and variance of the M-member ensemble forecast $\boldsymbol{x}_t$, and $\hat{a}_{NGR}, \hat{b}_{NGR}, \hat{c}_{NGR}$ and $\hat{d}_{NGR}$ are the point estimates obtained for the 'true' model parameters, calculated over a training sample of ensemble forecasts and observations. However, the sampling distribution of the parameter estimates adds an additional source of uncertainty to the probability forecast distribution for y_t. Even if the family of probability distributions is correctly specified for the verifying observations, therefore, the common practice of neglecting uncertainty in the model parameters will yield probability forecast distributions that, on average, underestimate the uncertainty in the verifying observations.

More formally, the importance of accounting for parameter uncertainty is motivated by the following discussion. As noted above, the probability forecast distributions that are typically issued for the future verifying observations take the form $f(y_t \mid \boldsymbol{x}_t, \hat{\psi}(D))$, where $f(\cdot)$ denotes the PDF of the chosen family of distributions, and is conditional on both the ensemble forecast $\boldsymbol{x}_t$ and the parameter estimates $\hat{\psi}(D)$. The notation $\hat{\psi}(D)$ has been chosen to indicate the dependence of the parameter estimates $\hat{\psi}$ on the training sample,

$$D = \begin{pmatrix} \boldsymbol{x}_1 & y_1 \\ \boldsymbol{x}_2 & y_2 \\ \vdots & \vdots \\ \boldsymbol{x}_N & y_N \end{pmatrix}.$$

Since the observations y_i in the training sample are random variables, it follows that D and $\hat{\psi}(D)$ are also random. On the other hand, ensemble post-processing methods that specify a parametric family of probability distributions assume that the 'true' data generating process for the future, verifying observations is a member of that family with 'true' model parameters ψ. We denote this 'true' PDF by $f(y_t \mid \boldsymbol{x}_t, \psi)$. The goal of ensemble post-processing is to approximate this PDF. Unlike the estimator $\hat{\psi}(D)$, however, the 'true' parameters ψ are fixed constants. With the exception of the simple (and usually inadequate) model output statistics (MOS) post-processing method (see section 2.4.4.1), analytic results for the sampling distribution of the parameter estimates, and how to use such results to account for parameter uncertainty in out-of-sample probability forecasts, are generally unknown.

In order to investigate the effect of uncertainty in the parameter estimates $\hat{\psi}(D)$ for more complicated statistical models than MOS, a method is required to acco-

modate the sampling distribution of the estimator in the post-processing statistical models. In this chapter we propose to account for the parameter uncertainty in our approximation of the true data generating process by integrating over the sampling distribution of the parameter estimates. The estimates are viewed as so-called 'nuisance parameters' which are integrated out to yield an approximation to the data generating process that accounts for the parameter uncertainty. In other words, by integrating over the sampling distribution of the parameter estimates we obtain a marginal probability distribution for the observations, and remove the additional source of uncertainty that is due to the random nature of the parameter estimates. Specifically, this integral is given by

$$\hat{f}(y_t \mid \boldsymbol{x}_t, \psi) = \int_{\Delta_D} f(y_t \mid \boldsymbol{x}_t, \hat{\psi}(D))\, dF(\hat{\psi}(D)), \tag{5.1}$$

where Δ_D denotes the support of the cumulative distribution function (CDF) of the parameter estimates, $F(\hat{\psi}(D))$, and the notation $\hat{f}(y_t \mid \boldsymbol{x}_t, \psi)$ is chosen to indicate our approximation to the 'true' PDF $f(y_t \mid \boldsymbol{x}_t, \psi)$. Equivalently, the integral can be viewed as the expectation of the probability forecast distribution $f(y_t \mid \boldsymbol{x}_t, \hat{\psi}(D))$ calculated with respect to the parameter estimates $\hat{\psi}(D)$, that is

$$E_{\hat{\psi}(D)}\{f(y_t \mid \boldsymbol{x}_t, \hat{\psi}(D)\}.$$

The expectation is a weighted average of the probability forecast distributions specified by the ensemble post-processing methods for all possible realisations of the parameter estimates, where the weights depend on the sampling distribution.

In practice, the CDF of the sampling distribution of the parameter estimates in ensemble post-processing models, such as nonhomogeneous Gaussian regression (NGR) and Bayesian model averaging (BMA) is generally unknown. The integral given in equation (5.1) must therefore be approximated. To do so requires approximating both the sampling distribution of the parameter estimates, and then using this approximation to approximate the integral. In this chapter we propose a means of approximating and accounting for the sampling distribution of parameter estimates that is applicable to a wide range of ensemble post-processing methods, and is easy to implement. We first use a bootstrap resampling approach to approximate the sampling distribution of the parameter estimates. Bootstrap resampling is a flexible approach that is applicable to the regression-type models that are typically encountered in ensemble post-processing applications, such as those described in chapter 2. If the statistical model is correctly specified, bootstrap resampling enables the user to obtain a representative sample of parameter estimates from the sampling distribution. See Davison and Hinkley [1997, chapter 6] for details on bootstrapping in regression models. Secondly, the integral given in equation (5.1) is readily approxi-

mated from the bootstrap sample, by employing a Monte Carlo approximation.

Our proposed approach for accounting for parameter uncertainty in ensemble post-processing has both similarities and differences to the natural alternative of forming a Bayesian model, for which parameter uncertainty is accommodated as an inherent feature of the statistical model. The key difference is in the approach used for estimating the distribution of the parameter estimates. In the frequentist philosophy adopted in this chapter, we use the sampling distribution of the parameter estimator, the distribution of the estimator that represents the typical variation of the parameter estimates around the unknown, 'true' parameter values. In the Bayesian philosophy, on the other hand, the model parameters are considered to be random variables, in order to represent the user's uncertainty about the 'true' parameter values. Bayesians would specify a prior distribution for the model parameters which, for example, would be based on their knowledge (or intuition) of the relationships between the ensemble forecasts and verifying observations. The prior distribution would then be updated by including the training sample, to result in a posterior distribution for the parameters. The similarity between the two approaches is that the probability forecast distribution for the verifying observation is obtained by integrating over the posterior distribution (in the Bayesian approach), or the sampling distribution (in our frequentist-based approach). As we discuss in section 5.5, there are advantages and disadvantages to both approaches. However, our proposed method is easy to implement, and will serve as a useful guide for determining the importance of accounting for parameter uncertainty. Friederichs and Thorarinsdottir [2012] has applied a Bayesian model for probability forecasts of peak wind speeds, and Siegert et al. [2015b] has proposed an approach that is appropriate for forecasts on long time scales in the climate literature.

The remainder of this chapter is organised as follows. In section 5.2 we review an analytic result for accounting for parameter uncertainty in the simple MOS model, and give details of our proposed bootstrap resampling procedure which is appropriate for more complicated models. In section 5.3 we describe the verification measures that are used to compare forecasts that neglect and account for parameter uncertainty. In section 5.4 we present results for three case studies, and in section 5.5 we conclude the chapter with a discussion and ideas for future research.

5.2 Parameter uncertainty: Analytic results and bootstrap approximations

5.2.1 Analytic results for model output statistics

We now present an analytic result that enables forecasts to take account of parameter uncertainty when using the model output statistics (MOS) model, described in section 2.4.4.1. Recall that the probability forecast distribution for a future observation, y_t, conditional on the corresponding ensemble forecast $\boldsymbol{x}_t$ is given by

$$y_t \sim N(a + b\bar{x}_t, c^2), \tag{5.2}$$

where a, b and c are the 'true' parameter values that we must estimate. As discussed in the previous section, the standard approach in the literature is to issue probability forecasts by direct substitution of $\hat{\psi}(D) = (\hat{a}, \hat{b}, \hat{c})$ for $\psi = (a, b, c)$, such that probability forecast distributions are issued as

$$y_t \sim N(\hat{a} + \hat{b}\bar{x}_t, \hat{c}^2). \tag{5.3}$$

Probability forecast distributions given by equation (5.3) have been used to post-process ensemble forecasts on seasonal scales [Kharin and Zwiers, 2003; Tippett et al., 2005], and for short-range weather forecasts [Glahn et al., 2009]. However, equation (5.3) takes no account of the uncertainty in $\hat{a}, \hat{b}$ and $\hat{c}$ as estimates of a, b and c. Glahn et al. [2009] noted the issue of parameter uncertainty, and even gave the result that we derive in this subsection, but asserted that it is not important for training samples larger than $N = 30$, and did not refer to the issue thereafter.

It is easily shown that the parameter estimates for a, b and c that minimise the negative log-likelihood (NLL) are

$$\hat{a} = \bar{y} - \hat{b}\overline{\bar{x}} \tag{5.4}$$

$$\hat{b} = \frac{S_{\bar{x},y}}{S_{\bar{x},\bar{x}}} \tag{5.5}$$

$$\hat{c}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{a} - \hat{b}\bar{x}_i\right)^2, \tag{5.6}$$

where $\overline{\bar{x}} = N^{-1}\sum_{i=1}^{N}\bar{x}_i$, $S_{\bar{x},y} = N^{-1}\sum_{i=1}^{N}(\bar{x}_i - \overline{\bar{x}})(y_i - \bar{y})$ and $S_{\bar{x},\bar{x}} = N^{-1}\sum_{i=1}^{N}(\bar{x}_i - \overline{\bar{x}})^2$, and N is the size of the training sample.

Draper et al. [1998, chapter 1] shows that the sampling distributions of the parameter

estimates $\hat{a}$ and $\hat{b}$ are

$$\hat{a} \sim N\left\{a, c^2\left(\frac{1}{N} + \frac{\overline{\overline{x}}^2}{S_{\bar{x},\bar{x}}}\right)\right\} \tag{5.7}$$

$$\hat{b} \sim N\left(b, \frac{c^2}{S_{\bar{x},\bar{x}}}\right). \tag{5.8}$$

An unbiased estimator for c^2 can be obtained by rescaling the likelihood estimator by a factor of $N/(N-2)$, in order to obtain the unbiased sample variance estimator for a sample of size N with $N-2$ degrees of freedom,

$$\hat{c}^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - \hat{a} - \hat{b}\bar{x}_i\right)^2. \tag{5.9}$$

Furthermore, a standard result from probability theory [Draper et al., 1998, chapter 1] is

$$\frac{(N-2)\hat{c}^2}{c^2} \sim \chi^2_{N-2}. \tag{5.10}$$

When issuing forecasts $\hat{y}_t = \hat{a} + \hat{b}\bar{x}_t$, therefore, the sampling distributions of the estimators $\hat{a}, \hat{b}$ and $\hat{c}^2$ should be accounted for. Draper et al. [1998, chapter 1] show that

$$\mathrm{var}(\hat{y}_t) = c^2\left(1 + \frac{1}{N} + \frac{\left(\bar{x}_t - \overline{\overline{x}}\right)^2}{S_{\bar{x},\bar{x}}}\right). \tag{5.11}$$

Accounting for uncertainty in the parameter estimates $\hat{a}$ and $\hat{b}$ therefore leads to probability forecast distributions with larger variance than those obtained through the usual practice of direct substitution of the estimates in to the forecast distribution. The variance of the forecast distribution for y_t is inflated by a term $1/N$, and a term that grows quadratically in the distance $|\bar{x}_t - \overline{\overline{x}}|$ of the ensemble mean $\bar{x}_t$ from the mean of the ensemble means $\bar{x}_i, i = 1, 2, \ldots, N$, in the training sample. The forecast variance is therefore larger for more extreme values of $\bar{x}_t$. This feature is intuitively appealing — it implies that forecast uncertainty increases with the distance of the ensemble mean $\bar{x}_t$ from the mean of the ensemble means in the training sample.

Finally, accounting for uncertainty in the parameter estimate $\hat{c}^2$ results in a change to the forecast distribution itself. Draper et al. [1998, chapter 1] show that the forecast distribution for the future observation y_t should instead take the form of a t-distribution with $N-2$ degrees of freedom, that is

$$\frac{y_t - \hat{y}_t}{\hat{c}\sqrt{1 + 1/N + (\bar{x}_t - \overline{\overline{x}})^2/S_{\bar{x},\bar{x}}}} \sim t_{N-2},$$

and so probability forecasts for y_t should be derived from the t_{N-2} distribution with expectation $\hat{y}_t$, and variance $\hat{c}^2(1 + 1/N + (\bar{x}_t - \overline{\overline{x}})^2/S_{\bar{x},\bar{x}})$. The tails of the t distribution are heavier than the corresponding tails of the Gaussian distribution. This fact, combined with the inflated variance derived above, results in larger forecast uncertainty than would be derived through the standard approach of Tippett et al. [2005]; Kharin and Zwiers [2003]; Glahn et al. [2009]. As the training sample size N increases, the t_{N-2} distribution tends to the Gaussian distribution.

5.2.2 Accounting for parameter uncertainty with the predictive bootstrap

5.2.2.1 Approximating the sampling distribution of parameter estimates

The analytic result derived in the previous subsection is based upon the assumption that the distribution of the observation y_t, conditional on the ensemble mean $\bar{x}_t$ is normal with expectation $a + b\bar{x}_t$ and constant variance c^2. Unfortunately, as noted in section 5.1, analogous results are not known for more complicated statistical models, such as the NGR model (see equations (2.19) and (2.20) on page 38) in which the forecast variance is nonhomogeneous, and requires the estimation of two parameters. Furthermore, ensemble post-processing methods whose probability forecast distributions are non-Gaussian, such as the methods developed for the forecasting of precipitation (e.g. Scheuerer [2014], described in section 2.4.5), will typically yield parameter estimates with non-Gaussian sampling distributions. In general, therefore, analagous results to those presented for the MOS model are likely to be either mathematically intractable or difficult to obtain. While it may be possible to derive results using asymptotic theory, the small training samples that are often encountered in real-world scenarios would limit the value of such work. Therefore, a method that is easily applicable to a wide range of post-processing methods, without the need for complicated mathematical analysis, seems highly desirable. Then, if the effect on forecast skill of accounting for parameter uncertainty appears to be significant, it may subsequently be worth attempting to derive analytic results for small training samples.

An appealing and easy-to-implement option is to estimate the sampling distribution of the parameter estimates by bootstrap resampling [Efron and Gong, 1983]. The bootstrap method is a popular means of approximating the sampling distributions of summary statistics of data samples. Using the original sample for which the summary statistic is calculated, the bootstrap method simulates N_B new, 'synthetic' data samples, for which the summary statistic is recalculated. Each of the N_B values of the summary statistic is referred to as a 'bootstrap replicate', and the distribu-

tion of the replicates is used to approximate the desired properties of the sampling distribution of the summary statistic. In particular, the bootstrap is widely used to estimate the sampling distribution of so-called pivotal statistics. The sampling distributions of pivots are by definition independent of the 'true parameter values', ψ, and can be used to approximate confidence intervals or to conduct hypothesis tests. Bootstrapping pivotal statistics is appealing, since there is no error (beyond the effects of sampling variation) in estimating properties of the distribution of the pivot by bootstrap resampling — in other words, the sample of bootstrap replicates of the pivotal statistic is a representative sample of its sampling distribution.

In the context of ensemble post-processing, we use bootstrap resampling to approximate the sampling distribution of the estimated model parameter vector, $\hat{\psi}(D)$, the model parameter estimates obtained by minimising the objective function over the training sample D. As with pivotal statistics, we would like the empirical distribution of the bootstrap replicates to approximate the sampling distribution of $\hat{\psi}(D)$. For example, in accounting for parameter uncertainty in the NGR model, we would like the empirical distribution of parameter estimates obtained from bootstrap resamplling, say $(\hat{a}^*_{NGR,j}, \hat{b}^*_{NGR,j}, \hat{c}^*_{NGR,j}, \hat{d}^*_{NGR,j})$, for $j = 1, 2, \ldots, N_B$ to provide a good approximation to the sampling distribution of the parameter estimates $\hat{\psi}(D) = (\hat{a}_{NGR}, \hat{b}_{NGR}, \hat{c}_{NGR}, \hat{d}_{NGR})$. Bootstrap resampling has been used to perform inference in regression models, and is discussed in detail in Davison and Hinkley [1997, chapter 6]. For example, users are more often interested in constructing confidence intervals (or regions) for parameter estimates, and bootstrap resampling provides an intuitive means of doing so when analytic results are not available or are difficult to obtain.

We now discuss some possible approaches for using bootstrap resampling to approximate the sampling distribution of the parameter estimates $\hat{\psi}(D)$. The simplest approach is known as 'case resampling'. Given the training sample,

$$D = \begin{pmatrix} \boldsymbol{x}_1 & y_1 \\ \boldsymbol{x}_2 & y_2 \\ \vdots & \vdots \\ \boldsymbol{x}_N & y_N \end{pmatrix}$$

we create new, synthetic training samples, D^*, of size N by sampling with replacement from the rows of D. The post-processing model is refitted for the training sample D^*, with the resulting parameter estimates denoted by $\hat{\psi}^*$. This process is repeated N_B times to form a set of parameter estimates $(\hat{\psi}^*_1, \hat{\psi}^*_2, \ldots, \hat{\psi}^*_{N_B})$, which are interpreted as representing a sample from the sampling distribution of the parameter estimates $\hat{\psi}(D)$.

A second approach is the so-called 'residual bootstrap'. In this case the statistical model is fitted, and standardised residuals are calculated for the fitted values over the training sample. For example, the standardised residuals for the NGR model are given by

$$\mathrm{sr}_i = \frac{y_i - \mu_i}{\sigma_i} \text{ for } i = 1, 2, \ldots, N, \tag{5.12}$$

where $\mu_i = \hat{a} + \hat{b}\bar{x}_i$ and $\sigma_i = \sqrt{\hat{c} + \hat{d}s_i^2}$ are the expectation and standard deviation of the ith NGR forecast distribution, respectively. The residual bootstrap then proceeds as follows. Firstly, a sample of size N is drawn (with replacement) from the N standardised residuals sr_i, denoted $\mathrm{sr}^* = (\mathrm{sr}_1^*, \mathrm{sr}_2^*, \ldots, \mathrm{sr}_N^*)$. This sample is then used to create synthetic observations, $y_i^*, i = 1, 2, \ldots, N$ by

$$y_i^* = \mu_i + \sigma_i \mathrm{sr}_i^*. \tag{5.13}$$

In other words, the sample of standardised residuals is added to the fitted values of the post-processing method, μ_i, and rescaled by σ_i. If the standardised residuals sr_i do indeed have unit variance, it follows that the ith synthetic observation, y_i^*, has variance σ_i^2. The statistical model is then refitted using the N ensemble forecasts, and the sample of synthetic observations $y_i^*, i = 1, 2, \ldots, N$. Again, this process is repeated N_B times to form a sample of N_B parameter estimates. Our implementation of the residual bootstrap is an extension of the approach described in Davison and Hinkley [1997, chapter 6], which assumes that the variance of the verifying observations is homoscedastic, rather than heteroscedastic. Note that, unlike the case resampling approach, the residual bootstrap does not sample from the ensemble forecasts — the ensemble forecasts remain fixed, while the standardised residuals are viewed as independent random variables that are interchangeable.

A third approach is the 'parametric bootstrap' which, unlike the residual bootstrap, assumes complete knowledge of the distribution of the observations. In our use of the residual bootstrap we assume only that the first and second moments of the forecast distributions are correctly specified — we do not make any distributional assumptions in the simulation procedure used to create synthetic training samples. On the other hand, the parametric bootstrap assumes that the statistical post-processing model is correctly specified. In the case of the NGR post-processing method, therefore, the standardised residuals sr_i defined above are assumed to be IID $N(0, 1)$, if the data really do arise from the fitted NGR model. In this case the parametric bootstrap proceeds by simulating samples of size N from the standard normal distribution, which are added to the fitted values μ_i and rescaled by σ_i, in the same manner as given by equation (5.13), such that

$$y_i^* = \mu_i + \sigma_i z_i^*, \text{ for } i = 1, 2, \ldots, N, \tag{5.14}$$

where the $z_i^* \sim N(0,1)$ are IID draws from a standard normal distribution. The variables y_i^* are then used as synthetic observations, and the model is refitted to obtain the bootstrap parameter estimates. Again this procedure is repeated N_B times. As with the residual bootstrap, the parametric bootstrap does not sample from the ensemble forecasts.

The three approaches to bootstrap resampling described above each place increasing confidence on the statistical model specified by the ensemble post-processing method. The parametric bootstrap assumes that the statistical model is correctly specified, and that we are only unaware of the 'true' parameter values, ψ. The approach of resampling the standardised residuals assumes that the parametric functions for the expectation and variance of the statistical model are correct, but does not make distributional assumptions. Case resampling, which samples both the predictor variables (such as the ensemble means and variances) and the observations, makes no modelling assumptions.

It seems that our parametric and residual-based resampling approaches are, in theory, preferable to case resampling. Firstly, by resampling from the predictor variables as well as the observations, the synthetic training samples, D^*, contain less information about the conditional distribution of the observations, given the predictor variables. For example, outlying predictor variables may not be present in the synthetic training samples. The synthetic training samples will also contain (with high probability) replicate predictor variable-observation pairs, and so the training samples D^* are effectively smaller than the original training sample, D. On the other hand, by maintaining all predictor variables in the synthetic training samples, the parameter estimates $\hat{\psi}^*$ that are obtained with the residual and parametric resampling approaches are subject to less sampling variation, and there is no loss of information in the predictor variables. We might reasonably expect, therefore, that the variability of the bootstrap replicates will be largest for case resampling, while residual and parametric resampling will yield less variable estimates. If the user wishes to use the bootstrap to estimate confidence intervals (or confidence regions) for the parameter estimates, therefore, we would expect intervals obtained by case resampling to be wider than those obtained with the residual and parametric approaches.

In practice the statistical post-processing models are an approximation to the 'true' data generating process, and so which of the approaches to bootstrap resampling is preferred is likely to depend on the accuracy of the statistical model as an approximation to the distribution of the observations. We might expect case resampling to be preferred for poorly specified models. For example, suppose the 'true' relationship between the ensemble means $\bar{x}_i$ and verifying observations y_i is nonlinear, rather than linear, as is assumed by the MOS and NGR models. In that case, the stan-

dardised residuals given by equation (5.12), sr_i, will each have non-zero expectation that depends on the forecast occasion, i. The assumption of identically distributed standardised residuals would therefore be violated. Intuitively, therefore, it seems that resampling the residuals to create synthetic samples of observations using equation (5.13) would yield observations, y_i^*, for which the assumed linear relationship is typically a worse approximation than it is for the original sample of observations, y_i.In turn, therefore, the sample of bootstrap parameter estimates will be a poor approximation to the sampling distribution of the parameter estimates in the linear model. On the other hand, sampling both the ensemble forecasts and verifying observations with case resampling merely affects the effective size of the synthetic training samples D^* (as mentioned above) — we do not introduce synthetic observations for which the assumed linear relationship is a worse approximation to the 'true' nonlinear data generating process.

Finally, we consider the use of the 'block bootstrap' [Davison and Hinkley, 1997, chapter 8]. If the standardised residuals given in equation (5.12) are temporally correlated, then the assumption that the observations (conditional on the ensemble forecasts) are independent and identically distributed is violated. In such cases, therefore, independent resampling approaches (such as the three described above) destroys the temporal dependence structure of the standardised residuals. The block bootstrap samples 'blocks' from the original training sample, D, rather than sampling independently, where the term 'block' refers to a series of length L of consecutive data from the training sample. The block length L is chosen so as to retain the predominant features of the temporal dependence structure. For example, in section 5.4.2 we experimented with blocks of length 2, as there was some evidence of autocorrelation of the standardised residuals at lag 1, but not at longer lags. To construct a new training sample, D^*, of size N, N/L blocks of length L are sampled from the training sample D. The statistical model is refitted to D^* to obtain parameter estimates $\hat{\psi}^*$, and the process repeated N_B times. In addition, we consider two possible approaches of constructing 'blocks'. The first is to sample blocks along the lines of case resampling, by sampling blocks of both ensemble forecasts and the corresponding observations. Alternatively, we also consider sampling blocks of standardised residuals, and not sampling from the ensemble forecasts (in keeping with the residual bootstrap).

5.2.2.2 Accounting for the sampling distribution of parameter estimates with the predictive bootstrap

We now explain how our bootstrap approximation to the sampling distribution of the parameter estimates is used to account for parameter uncertainty in the resulting

probability forecast distributions that are issued for the future observations. Firstly, recall the integral given in section 5.1 (see equation (5.1)),

$$\hat{f}(y_t \mid \boldsymbol{x}_t, \psi) = \int_{\Delta_D} f(y_t \mid \boldsymbol{x}_t, \hat{\psi}(D))\, dF(\hat{\psi}(D)),$$

where Δ_D denotes the support of the CDF of the sampling distribution of the parameter estimates $\hat{\psi}(D)$. Having used one of the bootstrap resampling approaches to obtain a representative sample of the sampling distribution, the integral is approximated with a Monte Carlo approach, as we now explain. Each of the N_B bootstrap replicates of the parameter vector, $\hat{\psi}_j^*, j = 1, 2, \ldots, N_B$, results in a probability forecast distribution with PDF $f(y_t \mid \boldsymbol{x}_t, \hat{\psi}_j^*)$. In order to account for the sampling distribution of the parameter estimates $\hat{\psi}$, we average over the N_B distributions $f(y_t \mid \boldsymbol{x}_t, \hat{\psi}_j^*)$ to obtain

$$\hat{f}(y_t \mid \boldsymbol{x}_t, \psi) = \frac{1}{N_B} \sum_{j=1}^{N_B} f(y_t \mid \boldsymbol{x}_t, \hat{\psi}_j^*). \tag{5.15}$$

Equation (5.15) is therefore a Monte Carlo approximation to the aforementioned integral given in equation (5.1). The sampling distribution of the parameter estimates, $\hat{\psi}(D)$, has been integrated out to yield the marginal distribution of the observation y_t, conditional on the corresponding ensemble forecast $\boldsymbol{x}_t$, and the 'true' parameter vector ψ. The resulting probability forecast distribution for a future observation y_t is thus a mixture distribution with N_B 'component distributions'.

This approach for accounting for parameter uncertainty, which is hereafter referred to as the 'predictive bootstrap', was proposed by Harris [1989]. While the author introduced the term 'predictive bootstrap', he did not provide examples of the use of bootstrap resampling to estimate the sampling distribution of the parameter estimates $\hat{\psi}(D)$. Rather, examples were presented for which the sampling distribution of parameter estimates could be calculated analytically. The author showed that under the correct model for the data generating process, the expectation of the log-likelihood for the predictive bootstrap forecast distributions is larger than the expected log-likelihood for the standard plug-in forecast distributions or, equivalently, that the negative log-likelihood is, on average, smaller for the predictive bootstrap forecast distributions. As the negative log-likelihood is closely related to the ignorance score (see section 2.5.3), we might reasonably expect to improve the ignorance score of out-of-sample probability forecasts by using the predictive bootstrap to account for parameter uncertainty.

5.3 Forecast verification

In section 5.4 we compare the skill of the probability forecasts obtained with the predictive bootstrap method, hereafter referred to as the 'bootstrap forecasts', and the standard forecasts that do not account for parameter uncertainty, hereafter referred to as the 'plug-in' forecasts. In sections 5.4.1 and 5.4.2 we consider probability forecasts, and in section 5.4.3 we consider the skill of ensemble forecasts that are sampled from probability forecast distributions. The skill of probability forecasts is assessed with the ignorance score (IGN), the continuous ranked probability score (CRPS), the coverage of 95% prediction intervals, PIT histograms and the Brier score, which is used to assess the skill of probability forecasts for binary events. We use the NGR post-processing method, and so the bootstrap forecast distributions are a mixture of N_B Gaussian component distributions, each with different expectations and variances that depend on the N_B parameter estimates. Specifically, therefore, the aforementioned verification measures are given as follows.

The ignorance score and CRPS

For the plug-in forecasts, the ignorance score and the CRPS are given by

$$\text{Ign}_{\text{Plugin}} = -\frac{1}{T}\sum_{t=1}^{T}\log_2\left\{\frac{1}{\sigma_t}\phi\left(\frac{y_t-\mu_t}{\sigma_t}\right)\right\} \tag{5.16}$$

$$\text{CRPS}_{\text{Plugin}} = \frac{1}{T}\sum_{t=1}^{T}\sigma_t\left[z_t\{2\Phi(z_t)-1\}+2\phi(z_t)-\frac{1}{\sqrt{\pi}}\right], \tag{5.17}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard Gaussian PDF and CDF, $z_t = (y_t - \mu_t)/\sigma_t$ is a (standardised) observation, and μ_t, σ_t depend on the post-processing method. For MOS, we have

$$\begin{aligned}\mu_t &= \hat{a} + \hat{b}\bar{x}_t\\ \sigma_t^2 &= \hat{c}^2,\end{aligned}$$

and for NGR we have

$$\begin{aligned}\mu_t &= \hat{a} + \hat{b}\bar{x}_t\\ \sigma_t^2 &= \hat{c} + \hat{d}s_t^2.\end{aligned}$$

Note that equation (5.17) for the CRPS is simply the out-of-sample equivalent of the objective function used for parameter estimation in chapter 3 (see equation (3.3)). The closed form expression was given in Gneiting et al. [2005].

As noted in the previous section, the predictive bootstrap gives rise to mixture distributions. The ignorance score for bootstrap forecasts using MOS and NGR post-processing methods, which give rise to Gaussian probability forecast distributions, is

$$\text{Ign}_{\text{Bootstrap}} = -\frac{1}{T}\sum_{t=1}^{T}\log_2\left\{\frac{1}{N_B}\sum_{j=1}^{N_B}\sigma_{j,t}^{*^{-1}}\phi\left(\frac{y_t-\mu_{j,t}^*}{\sigma_{j,t}^*}\right)\right\}, \tag{5.18}$$

where for MOS we have

$$\begin{aligned}\mu_{j,t}^* &= \hat{a}_j^* + \hat{b}_j^*\bar{x}_t\\ \sigma_{j,t}^* &= \hat{c}_j^*,\end{aligned}$$

and for NGR we have

$$\begin{aligned}\mu_{j,t}^* &= \hat{a}_j^* + \hat{b}_j^*\bar{x}_t\\ \sigma_{j,t}^* &= \sqrt{\hat{c}_j^* + \hat{d}_j^* s_t^2}.\end{aligned}$$

As noted in chapter 3 for the estimation of model parameters for the BMA and BMD post-processing methods, the CRPS for a mixture of Gaussian distributions was given by Grimit et al. [2006] as

$$\text{CRPS}_{\text{Bootstrap}} = \frac{1}{T}\sum_{t=1}^{T}\text{crps}\left(y_t, F(y_t \mid \boldsymbol{x}_t, \psi)\right), \tag{5.19}$$

where F denotes the CDF of the mixture distribution given by the predictive boostrap, and the CRPS for an individual forecast at time t is

$$\text{crps}\left(y_t, F(y_t \mid \boldsymbol{x}_t, \psi)\right) = \frac{1}{N_B}\sum_{j=1}^{N_B}\left\{A(y_t-\mu_{j,t}^*, \sigma_{j,t}^{*^2}) - \frac{1}{2}\sum_{k=1}^{N_B}A(\mu_{j,t}^*-\mu_{k,t}^*, \sigma_{j,t}^{*^2}+\sigma_{k,t}^{*^2})\right\} \tag{5.20}$$

where

$$A(\mu,\sigma^2) = 2\sigma\phi(\mu/\sigma) + \mu\left\{2\Phi\left(\mu/\sigma\right) - 1\right\}, \tag{5.21}$$

In section 5.4.1 we compare the results of the plug-in and predictive bootstrap forecasts with the analytic results presented in section 5.2.1 for the MOS post-processing method, hereafter referred to as the 'analytic forecasts'. The ignorance score for the analytic forecasts is given by

$$\text{Ign}_{\text{Analytic}} = -\frac{1}{T}\sum_{t=1}^{T}\log_2\tau_{N-2}\left\{y_t\middle|\mu_t, \hat{c}^2(1+1/N+(\bar{x}_t-\overline{\bar{x}})^2/S_{\bar{x},\bar{x}})\right\}, \tag{5.22}$$

where $\tau_{N-2}(\cdot)$ denotes the density of the t-distribution with $N-2$ degrees of freedom,

with expectation μ_t and the inflated variance derived in section 5.2.1. We are not aware of a published result for the CRPS of a t-distributed random variable, and so we use numerical integration to approximate the integral form of the CRPS (see equation (2.44) on page 58) for the CRPS of the analytic forecasts in section 5.4.1. We do not provide the CRPS for the BMA predictive bootstrap distributions, as we do not consider probability forecasts in our example (section 5.4.3).

The probability integral transform

The values of the probability integral transform (PIT) for the MOS and NGR plug-in forecast distributions, which are plotted in the PIT histograms in section 5.4.2 (see section 2.5.2.3 for details) are given by

$$F_{\text{Plugin},t} = \Phi\left(\frac{y_t - \mu_t}{\sigma_t}\right), \tag{5.23}$$

where $\Phi(\cdot)$ denotes the standard Gaussian CDF, and μ_t, σ_t are as given earlier in this section for the ignorance score and CRPS of the plug-in forecasts. Similarly, the PIT values for forecast occasion t of the MOS and NGR bootstrap forecast distributions are given by

$$F_{\text{Bootstrap},t} = \frac{1}{N_B}\sum_{j=1}^{N_B}\Phi\left(\frac{y_t - \mu^*_{j,t}}{\sigma^*_{j,t}}\right), \tag{5.24}$$

where $\mu^*_{j,t}$ and $\sigma^*_{j,t}$ are as given above for the ignorance score and CRPS of the bootstrap forecast distributions.

Prediction intervals

We calculate 95% prediction intervals using the 2.5% and 97.5% quantiles ($q_{.025}$ and $q_{.975}$) of the probability forecast distributions. For the Gaussian and t-distributed probability forecasts, these quantiles are calculated using the 'qnorm' and 'qt' functions implemented in the R language [R Core Team, 2015]. We are not aware of a closed form for the inverse of the CDF of a mixture of Gaussian distributions, however, and so we estimate quantiles of the bootstrap forecast distributions numerically using the 'uniroot' function of the R language. Specifically, to calculate the p-quantile for a general forecast occasion we find the value of u that solves the equation

$$F(u) - p = 0,$$

where $F(u)$ is the CDF of the bootstrap forecast distribution evaluated at u.

The Brier score

In section 5.4.2 we compare the skill of the plug-in and bootstrap forecast distributions in issuing probability forecasts of binary events, for the NGR post-processing method. For a fixed threshold q, probability forecasts of the binary event $\{y_t \leq q\}$ for the plug-in forecasts are given by

$$\Pr(y_t \leq q) = \Phi\left(\frac{q - \mu_t}{\sigma_t}\right), \tag{5.25}$$

and for the bootstrap forecasts by

$$\Pr(y_t \leq q) = \frac{1}{N_B}\sum_{j=1}^{N_B} \Phi\left(\frac{q - \mu^*_{j,t}}{\sigma^*_{j,t}}\right), \tag{5.26}$$

where $\mu_t, \sigma_t, \mu^*_{j,t}$ and $\sigma^*_{j,t}$ are as given earlier in this section. The Brier score and its decomposition (see sections 2.5.3.2 and 2.5.5) are used to assess the skill of these forecasts.

Verification for ensemble forecasts

Finally, in section 5.4.3 we sample ensemble forecasts from plug-in and bootstrap probability forecast distributions for the BMA post-processing method. For BMA, both the plug-in and bootstrap forecast distributions are mixtures of Gaussian component distributions. We sample M ensemble members as the equidistant quantiles $q_{m/(M+1)}$ for $m = 1, 2, \ldots, M$ with the numerical method described above. The skill of these ensemble forecasts is assessed using the empirical estimate of the CRPS (see equation (2.48) on page 60) for univariate quantities, and the energy score (see equation (2.47) on page 59) for forecasts over spatial fields.

5.4 Results

5.4.1 A simulation study

In this subsection we compare the skill of 'plug-in' and 'bootstrap' probability forecast distributions in an idealised simulation experiment, so that the observations are distributed according to the statistical model specified by the ensemble post-processing method. We begin by showing results for the NGR model, and then consider the MOS model for which we also have the analytic results given in section 5.2.1. We present results using case resampling. The residual and parametric

bootstrap approaches yield very similar results and, as the training data in this experiment are independent by construction, there is nothing to be gained by using the block bootstrap.

For the NGR post-processing method, the simulation experiment proceeds as follows.

1. Simulate 100 random variables ξ_i and θ_i^2, that are representative of the ensemble mean (ξ) and ensemble variance (θ^2).

2. For given 'true' parameters a, b, c and d, simulate 100 training observations y_i with distribution $N(a + b\xi_i, c + d\theta_i^2)$ for $i = 1, 2, \ldots, 100$.

3. Simulate out-of-sample random variables ξ_0 and θ_0^2 and a corresponding observation $y_0 \sim N(a + b\xi_0, c + d\theta_0^2)$.

4. Fix a training sample size, N.

5. Compute the 'plug-in' parameter estimates $\hat{a}_{NGR}, \hat{b}_{NGR}, \hat{c}_{NGR}$ and $\hat{d}_{NGR}$ by optimisation of the negative log likelihood.

6. Also compute N_B sets of parameter estimates $(\hat{a}_j^*, \hat{b}_j^*, \hat{c}_j^*, \hat{d}_j^*), j = 1, 2, \ldots, N_B$ using bootstrap resampling.

7. Evaluate the probability integral transform, the ignorance score and the CRPS for the plug-in and bootstrap forecast distributions. Determine whether y_0 lies in the prediction interval of the plug-in and bootstrap forecast distributions.

8. Repeat steps 4–7 for N in the set $\{10, 20, \ldots, 100\}$.

9. Repeat steps 1–8 N_{sim} times to obtain N_{Sim} measures of forecast skill for the various sizes of training sample.

In this study we used parameter values $a = 1/2, b = 5/4, c = 1/2$ and $d = 3/2$. The predictor variables ξ_i were simulated as Gaussian-distributed random variables with distribution $N(0, 6^2)$, and the variables θ_i^2 were taken as the absolute value of simulations with distribution $N(0, (1/2)^2)$. The chosen parameter values are representative of biases in ensemble location and dispersion that are commonly observed in practical scenarios. The average variance of the observations is approximately 2% of the variance of the (synthetic) ensemble means, ξ, which is representative of typical ensemble forecasts of temperature. However, we found that the qualitative features of the results shown below are similar regardless of the marginal distributions of ξ and θ^2, and the values of the true parameters a, b, c and d. We used $N_{sim} = 10\,000$ simulations for each training sample size and $N_B = 100$ bootstrap replicates. We also investigated simulations with 50 and 200 bootstrap replicates. In our experience there is little to be gained in using more than 50 replicates — for

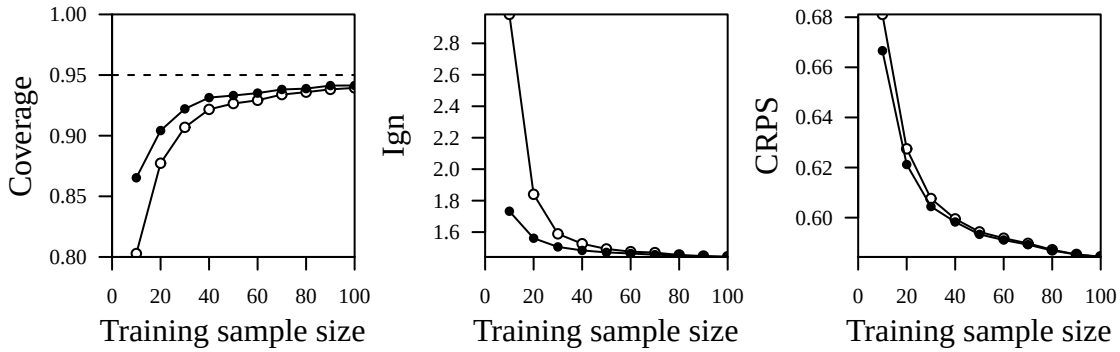

Figure 5.1 Coverage of the 95% prediction intervals, the ignorance score (Ign) and the CRPS for the plug-in forecasts (open circles) and the predictive bootstrap forecasts (filled circles) as a function of training sample size, N, for simulated observations that follow the NGR model.

example, improvements in the ignorance score occur only at the third decimal place, while significantly increasing the computational cost.

Figure 5.1 shows the coverage of the 95% prediction intervals, the ignorance scores and the continuous ranked probability scores for the standard plug-in forecasts and the bootstrap forecasts (equation (5.15)) as a function of training sample size, N. The predictive bootstrap forecasts improve on the plug-in forecasts under all three measures shown in figure 5.1. The improvements are most significant for small training samples, although are still evident for training samples of size 50 and 60, which are commonly used in real-world scenarios. Indeed, the coverage of the bootstrap forecasts is closer to 95% than the plug-in forecasts for all training sample sizes considered. The relatively large improvements in the ignorance score compared to the continuous ranked probability score are due to the sensitivity of that score to observations that lie in the tails of the probability forecast distributions. The bootstrap method does not fully correct for the effect of parameter uncertainty on the width of prediction intervals, which, on average, remain too narrow.

In figure 5.2 we show the PIT histograms for the plug-in and bootstrap forecasts for training samples of size 30 and 60. In keeping with the coverage of the 95% prediction intervals discussed above, the PIT histograms show that the predictive bootstrap does not fully correct for the underdispersion of the standard plug-in forecasts. It is a little disappointing that the predictive bootstrap yields only a small improvement in the PIT histograms for training samples of size $N = 60$. Nonetheless, considering figures 5.1 and 5.2 in combination, the gain in forecast skill that can be attributed to accounting for parameter uncertainty with the predictive bootstrap in this idealised setting is sufficiently encouraging to suggest that the method may be beneficial in real-world scenarios.

We conclude our simulation experiment with an illustration of accounting for param-

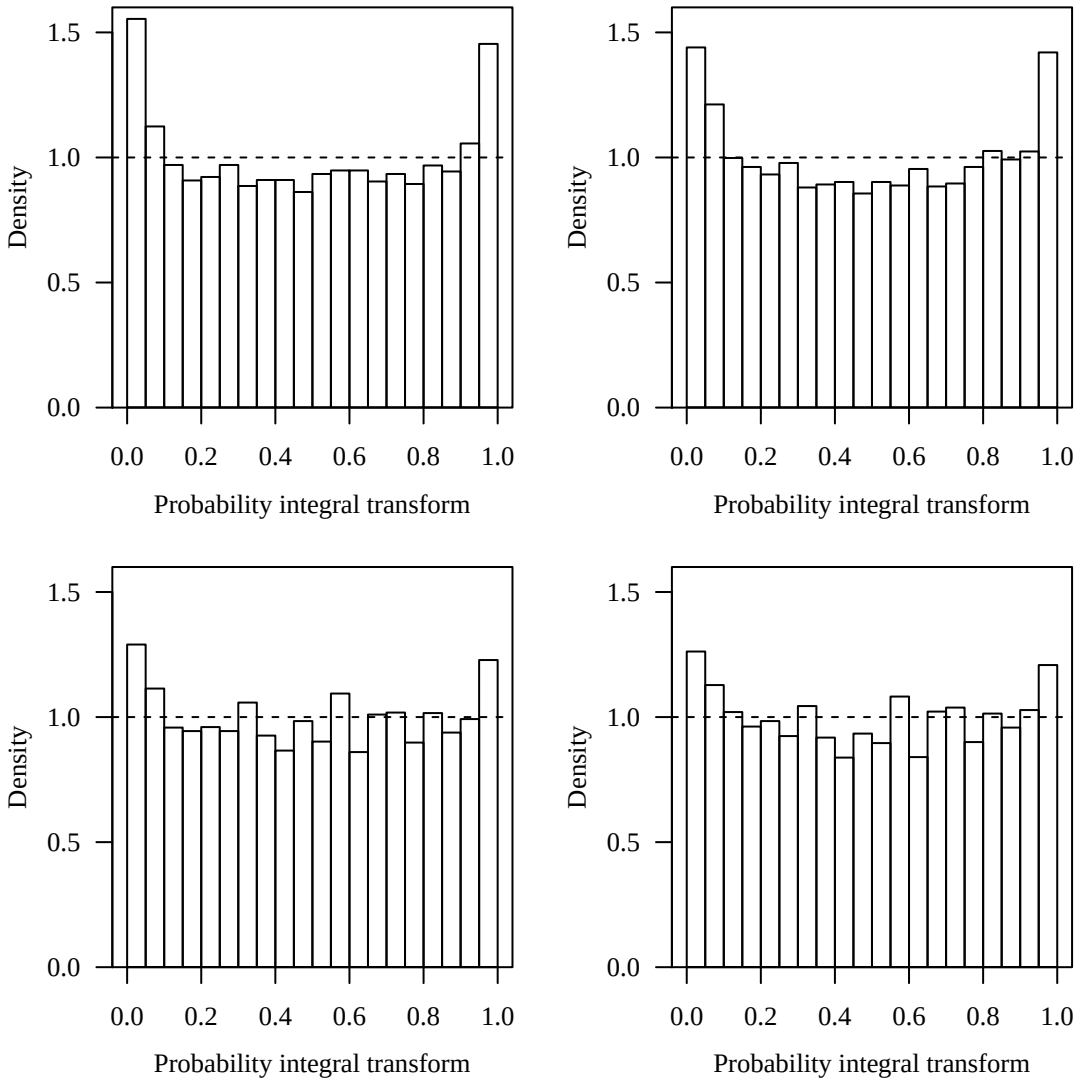

Figure 5.2 PIT histograms for the plug-in NGR forecast distributions (left) and bootstrap forecast distributions (right), for training samples of size $N = 30$ (top row) and $N = 60$ (bottom row).

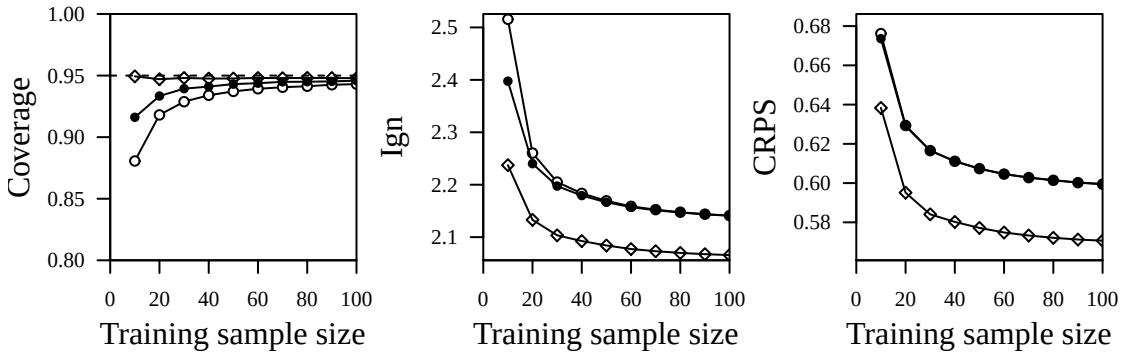

Figure 5.3 As for figure 5.1, but for observations that are distributed according to the MOS statistical model. The results of the analytic forecasts are shown as open diamonds.

eter uncertainty with the predictive bootstrap for the more simplistic MOS post-processing method, for which we also have the analytic result for accounting for parameter uncertainty given in section 5.2.1. The simulation experiment proceeds as described above, except that the observations are drawn from Gaussian distributions with constant variance, c^2, in keeping with the assumptions of the MOS method. We set $c^2 = 1.098$, the mean variance of the observations y_i in the simulation experiment for the NGR post-processing method described above.

In figure 5.3 we show the coverage of 95% prediction intervals, the ignorance score and the CRPS for the plug-in, bootstrap and analytic forecast distributions, where 'analytic' refers to the result presented in section 5.2.1 for MOS forecasts. The results indicate that accounting for parameter uncertainty with the predictive bootstrap method yields more skilful probability forecasts than the standard plug-in forecasts, although the analytic T-distributed forecasts are superior in skill. Again the predictive bootstrap does not fully correct for the underdispersion of the plug-in forecasts — on average the prediction intervals remain too narrow. The skill of the predictive bootstrap and plug-in probability forecasts converge more rapidly for the MOS post-processing method than for NGR, and the assertion by Glahn et al. [2009] that accounting for parameter uncertainty is not important when training samples are larger than $N = 30$ appears to be justified. However, were the analytic result given in section 5.2.1 not known, the predictive bootstrap would remain a useful addition to the out-of-sample probability forecasts.

5.4.2 Parameter uncertainty in 2-metre temperature forecasts

We now present the results of an investigation in to the effect of accounting for parameter uncertainty with the predictive bootstrap method for probability forecasts

of 2-metre temperature observations. The observations and 10-member ensemble forecasts were taken from version 2 of the GEFS reforecast project [Hamill et al., 2013], described in section 2.6. The forecast lead time was chosen as 48 hours, at a grid point located near to New York City, USA (40 degrees North, 74 degrees West). Ensemble forecasts were issued at 00 UTC, which corresponds to 20:00 or 19:00 local time, depending on the time of year. Out-of-sample probability forecast distributions were issued for the period 26 May 1990 – 15 September 2014 inclusive, using a rolling training sample of the previous N ensemble forecasts and observations, where N denotes the training sample size. The results presented below are therefore calculated using a total of $T = 8879$ out-of-sample forecasts and verifying observations. We show results using $N_B = 100$ bootstrap replicates. As mentioned in the previous subsection, we found forecast skill to be similar for $50, 100$ and 200 replicates, and so do not comment further on the value of N_B. Many of the following results are also presented in Siegert et al. [2015a, section 3.3].

Preliminary investigations (not shown) demonstrated an improvement in the skill of probability forecasts when using the NGR post-processing method compared with the more simplistic MOS method discussed in section 5.2.1. The so-called plug-in NGR forecasts are more skilful than MOS forecasts when both neglecting and accounting for the effect of parameter uncertainty in the MOS forecasts. This provides clear motivation for investigating the effect of accounting for parameter uncertainty with the predictive bootstrap on probability forecasts issued with the NGR post-processing method.

5.4.2.1 Choosing the bootstrap resampling approach

We found that the 'case resampling' approach for estimating the sampling distribution of the parameter estimates resulted in probability forecasts that were slightly more skilful than those obtained with either the approach of residual resampling or the parametric bootstrap. This is thought to be a result of some misspecification of the statistical model issued by the NGR post-processing method. We provide evidence for this claim later in this section (see figure 5.8). As discussed in section 5.2.2, the approach of case resampling places the least confidence in the statistical model, and is therefore the most resilient to violations of modelling assumptions. However, we stress that the differences in the measures of forecast skill are small, and that all three approaches improve on the standard 'plug-in' forecasts. At lag 1 the sample mean of our correlation values was approximately 0.15, and correlations decayed to 0 for longer lags. Therefore, while there is evidence of some temporal correlation in the standardised residuals, the correlations do not amount to a statistically significant violation of the NGR model assumptions. However, we applied the block bootstrap,

using a block length of 2, so as to investigate whether there was anything to be gained in its use. As with the residual and parametric bootstrap approaches, however, we found that the skill of the out-of-sample probability forecasts was slightly worse than when using the simple case resampling approach. This was also the case for longer block lengths. In what follows, therefore, we show results for the approach of case resampling only. We suggest that the block bootstrap may yield improved probability forecasts if a time series model that allows for temporal correlation in the residuals was used in place of the standard NGR post-processing model, which assumes implicitly that the standardised residuals are independent. Indeed, the approach of case resampling, in which the ensemble-observation pairs are sampled independently from the original training sample, adopts this assumption, which may serve to explain the improved forecast skill over the block bootstrap approach.

5.4.2.2 Comparing plug-in and bootstrap probability forecasts for 2-metre temperature

In figure 5.4 we show the coverage of the 95% prediction intervals, the ignorance score and the continuous ranked probability score for the out-of-sample probability forecasts as a function of the training sample size, N, before and after accounting for parameter uncertainty with the predictive bootstrap method. The coverage of the 95% prediction intervals is more accurate for the bootstrap forecast distributions, particularly for small training samples, which are also the training samples that are prefered by the ignorance score and CRPS (we discuss this point further below). However, as we found in our simulation experiment, the bootstrap forecast distributions on average remain underdispersed — the 95% prediction intervals remain too narrow. The improvements in both the ignorance score and the CRPS for the bootstrap forecasts are most evident for small training samples, but also persist for large training samples. It is encouraging to note that the predictive bootstrap forecasts yield an ignorance score for training samples of size $N = 20$ that is less than that of the plug-in forecasts with training samples of size $N = 50$. As noted in section 5.4.1, the large improvements in the ignorance score relative to the CRPS are due to the sensitivity of the ignorance score to the overpopulation of observations in the tails of the plug-in forecast distributions which, to some extent, is corrected by the predictive bootstrap (as evidenced by the coverage of the prediction intervals). Using the quantity $2^{\text{Ign}_A - \text{Ign}_B}$, where Ign_A and Ign_B denote the ignorance scores for forecasts before and after accounting for parameter uncertainty, respectively, we find that for the optimal training samples (discussed further below) the forecasts that account for parameter uncertainty assign on average 6% more probability density to the verifying observations than the plug-in forecasts.

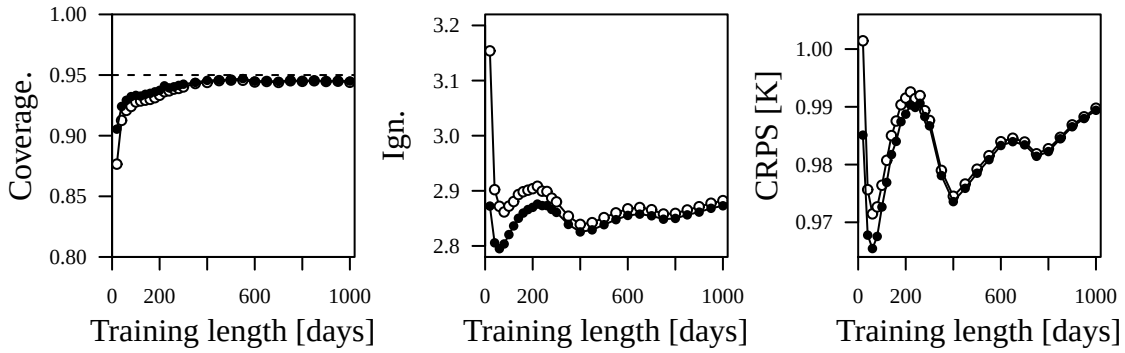

Figure 5.4 Coverage of the 95% prediction intervals, the ignorance score (Ign) and the CRPS for the plug-in forecasts (open circles) and the predictive bootstrap forecasts (filled circles) as a function of training sample size, N, for probability forecasts of 2-metre temperature.

The plots of the ignorance score and CRPS as a function of training sample size presented in figure 5.4 indicate that the training sample size is a highly influential factor in the skill of the out-of-sample probability forecasts, for both the standard plug-in and the predictive bootstrap forecast distributions. Contrary to the belief that larger training samples are preferable, the scores indicate a strong cyclical, or seasonal effect of training sample size on the forecast skill. Training samples that contain forecasts that are relevant to the time of year perform well in comparison to training samples that contain, for example, the previous six months of forecasts and observations. To illustrate this point, consider the plug-in forecasts. The optimal training sample size as measured by the CRPS is $N = 70$, while $N = 400$ (i.e. a similar time frame, but with an additional year of training data) is optimal for the ignorance score. When accounting for parameter uncertainty with the predictive bootstrap, however, training samples of size $N = 50$ are optimal for the CRPS, while $N = 60$ is optimal for the ignorance score. Indeed, for the predictive bootstrap forecasts, the values of both scores are very similar for training samples of size 40–70. Our findings therefore indicate that accounting for parameter uncertainty with the predictive bootstrap may help to reduce the optimal training sample size. This could prove an important benefit for users who may have access to limited data and/or computational resources. For example, while in this thesis we have concentrated on short-range weather forecasts, it is also necessary to post-process ensemble forecasts of seasonal or climate forecasts, for which far less training data is typically available. See Siegert et al. [2015a, sections 3.1 and 3.2] for examples of the method in post-processing ensemble forecasts on seasonal and decadal time scales.

Figure 5.5 shows PIT histograms for the plug-in and bootstrap probability forecasts for training samples of size $N = 60$, which were found to be almost optimal as measured by the ignorance score and CRPS. In figure 5.6 we show the corresponding

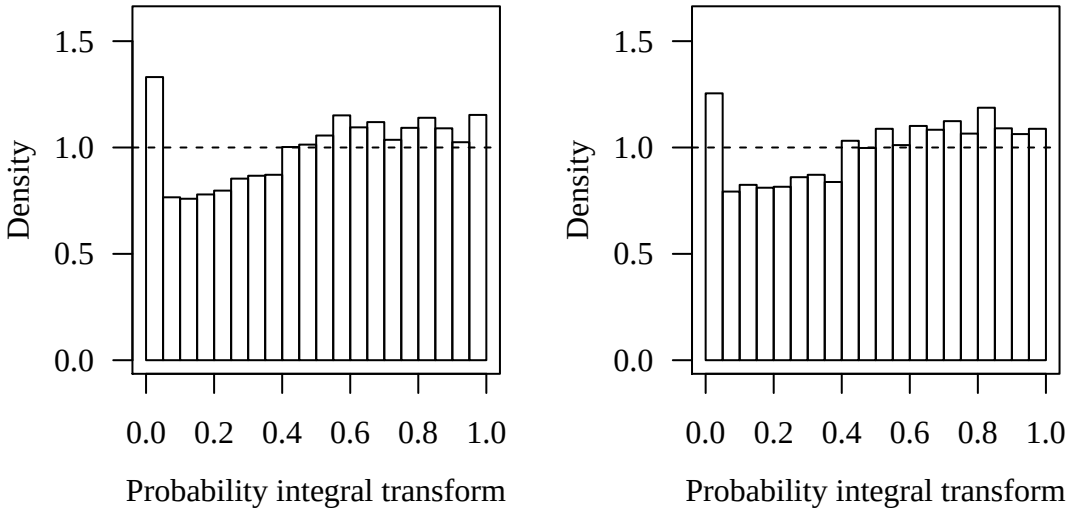

Figure 5.5 PIT histograms for probability forecasts of 2-metre temperature using plug-in (left) and predictive bootstrap (right) forecasts. Model parameters are estimated using rolling training samples of the previous 60 ensemble forecasts and observations.

PIT histograms for the smaller training samples of size $N = 30$. For both sizes of training sample the outermost PIT histogram bins for the plug-in forecasts are significantly overpopulated, particularly for the shorter training samples. This is indicative of probability forecast distributions whose tails are too light, in keeping with our findings in the simulation experiment and the explanation provided in section 5.1. Furthermore, the non-uniformity of the central bins of the PIT histograms suggests remaining biases in the location and/or dispersion of the NGR probability forecast distributions — more observations fall in the upper tails of the forecast distributions than the lower tails. In keeping with the coverage of the prediction intervals shown in figure 5.4, the PIT histograms for the bootstrap forecasts indicate that accounting for parameter uncertainty with the predictive bootstrap has not fully corrected for the aforementioned underdispersion of the plug-in forecasts, although the overpopulation of the outermost bins of the PIT histograms is reduced.

Finally, we turn to an assessment of the skill of the plug-in and bootstrap probability forecast distributions in issuing probability forecasts of the form $p = \Pr(y \leq q)$, where, as in chapter 3, y is the verifying observation and q is a threshold of interest. In table 5.1 we show the Brier score and the reliability and resolution components of its decomposition (see section 2.5.5 for details) for five thresholds of interest, namely the quantiles $q_{1/20}, q_{1/10}, q_{1/5}, q_{1/3}$ and $q_{1/2}$ of the climatological distribution of the observations over the period of investigation, where the notation q_α refers to the $100 \times \alpha\%$ quantile. Recall that smaller Brier and reliability scores, and larger resolution scores, are preferred. We show the scores for training samples of size $N = 60$. The qualitative features of the scores are similar for other training sample sizes (not shown).

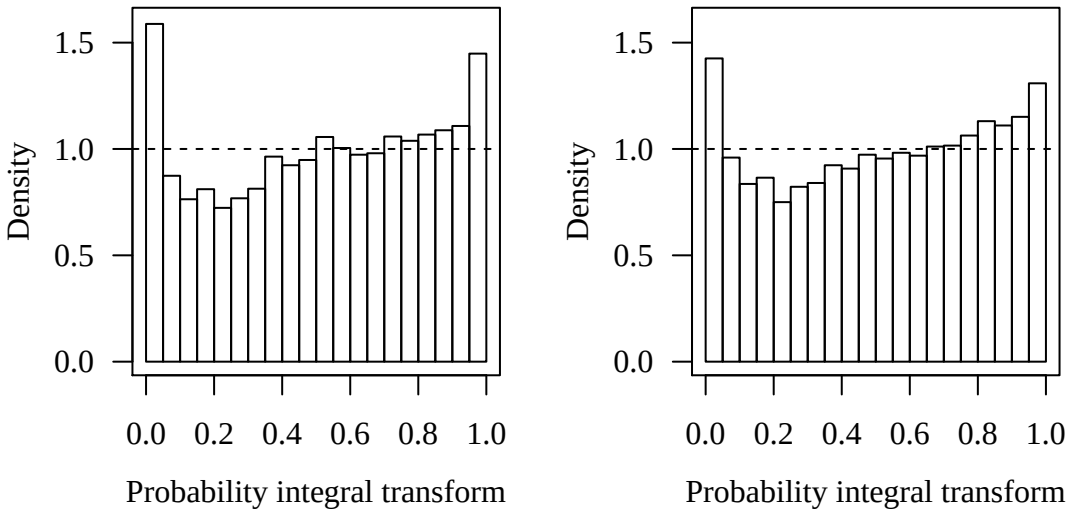

Figure 5.6 As for figure 5.5, but for training samples of size $N = 30$.

		Plug-in			Bootstrap	
	Brier	Rel	Res	Brier	Rel	Res
$q_{1/20}$	186	200	289	185	150	289
$q_{1/10}$	297	379	603	295	341	603
$q_{1/5}$	353	229	125	350	160	125
$q_{1/3}$	342	123	1877	340	105	1879
$q_{1/2}$	283	147	2214	280	137	2217

Table 5.1 Brier scores and the reliability (Rel) and resolution (Res) components of their decomposition, calculated at five thresholds of interest, for training samples of size $N = 60$. The Brier scores and the resolution components are scaled by 10^4, and the reliability components are scaled by 10^6.

The effect of the predictive bootstrap is most evident for the reliability scores, which indicate that accounting for parameter uncertainty improves the forecast reliability at each of the five thresholds considered here. It is encouraging that the Brier scores all improve when accounting for parameter uncertainty with the predictive bootstrap, although we note that the differences between the Brier scores for the bootstrap and plug-in forecasts are small relative to the differences in the ignorance scores and the CRPS shown in figure 5.4. Interestingly, despite the fact that the bootstrap forecasts yield probability forecast distributions that are more dispersed than the corresponding plug-in distributions, the resolution components of the Brier score decomposition are very similar. Unfortunately, the standard errors of the scores reported here are too large for us to report a statistically significant improvement in the scores for the predictive bootstrap forecasts. The scores exhibit some temporal correlation (not shown), and so we approximate the standard errors using the expression given in section 2.5.3 [Wilks, 2006b, pp 144-145]. However, in combination with the other results presented in this subsection, it is evident that accounting for parameter uncertainty with the predictive bootstrap in this real-world scenario results in more skilful probability forecasts than are issued with the standard plug-in approach.

5.4.2.3 Analysis of forecast residuals

Earlier in this subsection we commented that the statistical model specified by the NGR post-processing method is in some cases an inadequate fit to the data. We now illustrate this claim with two diagnostic plots. The use of such plots was discussed in section 2.5.2.1. In figure 5.7 we plot the residuals $r_t = \mu_t - y_t$ of the out-of-sample forecasts, where y_t and μ_t are the verifying observations and the expectation of the NGR plug-in forecasts, respectively, against the ensemble mean $\bar{x}_t$, for $t = 1, 2, \ldots, T$, where $T = 8879$ denotes the size of the test dataset for which out-of-sample forecasts are issued. A nonparametric approximation to the expectation of the residuals is shown in red, using the Loess function described in section 2.5.2.1. The residuals r_t should be symmetrically distributed around the line $r = 0$ if the forecast means $\mu_t = \hat{a} + \hat{b}\bar{x}_t$ are well-calibrated with the observations. Figure 5.7 indicates that the observations are generally well estimated by μ_t, although there is some evidence of miscalibration for small values of $\bar{x}$.

More interestingly, in figure 5.8 we plot the squared standardised residuals,

$$\text{ssr}_t = \left(\frac{y_t - \mu_t}{\sigma_t} \right)^2,$$

against the ensemble standard deviation, s_t. If the NGR forecast variance, $\sigma_t^2 =$

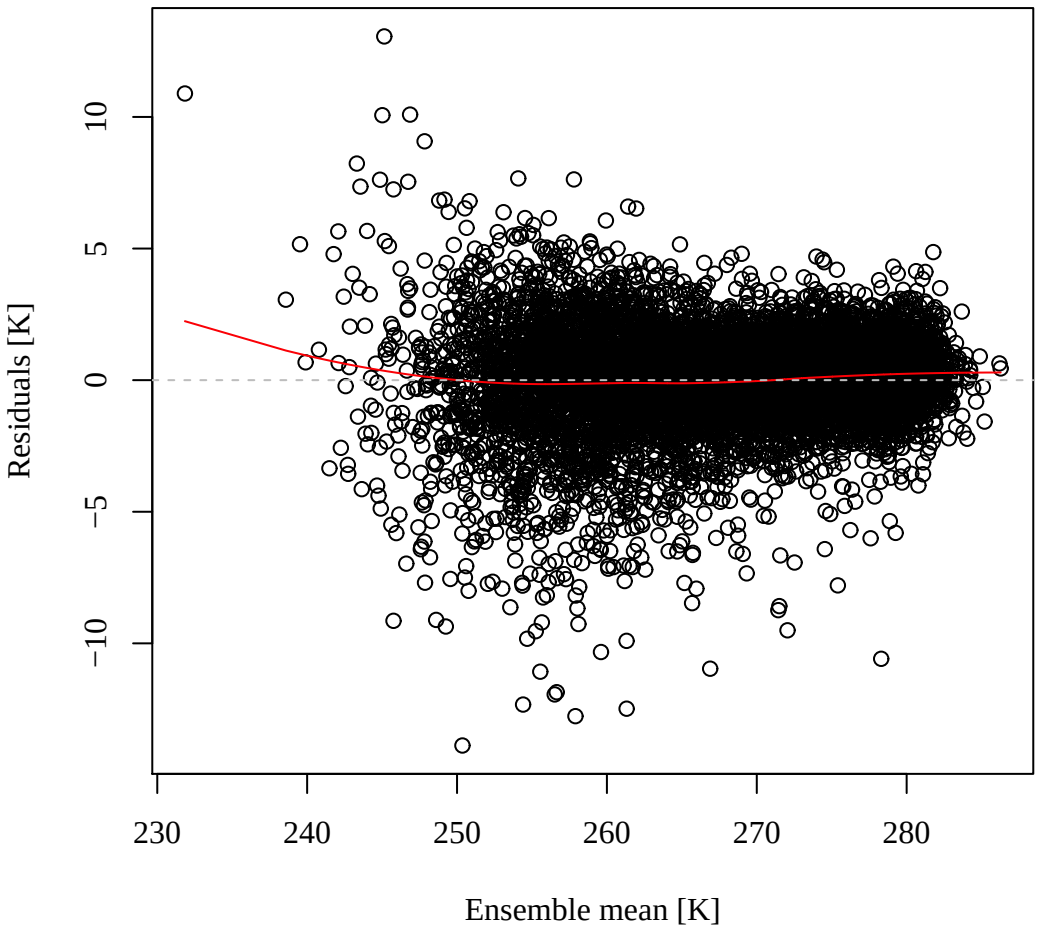

Figure 5.7 Residuals r_t as a function of the ensemble mean $\bar{x}_t$. A nonparametric Loess approximation to the expectation of the residuals is shown in red.

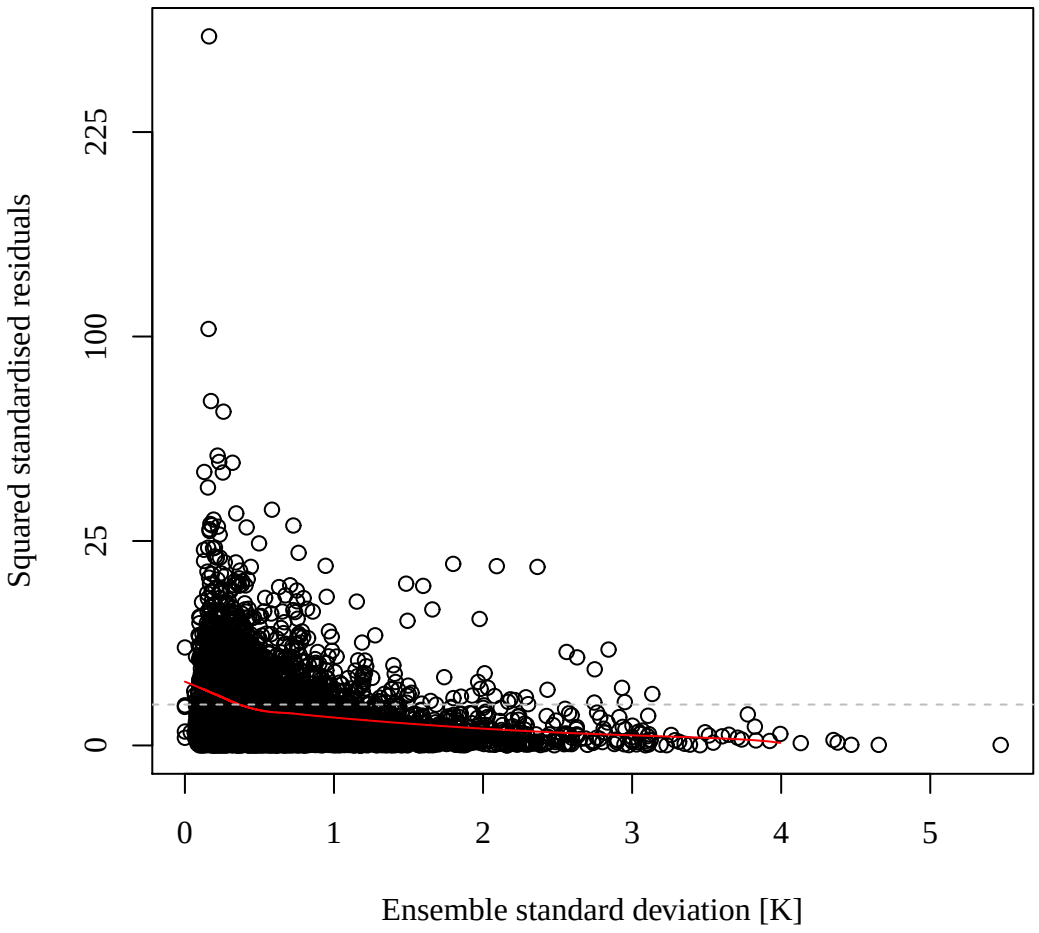

Figure 5.8 Squared standardised residuals ssr_t as a function of the ensemble standard deviation s_t. A Loess approximation to the expectation of ssr is shown in red. The vertical axis is plotted on a square root scale.

$\hat{c} + \hat{d}s_t^2$ is a well-calibrated estimate of the squared residuals, $r_t^2 = (\mu_t - y_t)^2$, it follows that the expectation of ssr_t is 1. Significant deviations from the line ssr = 1 are indicative of a misspecified statistical model. The plot in figure 5.8 shows that while this relationship is satisfied for much of the data, a significant number of the squared standardised residuals lie far from the line ssr = 1. In chapter 6 we illustrate how such plots can be used to improve the statistical model assumed by the NGR post-processing method.

5.4.3 Accounting for parameter uncertainty in post-processed ensemble forecasts of temperature and air pressure

As a final example we consider post-processed ensemble forecasts, rather than probability forecasts, where the ensemble forecasts are sampled from probability forecast distributions as described in section 2.4.6 and as also illustrated in chapter 4. We compare ensemble forecasts that are sampled from the plug-in and predictive bootstrap forecast distributions, using the Bayesian model averaging (BMA) mixture distributions (see section 2.4.3.2). We use $M = 50$-member ensemble forecasts issued by the European Centre for Medium-Range Weather Forecasts (ECMWF) for Berlin, Frankfurt and Hamburg airports, for the period 1 May 2010 – 30 April 2011 inclusive. Observations were provided by the Deutscher Wetterdienst (DWD). This dataset was also used in the article Schefzik et al. [2013], which presented the idea of ensemble copula coupling (ECC), described in section 2.4.6 for producing calibrated ensemble forecasts over spatial fields. As in Schefzik et al. [2013], parameter estimates for the plug-in and bootstrap probability forecast distributions are estimated using rolling training samples of the previous 30 days. The results presented in the previous two subsections therefore indicate that we might reasonably expect to improve the forecast skill with the predictive bootstrap method, due to the relatively small training sample size. We again use case resampling with $N_B = 100$ bootstrap replicates.

We correct ensemble forecasts for biases in their location using the linear correction (LC) ensemble adjustment scheme, introduced in section 3.3.2. Specifically, the CDF of the plug-in forecast distribution at forecast time t is

$$F(y_t \mid \boldsymbol{x}_t, \hat{\psi}(D)) = \frac{1}{M} \sum_{m=1}^{M} \Phi\left(\frac{y_t - \hat{x}_{m,t}}{\sigma_t^{\text{BMA}}}\right), \tag{5.27}$$

where $\Phi(\cdot)$ denotes the standard Gaussian CDF, σ_t^{BMA} is the dressing kernel variance estimated from the previous N ensemble forecasts and verifying observations, and

		ECMWF	Plug-in	Bootstrap
Temperature	Berlin	1.249	0.964	0.956
	Frankfurt	1.263	0.922	0.923
	Hamburg	1.062	0.910	0.905
Pressure	Berlin	0.796	0.714	0.714
	Frankfurt	0.782	0.715	0.707
	Hamburg	0.771	0.700	0.689

Table 5.2 The CRPS for the raw ensemble forecasts, and ensembles sampled from plug-in and bootstrap BMA forecast distributions at Berlin, Frankfurt and Hamburg airports for the period 1 May 2010 – 30 April 2011.

the adjusted ensemble member $\hat{x}_{m,t}$ is given by

$$\hat{x}_{m,t} = x_{m,t} + \hat{a} + (\hat{b} - 1)\bar{x}_t.$$

The CDF of the bootstrap forecast distributions is

$$F(y_t \mid \boldsymbol{x}_t, \psi) \approx \frac{1}{M \times N_B} \sum_{j=1}^{N_B} \sum_{m=1}^{M} \Phi\left(\frac{y_t - \hat{x}^*_{j,m,t}}{\sigma^{\text{BMA}^*}_{j,t}}\right), \tag{5.28}$$

where $\hat{x}^*_{j,m,t}$ and $\sigma^{\text{BMA}^*}_{j,t}$ denote the adjusted ensemble members and the estimated dressing kernel variance for the jth bootstrap replicate. Preliminary investigations (not shown) indicated that forecast skill did not improve under the more flexible LCR scheme that also permits a rescaling of the ensemble forecasts, and so we use the LC scheme for reasons of parsimony.

Ensemble forecasts were sampled from the plug-in BMA and predictive bootstrap probability forecast distributions. The M ensemble members were taken as the equidistant $100 \times m/(M+1)\%$ quantiles of the probability forecast distributions, for $m = 1, 2, \ldots, M$. The quantiles were calculated numerically as described in section 5.3. To facilitate a further comparison of our results with those of Schefzik et al. [2013], the ECC methodology was applied to the post-processed ensemble forecasts at the three airports. Recall from section 2.4.6 that Schefzik et al. [2013] found that by preserving the rank dependence structure of the raw ensemble forecasts with the ECC methodology, the multivariate ensemble forecasts improved in skill compared to those that did not preserve the rank dependence, particularly for forecasts of air pressure. In table 5.2 we show the mean CRPS for the three individual airports, and in table 5.3 we show the mean energy scores for multivariate forecasts of the spatial field defined by their location, for both temperature and air pressure.

The results presented in tables 5.2 and 5.3 indicate that the predictive bootstrap leads to improvements in the skill of ensemble forecasts, at both individual locations (as measured by the CRPS), and over the spatial field defined by the three airports

		Before ECC		After ECC	
	ECMWF	Plug-in	Bootstrap	Plug-in	Bootstrap
Temperature	2.342	1.979	1.969	1.875	1.865
Pressure	1.478	1.401	1.382	1.371	1.353

Table 5.3 The energy scores for the raw ensemble forecasts and ensembles sampled from plug-in and bootstrap BMA forecast distributions for the spatial field defined by Berlin, Frankfurt and Hamburg airports, for the period 1 May 2010 – 30 April 2011.

(as measured by the energy scores). For spatial forecasts, the improvements in forecast skill (as measured by the energy scores) due to the use of the predictive bootstrap are small relative to those due to the ECC methodology for temperature, but of similar order for air pressure. For individual airports, the standard errors of the CRPS values reported in table 5.2 (not shown) are too large to conclude that the predictive bootstrap yields a statistically significant improvement for any of the six forecasts considered. Typically the standard errors are of the order 3×10^{-2} for air pressure forecasts and 5×10^{-2} for temperature forecasts. Statistically significant results may exist for some forecasts (most noteably for air pressure) if the analysis was extended to a larger dataset of approximately eight years of forecasts and observations, provided that the improvements in forecast skill were similar for that dataset. Similarly, the standard errors of the energy scores preclude us from reporting a statistically significant improvement in forecast skill when using the predictive bootstrap. The standard errors are of the order 4×10^{-2} for forecasts of air pressure and 5×10^{-2} for forecasts of temperature. However, a statistically significant improvement in the energy score of air pressure forecasts would exist with the combination of the predictive bootstrap and ECC methodology, were an additional two years of data available for analysis and the improvements in forecast skill were to persist. The ECC methodology alone yields a statistically significant improvement in the energy score for temperature forecasts, but not for air pressure forecasts.

The results of our plug-in forecasts differ slightly from those of Schefzik et al. [2013]. In particular, our plug-in BMA forecasts for temperature produce a worse energy score, but improve significantly with the application of the ECC methodology, unlike the results of Schefzik et al. [2013] who reported that ECC did not improve the skill of temperature forecasts. We note with interest that the energy scores of our plug-in forecasts for both temperature and air pressure improve on those given in Schefzik et al. [2013] after ECC, while without ECC we achieve a worse energy score. These differences presumably derive from the model fitting procedure. Schefzik et al. [2013] use the 'EnsembleBMA' R package [Fraley et al., 2007] to estimate model parameters. In this case, each member of ensemble forecast $\boldsymbol{x}_t$ is recalibrated to $\hat{x}_{m,t} = \hat{a} + \hat{b}x_{m,t}$, for $m = 1, 2, \ldots, M$, such that the sample variance of the adjusted ensemble forecast, $\hat{\boldsymbol{x}}_t$ is $\hat{b}^2 s_t^2$, where s_t^2 is the sample variance of $\boldsymbol{x}_t$, and

$\hat{b}$ is the parameter estimate found by regressing the training observations y_i on the ensemble members x_{im}, for $i = 1, 2, \ldots, N$. On the other hand, we form adjusted ensemble forecasts using the LC scheme introduced in section 3.3.2, such that $\hat{x}_{m,t} = x_{m,t} + \hat{a} + (\hat{b} - 1)\bar{x}_t$, where $\bar{x}_t$ is the sample mean of $\boldsymbol{x}_t$. In this case the ensemble variance remains unaltered (see section 3.3.2). Furthermore, Schefzik et al. [2013] use the expectation-maximisation algorithm for parameter estimation, whereas we use the Nelder-Mead and BFGS algorithms as described in section 2.4.7.3.

5.5 Discussion and conclusions

In this chapter we have highlighted that users of ensemble post-processing methods should acknowledge and account for the effects of parameter uncertainty on the probability forecast distributions that are issued for the future, verifying observations. We have proposed an approach for accommodating parameter uncertainty in the statistical models that are used for ensemble post-processing that is based on bootstrap resampling, known as the predictive bootstrap, that is easy to implement and is applicable to a large variety of ensemble post-processing methods. In three case studies it was shown that accounting for parameter uncertainty with the predictive bootstrap yielded more skilful probability forecasts than were obtained from the standard practice of direct substitution of parameter estimates in to the chosen parametric family of probability distributions. Accounting for parameter uncertainty with the predictive bootstrap method resulted in significant improvements to the ignorance score, due to the sensitivity of this score to observations that lie in the tails of the probability forecast distributions. Encouragingly, the predictive bootstrap was also shown to result in more uniform PIT histograms, and improvements to the reliability component of the Brier score decomposition, calculated at various thresholds of interest. Unlike Glahn et al. [2009], who asserted that parameter uncertainty was not important for training samples larger than 30 forecasts and observations, we found that the predictive bootstrap resulted in improvements to forecast skill for both small and large training samples. Our investigation of 2-metre temperature forecasts also indicated that the predictive bootstrap may reduce the optimal training sample size, which could be an important result for forecasters with limited data resources.

Three approaches (case and residual resampling, and the parametric bootstrap) were proposed for estimating the sampling distribution of the model parameter estimates. Case resampling was found to be the best option for the two real-world examples provided in this chapter, which we attribute to some misspecification of the underlying statistical models. However, we suggest that users should also investigate the other approaches suggested, in order to determine the approach that is most appropriate

for their particular case. The block bootstrap was also suggested as an approach that could be used if residuals from the fitted model exhibit significant temporal autocorrelation, although we found it to be ineffective for our examples. We suggest that statistical models that incorporate temporally correlated residuals may be of benefit here, and expect that the block bootstrap may then prove beneficial.

We were slightly disappointed that the predictive bootstrap did not yield more significant improvements to the probabilistic forecast skill, particularly for the simulation experiment in which the statistical model was correctly specified. However, we have at least shown that parameter uncertainty is a potentially important factor in the skill of probability forecasts, and one that should in general not be ignored, as has historically been the case. We suggest that refinements to the resampling procedures for approximating the sampling distribution of parameter estimates may result in more skilful bootstrap forecast distributions. For example, ensemble forecasts in the training sample that are similar to the current, out-of-sample ensemble forecast could be weighted such that they are resampled more frequently than those ensemble forecasts that bear little resemblance to the current forecast. We encourage the investigation of such resampling schemes.

An alternative approach to estimating and accounting for parameter uncertainty is to formulate a Bayesian model, in which case parameter uncertainty is an inherent feature. Friederichs and Thorarinsdottir [2012] introduced a novel Bayesian approach for probability forecasts of extreme windspeeds using the generalised extreme value distribution, and Siegert et al. [2015b] proposed a Bayesian approach for the post-processing of ensemble forecasts on longer time scales that are associated with climate science. Comparisons between the predictive bootstrap and Bayesian models are strongly encouraged. However, unlike the predictive bootstrap, a possible disadvantage of Bayesian models is the need to specify prior distributions for the parameter estimates. As in Friederichs and Thorarinsdottir [2012] informative prior distributions may be difficult to obtain, and the choice of prior may significantly affect the skill of the probability forecasts. On the other hand, an appealing feature of the predictive bootstrap is its flexibility and applicability to a wide range of forecast distributions. Bayesian models are also likely to require the use of Markov Chain Monte Carlo algorithms, which could prove computationally burdensome for forecasting systems of many weather variables, locations, and forecast lead times.

A third possibility of estimating the sampling distribution of the parameter estimates is to exploit known asymptotic properties of their sampling distributions in special cases, such as the well-known asymptotic normality of likelihood parameter estimates. With the exception of some irregular cases, likelihood-based parameter estimates are asymptotically unbiased and normally distributed, with a covariance matrix that can be calculated directly from the likelihood function. We could there-

fore simulate parameter estimates from the asymptotic sampling distribution, and obtain forecast distributions in an analogous manner to those of the predictive bootstrap method. However, preliminary results indicate that this approach results in less skilful forecasts.

6 Improving model specification, effects of ensemble member dependence, and closing remarks

6.1 Improving model specification with diagnostic plots

6.1.1 Introduction and motivation

In this section we provide an example of using diagnostic plots to improve the skill of probability forecasts issued with the NGR ensemble post-processing method. The potential for using such plots to diagnose and improve the specification of statistical models is discussed in section 2.5.2.1. We construct revised models for the variance of the NGR forecast distributions, and compare the skill of the resulting Gaussian probability forecast distributions with those issued by the 'standard' NGR post-processing method introduced by Gneiting et al. [2005] (see equations (2.19) and (2.20) on page 38).

In chapter 5 we suggested that the statistical model assumed by the NGR post-processing method could be improved, as in some instances the model is a poor fit to the data. We used diagnostic plots of the residuals and squared standardised residuals,

$$r_t = \mu_t - y_t,$$
$$\text{ssr}_t = \left(\frac{\mu_t - y_t}{\sigma_t}\right)^2,$$

where t indexes the forecast occasions in the test dataset, y_t is the verifying observation for forecast occasion t, and μ_t and σ_t^2 are the expectation and variance of the NGR probability forecast distribution, respectively. While we concluded that the

expectation of the NGR forecast distributions, μ, is sufficiently well specified, plotting the squared standardised residuals against the ensemble standard deviation, s (see figure 5.8) illustrated that the form of the NGR forecast variance, $\sigma^2 = \hat{c} + \hat{d}s^2$, is inadequately specified.

In order to illustrate how the statistical models specified by ensemble post-processing methods might be improved, we return to forecasting in the Lorenz 1996 system (Lorenz [1996], see section 2.6.1 for details). The Lorenz 1996 system was chosen so as our revised models could be formed in a data-rich setting for which we can be confident that our conclusions are not due to random chance and/or anomalous data. Our example is based on forecasts at lead time 4, although the procedure we follow and describe below is equally applicable to other lead times. As in chapter 3, model parameters are estimated using training samples from the first of the datasets described in section 2.6.1, that exhibits temporal correlation. The skill of the resulting probability forecast distributions is then assessed using the second of our datasets, that contains 190 000 ensemble forecasts and observations that are effectively independent. To begin, we show diagnostic plots for parameter estimates obtained with a training sample of size 100 000 forecasts and observations. This extremely large sample size was chosen in order to ensure that the effects of parameter uncertainty on the resulting out-of-sample probability forecasts are negligibly small. We show results for parameter estimates obtained by minimising the negative log-likelihood (NLL). However, in keeping with our findings in chapter 3, we found that the analogous results for parameter estimates obtained by minimising the continuous ranked probability score (CRPS, not shown), are qualitatively and quantitatively similar.

Recall that if μ_t and σ_t^2 correctly specify the expectation and variance of the distribution of the verifying observation y_t, then it follows that the expectation of the residual is $E(r_t) = 0$, and the expectation of the squared standardised residual is $E(\text{ssr}_t) = 1$. Significant deviations of the mean residuals from the line $r = 0$, and the mean squared standardised residuals from the line ssr $= 1$ are therefore indicative of a misspecified statistical model. In the following figures we approximate these expectations with a nonparametric fit to the scatter plots of residuals and squared standardised residuals, using the Loess function implemented in the R language [R Core Team, 2015] and described in section 2.5.2.1.

6.1.2 Results

In figure 6.1 we show the residuals r_t as a function of the ensemble mean $\bar{x}_t$, for $t = 1, 2, \ldots, T$, where $T = 190\,000$ denotes the size of the test dataset. The Loess curve indicates that the residuals r_t are approximately centred around the line $r = 0$, and the scatter plot indicates that the residuals are evenly spread throughout the

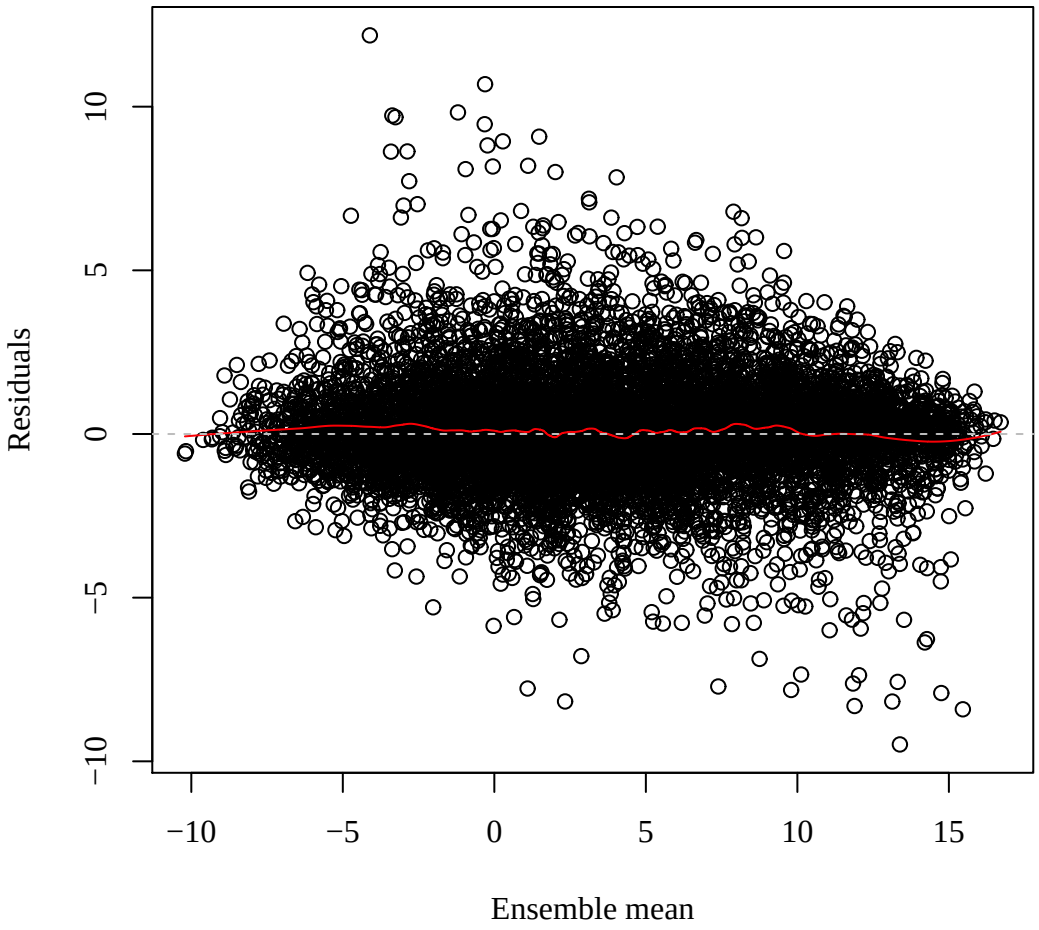

Figure 6.1 Forecast residuals r as a function of ensemble mean $\bar{x}$. A nonparametric approximation to the expectation $E(r)$ is shown in red.

range of $\bar{x}$. The distribution of the residuals r_t therefore appears to be independent of the ensemble mean $\bar{x}_t$ — there is no evidence of correlation between the value of $\bar{x}$ and either the expectation or variance of the residuals. We can conclude from this figure, therefore, that the forecast expectations $\mu_t = \hat{a} + \hat{b}\bar{x}_t$ are adequately specified.

With the forecast mean μ_t well specified, we now assess the adequacy of the forecast variance $\sigma_t^2 = \hat{c} + \hat{d}s_t^2$. Figure 6.2 shows the squared standardised residuals ssr_t as a function of the ensemble standard deviation s_t. In this case, the Loess curve deviates significantly from the line ssr = 1. The Loess curve indicates that the estimator $\sigma^2(s)$ is too large for small and large values of s, and too small for the remaining values. The linear specification of the NGR forecast variance is therefore an inadequate fit to the data under consideration.

We now illustrate how figure 6.2 can be used to improve the NGR post-processing

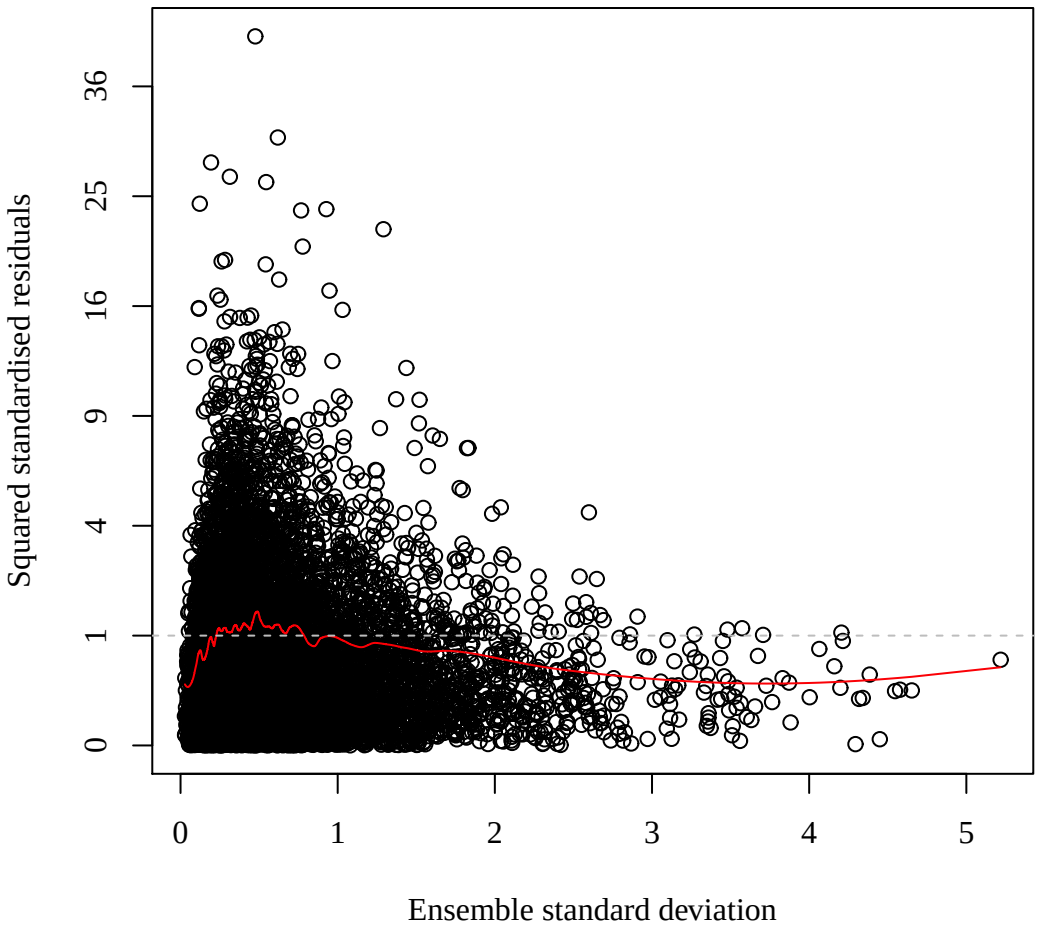

Figure 6.2 Squared standardised residuals ssr as a function of ensemble standard deviation s. A nonparametric approximation to the expectation $E(\text{ssr})$ is shown in red. The vertical axis is plotted on a square root scale.

method. We denote by $f(s)$ a functional form for the nonparametric Loess approximation to the expectation $E(\text{ssr}(s))$ of the squared standardised residuals as a function of the ensemble standard deviation s. If we can find a function $f(s)$ that is a good approximation to the Loess curve, therefore, we can form a revised model for the forecast variance, say σ_{f1}^2 as

$$\sigma_{f1}^2 = (c + ds^2)f(s), \tag{6.1}$$

where in general $f(s)$ depends on further parameters that are to be estimated. We can then form an estimate to σ_{f1}^2 as

$$\hat{\sigma}_{f1}^2 = (\hat{c} + \hat{d}s^2)\hat{f}(s), \tag{6.2}$$

where the notation $\hat{f}$ indicates our estimate of the function f by optimisation of an objective function. Returning to figure 6.2, a plausible approximation to the nonparametric estimate of $E(\text{ssr}^2(s)) = f(s)$ is

$$f(s) = 1 - (1 - ps)e^{-qs}, \tag{6.3}$$

where p and q are parameters that are to be estimated. Observe that the function $f(s)$ tends to 0 (1) as s tends to 0 (∞), implying that the estimator $\hat{\sigma}_{f1}^2(s)$ tends to 0 as s tends to 0, and to $c + ds^2$ as s tends to ∞. Despite the indication in figure 6.2 that we should specify a function $f(s)$ that tends to 0 as s tends to ∞, this seems unreasonable in the context of ensemble post-processing — it is counterintuitive to model the forecast variance as a decreasing function of the ensemble variance. In any case, we suggest that the Loess estimates for large s should be treated with caution due to the sparsity of data for such values.

Figure 6.3 shows the squared standardised residuals as a function of ensemble standard deviation, for the revised estimate of σ^2,

$$\hat{\sigma}_{f1,t}^2 = (\hat{c} + \hat{d}s_t^2) \times \left\{1 - (1 - \hat{p}s_t)e^{-\hat{q}s_t}\right\}, \text{ for } t = 1, 2, \ldots, T.$$

The Loess approximation to the expectation $E(\text{ssr}(s))$ now closely follows the line $\text{ssr}(s) = 1$, indicating an improvement to the variance of the Gaussian NGR forecast distributions.

Interestingly, upon convergence of both the Nelder-Mead and BFGS numerical algorithms (see section 2.4.7), the parameter estimate $\hat{q}$ is negligibly small. It therefore follows that our estimated function $\hat{f}(s)$ is approximately equal to $\hat{p}s$, and so $\hat{\sigma}_{f1,t}^2$ is approximately equal to $(\hat{c} + \hat{d}s_t^2)\hat{p}s_t$. For small values of s, such as the majority of those encountered in this example, the term $\hat{c}\hat{p}s$ dominates the term $\hat{d}\hat{p}s^3$. This

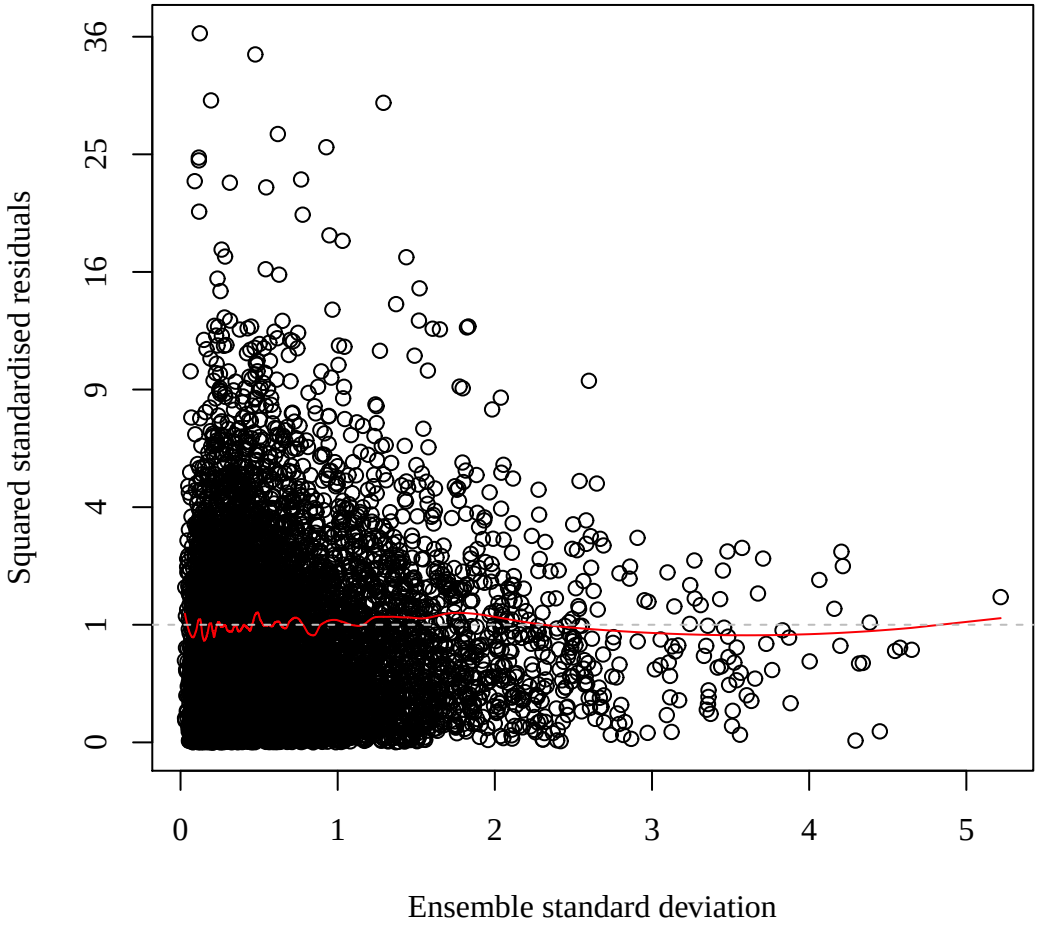

Figure 6.3 Squared standardised residuals ssr as a function of ensemble standard deviation s, for NGR forecast distributions with variance given by equations (6.1) and (6.3). A nonparametric approximation to the expectation $E(\text{ssr})$ is shown in red. The vertical axis is plotted on a square root scale.

motivates the idea of a second revised model for the NGR forecast variance,

$$\sigma_{f2}^2 = c + ds + es^2 + gs^3, \tag{6.4}$$

where now c, d, e and g are parameters that are to be estimated. However, both the Nelder-Mead and BFGS algorithms failed to converge for this model, with the 'optim' function reporting degeneracy of the Nelder-Mead simplex. We therefore reverted to the more parsimonious model

$$\sigma_{f3}^2 = c + ds + es^2. \tag{6.5}$$

The corresponding plot of squared standardised residuals for this model (not shown) is very similar to figure 6.3. The parameter estimate $\hat{c}$ is of the order 10^{-5}, and so the NGR forecast variance is approximately

$$\hat{\sigma}_{f3,t}^2 \approx 1.873 s_t + 0.636 s_t^2.$$

The estimate $\hat{\sigma}_{f3}^2$ therefore tends to 0 for small s, and is an increasing function of s. The negligibly small estimate $\hat{c}$ is of particular interest. Previously, in chapter 3 and Williams et al. [2014], we found that $\hat{c}$ was a crucial parameter in issuing skilful probability forecasts for the NGR model proposed by Gneiting et al. [2005] — we found that an NGR model with the constraint $c = 0$ was far less skilful than the standard linear function of the ensemble variance. However, our findings presented in the above figures indicate that allowing additional flexibility in the model for the forecast variance may lead to alternative conclusions, and yield forecast distributions whose moments are better calibrated with the properties of the verifying observations.

Note further that a similar form to the revised function for the forecast variance given in equation (6.5) would also be recovered if the forecast standard deviation σ were modelled as a linear function of the ensemble standard deviation s, in an analogous manner to the standard NGR formulation for the forecast variance [Gneiting et al., 2005]. Specifically, if we set $\sigma = c + ds$, then $\sigma^2 = c^2 + 2cds + d^2 s^2$, which is of the same functional form as equation (6.5), except that we allow additional flexibility in our model by using three, rather than two parameters.

As a further assessment of the forecast skill of our revised NGR post-processing method, we calculate the mean ignorance score and CRPS (see section 2.5.3.2 for details). Firstly, we consider the parameter estimates obtained from the large training sample of size $N = 100\,000$, and assess the skill of out-of-sample probability forecasts using the second, test dataset of size $T = 190\,000$. The values of the CRPS for the standard [Gneiting et al., 2005] and revised (equation (6.5)) NGR forecasts

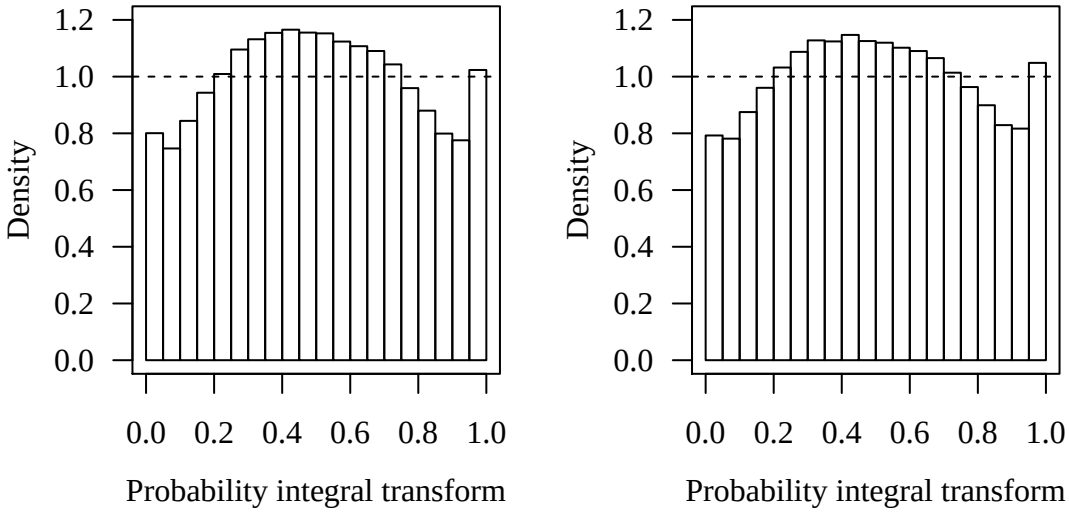

Figure 6.4 Pit histograms for the standard NGR model (left) and the revised model with variance given by equation (6.5) (right).

are 2.785 and 2.779, respectively, and the ignorance scores are 2.441 and 2.410. The standard errors of these scores are sufficiently small such that the differences are statistically significant at the 5% level.

In figure 6.4 we show the PIT histograms for the standard NGR model and our revised model with forecast variance given by equation (6.5). Perhaps surprisingly, differences in the PIT histograms are small, although the histogram for our revised model is slightly closer to uniformity than that of the standard linear model. Both histograms display an inverted U-shape pattern, and the rightmost histogram bin is overpopulated with observations. Given that parameter uncertainty is negligibly small in this example, and that we are confident in the calibration of the expectation and variance of the Gaussian forecast distributions issued by our revised NGR model, the PIT histograms may indicate misspecification of either higher moments of the forecast distributions, or the form of the distribution itself.

Finally, we compare the forecast skill of our revised NGR model with the standard model given in Gneiting et al. [2005] for small training samples that are likely to be encountered in practice. As in chapter 3 and Williams et al. [2014], we fit 500 instances of each model to training samples of size 100, where each training sample is effectively independent, while the data within the training sample are temporally correlated. We calculate verification scores for each of the 500 sets of parameter estimates over the test dataset of size 190 000, and report the mean of the 500 scores for the two models. The CRPS of the standard (revised) forecasts is 0.819 (0.805), and the ignorance scores are 2.615 (2.561). Again, the standard errors are sufficiently small so as the differences in these scores are statistically significant.

Using the quantity

$$2^{2.615-2.561},$$

the revised NGR forecast distributions on average assign 3.79% more density to the verifying observations than the standard NGR model.

6.1.3 Further comments and recommendations

In this section we have provided an illustrative example of the use of diagnostic plots to improve the statistical models that are specified by ensemble post-processing methods. Our example was based on a data-rich setting for which we had a clear indication of how the NGR forecast variance could be improved. It may be more difficult to suggest revisions to the statistical models in other scenarios, such as that presented in chapter 5, for which there is less data and no obvious indication of a functional form of ensemble covariates that captures the qualitative properties of the residual-based diagnostic plots. Indeed, the revised model given by equation (6.5) resulted in only minor improvements (not shown) to the forecast skill of the standard NGR model for those data. Nonetheless, we suggest that forecast users and other researchers should conduct similar diagnostic analyses, in order to determine whether their statistical models can be readily improved. While we do not know whether such analyses are conducted prior to publication, we are not aware of similar discussions in the literature. We therefore encourage authors to show, or at least comment on, diagnostic plots as well as the popular qualitative and quantitative assessments of forecast skill that are usually presented.

Our example has concentrated on revising the statistical model used for the NGR forecast variance. We have also conducted preliminary investigations in to the calibration of the skewness of the NGR forecast distribution, which (by definition of being Gaussian) is assumed to be 0. This was accomplished by plotting values of the probability integral transform (PIT), given by $F(y_t \mid \mu_t, \sigma_t^2)$, against the ensemble skewness, say τ_t. Nonparametric estimates of the quantiles of the distribution of PIT values as a function of ensemble skewness were then plotted (see section 2.5.2.4 for a brief discussion), in order to assess departures from uniformity as a function of the ensemble skewness. We found there to be little evidence of an obvious systematic relationship, and so did not pursue this line of enquiry. However, forecasters may wish to conduct similar analyses for other forecast scenarios.

Forecasters may also wish to check the assumption of temporally independent residuals — recall that the verifying observations should appear to be independent, conditional on the corresponding ensemble forecasts. This can be achieved by plotting the autocorrelation and partial autocorrelation functions of the residuals or standard-

ised residuals for different forecast lags. Significant values of the autocorrelation function indicate dependence in the forecast residuals, which could be addressed, for example, by fitting an autoregressive model. We found that the assumption of conditional independence was satisfied in the examples considered in this thesis.

6.2 Ensemble member dependence and forecast verification

As noted in section 2.2.4 and as used frequently throughout this thesis, it is popular to interpret ensemble members as independent and identically distributed (IID) realisations of underlying ensemble distributions. In section 2.5.4 we discussed fair scoring rules, which are appropriate for assessing the skill of ensemble forecasts under this interpretation. However, in that section we also stated that it is possible to 'hedge' ensemble forecasts in order to attain improved scores. In this section we therefore highlight the effect of dependencies between ensemble members on popular verification measures, specifically the rank histogram and the fair analogue of the continuous ranked probability score, denoted FCRPS (see sections 2.5.2.3 and 2.5.4).

In this section we illustrate the complexity of verifying post-processed ensemble forecasts. As mentioned in section 4.2.7, correcting for biases in the location of ensemble forecasts can induce dependence between the ensemble members, even if the members of the 'raw' ensemble forecasts truly are independent and identically distributed (IID). As we demonstrate below, such inter-member dependencies may lead to verification results that cause the user to make misleading conclusions. In this section we demonstrate the effect of inter-member dependence with the rank histogram and the fair analogue of the continuous ranked probability score (FCRPS).

We begin with a simulation experiment in which the ensemble members are drawn from a symmetric M-dimensional distribution. The ensemble members are therefore exchangeable (see definition 2.2.1). Specifically, we draw ensemble members from the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ, where

$$\Sigma = \begin{pmatrix} 1 & r & r & \dots & r \\ r & 1 & r & \dots & r \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r & r & r & \dots & 1 \end{pmatrix}.$$

Here the constant parameter r governs the covariance between ensemble members x_m and x_n for $m \neq n$. We draw $N = 10^6$ ensemble forecasts with $M = 10$ members.

In figure 6.5 we show rank histograms for two examples. Firstly, the observations

$y_i, i = 1, 2, \ldots, N$ are IID draws with distribution $N(0, 1)$. The observations therefore share the same marginal distribution of the ensemble members, but do not share the same dependence structure and so are not exchangeable with the members. Secondly, we show the rank histogram for the case when the observation and ensemble members are drawn from the $M + 1$-dimensional multivariate normal distribution with covariance matrix Σ. In both cases we set $r = 0.25$. In the first case, the rank histogram is U-shaped, which is usually taken as an indication of ensemble underdispersion. However, as noted above the marginal distributions of the ensemble members and observations are equal — the U-shaped histogram is a result of the inter-member dependence. The positive correlation of the ensemble members means that, in a sense, the members 'cluster together', and so a larger than desired proportion of observations populate the outer histogram bins. On the other hand, the rank histogram for the second example, in which the ensemble members and observations are exchangeable, is flat. The fact that the observations and ensemble members are multivariate draws from the same distribution means that the observation is equally likely to populate the $M+1$ rank histogram bins. Bröcker and Kantz [2011] reported that the rank histogram can still be used when evaluating ensemble forecasts whose members are exchangeable but not independent. As we have seen in this example, however, it is also necessary for the observation to be exchangeable with the ensemble forecasts — that is, for the observation and ensemble members to be draws from a symmetric $(M+1)$-dimensional multivariate distribution. Furthermore, as noted earlier and in section 4.2.7, the post-processing of ensemble forecasts may induce inter-member dependencies. In that case, the rank histograms of the post-processed ensemble forecasts may not be flat, even if, for example, biases in ensemble location are corrected such that the expectation of the ensemble members and verifying observations are equal.

We now turn to the more interesting scenario of post-processing ensemble forecasts. Recall that in chapter 4 we used a linear function of the ensemble mean to correct for bias in the ensemble location, and a linear function of the ensemble variance to correct for bias in the ensemble dispersion (see equation (4.52)). In this section we consider a simpler example in which the ensemble variance is correctly specified, and we correct the bias in ensemble location with a linear function of the ensemble mean only. Specifically, ensemble forecasts $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$ are adjusted to

$$\hat{x}_m = a + (b - 1)\bar{x} + x_m \text{ for } m = 1, 2, \ldots, M, \tag{6.6}$$

where $\bar{x} = M^{-1} \sum_{m=1}^{M} x_m$ is the aforementioned ensemble mean, and the constants a and b are model parameters. In this example we do not estimate a and b — they are considered to be known exactly. The sample mean of the post-processed ensemble is $\overline{\hat{x}} = a + b\bar{x}$, while the ensemble variance, s^2, remains unchanged. To ease

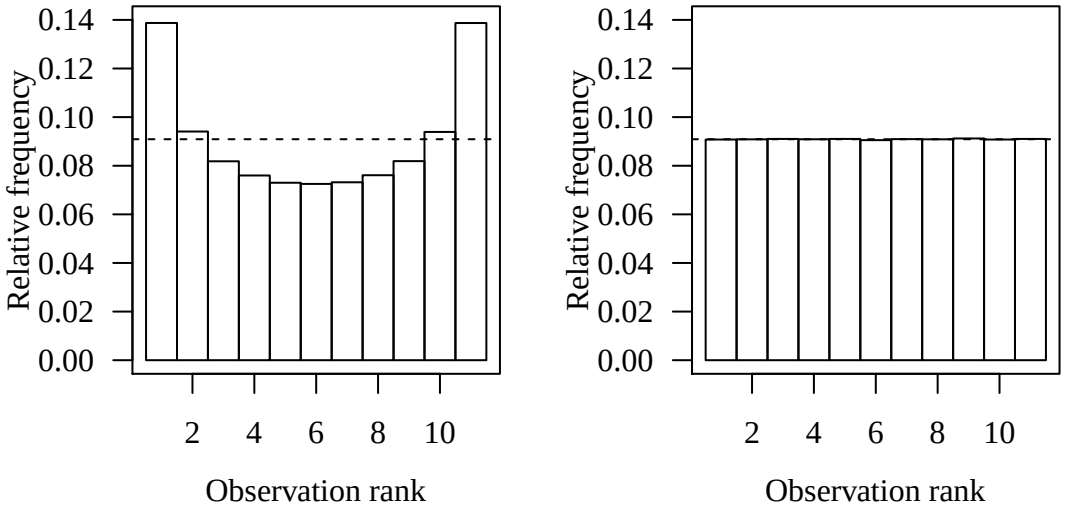

Figure 6.5 Rank histograms for ensemble forecasts whose members are dependent. Observations are independent of the ensemble members (left), and share the same multivariate distribution (right).

notation, in this example we consider ensemble members and observations with unit variance, although our conclusions easily transfer to the case of a constant variance, say c^2. We illustrate the effects of inter-member dependencies on the FCRPS. Recall from section 2.5.4 that the FCRPS for a general ensemble forecast, $\boldsymbol{x}$, with verifying observation, y, is

$$\text{FCRPS}(\boldsymbol{x}, y) = \frac{1}{M}\sum_{m=1}^{M}|y - x_m| - \frac{1}{2M(M-1)}\sum_{m=1}^{M}\sum_{n=1}^{M}|x_m - x_n|\,.$$

The inclusion of the ensemble mean $\bar{x}$ in equation (6.6) induces dependence between the members of the post-processed ensemble forecast $\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_M)$. It can be shown that the correlation between ensemble members $\hat{x}_m$ and $\hat{x}_n$ for $m \neq n$ is

$$\rho(\hat{x}_m, \hat{x}_n)_{m\neq n} = \frac{b^2 - 1}{b^2 + M - 1}. \tag{6.7}$$

Observe, therefore, that $b < 1$ ($b > 1$) induces negative (positive) correlation between the ensemble members, and that the correlation increases with b^2. We explain the relevance of this result in the example below.

We consider a second simulation experiment, which proceeds as follows.

- Simulate N realisations of a latent random variable, ξ, with distribution $N(10, 6^2)$.
- Simulate M-member ensemble forecasts $\boldsymbol{x}_i, i = 1, 2, \ldots, N$, whose members are IID realisations with distribution $N(\xi_i, 1)$.

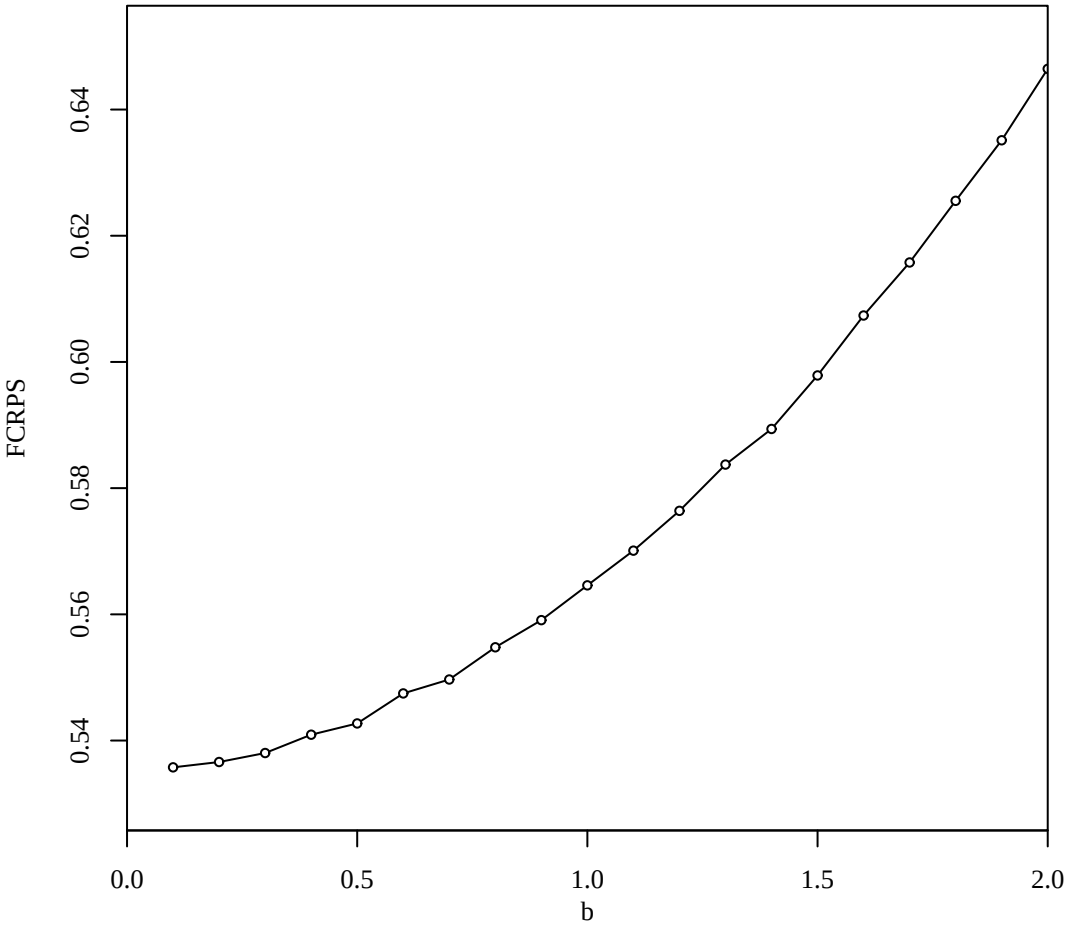

Figure 6.6 The fair CRPS as a function of the model parameter b, for simulated data and ensemble forecasts post-processed with equation (6.6).

- Simulate observations $y_i, i = 1, 2, \ldots, N$, with distribution $N(a + b\xi_i, 1)$.
- Post-process the ensemble forecasts $\boldsymbol{x}_i$ such that $\hat{x}_{i,m} = a + (b-1)\bar{x}_i + x_{i,m}$, for $i = 1, 2, \ldots, N$ and $m = 1, 2, \ldots, M$.
- Repeat the above steps for chosen values of the parameter b.

In this experiment we set $N = 100\,000$, $M = 10$ and $a = 0$. We investigate parameter values b in the set $\{0.1, 0.2, \ldots, 2\}$. In figure 6.6 we show the FCRPS as a function of the parameter b. The FCRPS is an increasing function of b, and is therefore also an increasing function of the inter-member correlation ρ (recall that ρ is also an increasing function of b).

At first sight the plot in figure 6.6 appears to be a surprising result — the expectation of the ensemble members is equal to the expectation of the observations, and the

expectation of the ensemble variance s^2 is equal to the variance of the observations. However, in this case the ensemble variance is a misleading estimate of the variance of the ensemble members, since the ensemble members are no longer independent. It can in fact be shown that the variance of an ensemble member, say $\hat{x}_1$, post-processed with the linear function of the ensemble mean described above, is

$$\text{var}(\hat{x}_1) = \sigma_x^2 = \frac{b^2 + M - 1}{M},$$

where (as in our simulation experiment) the 'raw' ensemble member x_1 has unit variance. Furthermore, the covariance between ensemble members $\hat{x}_m$ and $\hat{x}_n$, for $m \neq n$, is

$$\text{cov}(\hat{x}_m, \hat{x}_n)_{m \neq n} = \sigma_{x_m, x_n} = \frac{b^2 - 1}{M}.$$

In our simulation experiment, therefore, the ensemble members on the ith forecast occasion, $\hat{x}_{im}, m = 1, 2, \ldots, M$, are now realisations of a multivariate normal distribution with mean vector $\boldsymbol{\xi}_\mathbf{i}$, and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{x_m,x_n} & \cdots & \sigma_{x_m,x_n} \\ \sigma_{x_m,x_n} & \sigma_x^2 & \sigma_{x_m,x_n} & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{x_m,x_n} & \cdots & \sigma_{x_m,x_n} & \sigma_x^2 \end{pmatrix}.$$

The plot in figure 6.6 can be explained as follows. Recall from the kernel representation of the continuous ranked probability score (see equation (2.45) on page 58) that the first term in the expression for the FCRPS is the sample estimate of the expectation $E(|y - x|)$, where x is a random variable whose distribution is a probability forecast distribution for the verifying observation, y. For our simulation experiment, however, analytic results exist for this expectation. Observe that the random variable $y - \hat{x}_1$ is normally distributed, with expectation 0 and variance

$$\sigma_{\hat{x}_1, y} = \frac{b^2 + 2M - 1}{M}.$$

Analytic results in Gneiting et al. [2005] show that the expectation $E(|y - \hat{x}_1|)$ is given by

$$E\left(|\hat{x}_1 - y|\right) = \sigma_{\hat{x}_1,y}\left[\frac{y - \hat{x}_1}{\sigma_{\hat{x}_1,y}}\left\{2\phi\left(\frac{y - \hat{x}_1}{\sigma_{\hat{x}_1,y}}\right) - 1\right\} + 2\Phi\left(\frac{y - \hat{x}_1}{\sigma_{\hat{x}_1,y}}\right)\right].$$

This is an increasing function of $\sigma_{\hat{x}_1,y}$, and thus also an increasing function of b. Observe further that the second term in the expression for the FCRPS is independent

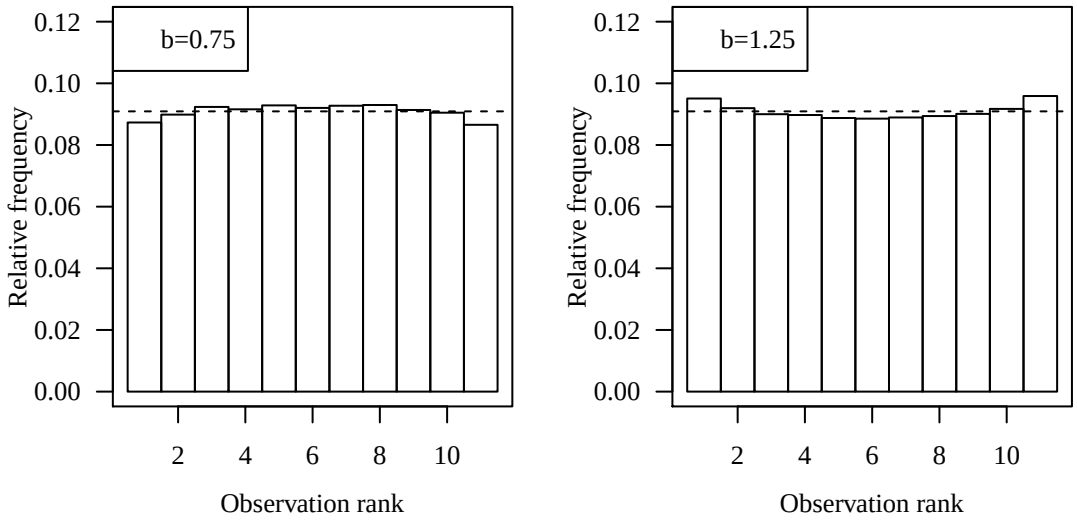

Figure 6.7 Rank histograms for simulated data and ensemble forecasts post-processed with equation (6.6).

of the ensemble post-processing, since

$$|\hat{x}_m - \hat{x}_n| = |x_m - x_n|.$$

It therefore follows that the FCRPS is an increasing function of b, in keeping with the plot shown in figure 6.6.

Figure 6.6 and the above working provides a further illustration of the complicated nature of assessing the skill of ensemble forecasts. We have shown that the fair CRPS attains different values, depending on the value of the parameter b that is used to correct the bias in ensemble location. Intuitively it seems that the post-processed ensemble forecasts are equally skilful regardless of the value of the parameter b — the expectations of the post-processed ensemble forecasts and the verifying observations are equal, and the expectation of the ensemble variance is equal to the variance of the observations. However, larger values of b result in larger variances of the post-processed ensemble members $\hat{x}_m$, as well as larger values of the covariances $\sigma_{\hat{x}_m,\hat{x}_n}$. The larger variances yield larger values of the FCRPS.

Finally, we show rank histograms for the cases $b = 0.75$ and 1.25, which corresponds to $\rho \approx -0.046$ and 0.053, respectively. As with figure 6.5, positive inter-member correlation yields U-shaped rank histograms, while (as we expect) negative inter-member correlation yields inverted U-shaped rank histograms.

The examples provided in this section illustrate that care should be taken when directly post-processing ensemble forecasts, that is, when not sampling ensemble forecasts from probability forecast distributions. In particular, forecasters should be careful when using verification measures such as the rank histogram and fair CRPS

to assess the calibration of ensemble forecasts. Our examples serve to emphasise our point that forecasters should use multiple measures of forecast skill — we have provided an example in which ensemble forecasts whose members exhibit negative inter-member correlation yield inverted U-shaped rank histograms but an improved FCRPS compared with ensemble forecasts whose members are independent (when $b = 1$) and so yield flat rank histograms, but a worse FCRPS. Users of real-world forecasts are unlikely to know the precise nature of any inter-member dependence, which depends on the method used to generate the ensemble forecasts (see the discussion in section 2.2.2).

6.3 Closing remarks

In this thesis we have presented novel work that has served to address some distinct and important topics in the post-processing of ensemble weather forecasts. Firstly, in chapter 3 we presented an investigation in to the comparative skill of competing ensemble post-processing methods in issuing probability forecasts for extreme events. Using the Lorenz 1996 system as a surrogate for the atmosphere, we showed that ensemble post-processing adds significant value to the probabilistic forecast skill. Further investigations in to the benefits of ensemble post-processing for probability forecasts of extreme events are strongly encouraged. We also demonstrated that two popular ensemble post-processing methods, NGR and BMA, exhibit similar probabilistic forecast skill provided that the first and second moments of the probability forecast distributions are allowed to take equally flexible functional forms. It was shown that allowing the bias in ensemble location to be corrected by a linear function of the ensemble mean yielded significant improvements to the forecast skill for extreme events, compared with the seemingly more common practice of a simple constant bias correction.

In chapter 4 we detailed a novel, 'distribution-free' ensemble post-processing method, that could serve as a more useful baseline for comparing new models than the usual baseline forecasts that are based on relative frequencies of the raw ensemble members. Our distribution-free method allows the biases in ensemble location and dispersion to be corrected by linear functions of the ensemble mean and variance, in keeping with the parameterisations of the expectation and variance of forecast probability distributions that we considered in chapter 3. Parameter estimation was performed using the method of moments which, unlike the popular approach of estimation by optimisation of a proper scoring rule, does not require the user to specify a parametric family of distributions with which to model the verifying observations. Rather than producing probability forecast distributions, our distribution-free method leads directly to post-processed ensemble forecasts. The ensemble forecasts also preserve

the rank dependence structure of the raw forecasts, which has been shown to be benefitial in forecasts of multivariate predictands. Perhaps surprisingly, we found that the popular approach of resampling ensemble forecasts from probability forecast distributions issued by ensemble post-processing methods yielded ensemble forecasts that were only slightly more skilful than the ensemble forecasts produced by our distribution-free method. When appropriate we suggest that forecasters should compare the skill of post-processing methods that assume a parametric family of distributions for the verifying observations with our distribution-free method, in order to ensure that the specification of a distribution adds value to the forecast. While we found that parameter estimation by the method of moments was sufficient for correcting the biases in ensemble location, parameter estimation for the bias corrections in the ensemble dispersion was more problematic. We therefore encourage efforts in the direcction of distribution-free parameter estimation with more robust techniques than the method of moments. The development of robust parameter estimates would enable users to compare the skill of different probability distributions with the same expectation and variance.

In chapter 5 we made the first formal investigation in the field in to the effects of uncertainty in the model parameter estimates on the probability forecasts issued by ensemble post-processing methods. We proposed to account for parameter uncertainty by integrating over the sampling distribution of the parameter estimates. We used a bootstrap resampling approach to approximate the sampling distribution, and a Monte Carlo approximation to the aforementioned integral. We demonstrated that accounting for parameter uncertainty yielded more skilful forecasts than the standard approach in the field, in which the parameter estimates are treated as 'truth', and parameter uncertainty is ignored. While we are confident that more sophisticated methods to account for parameter uncertainty can be developed, such as with Bayesian models, our bootstrap approach is easy to implement and applicable to a wide range of ensemble post-processing methods. Our method therefore serves as a useful starting point that can be used to assess the importance of parameter uncertainty for forecasts of a given predictand. It may then subsequently be worth employing a Bayesia model, or attempting to derive analytic results for the sampling distribution of parameter estimates, for specific applications in which accounting for parameter uncertainty is important.

In this chapter we have demonstrated that forecasters should explicitly consider the calibration of the moments of their probability forecast distributions or ensemble forecasts, as well as the usual measures of forecast skill that are typically presented in the literature. In another application to forecasting in the Lorenz 1996 system, we have shown that the standard linear functions of ensemble covariates can be improved upon, to issue probability forecasts distributions whose moments are better

calibrated with the equivalent moments of the verifying observations. Our revised statistical model also leads to improvements in the ignorance and continuous ranked probability scores. We suggest that developers of new statistical models should discuss the features of diagnostic plots, such as those presented in section 6.1. In our view, the combination of the results presented in chapters 3, 5 and section 6.1 of this chapter, should aid in the improvement of the statistical models that are specified by ensemble post-processing methods. Indeed, we intend to consider the combination of these distinct aspects of ensemble post-processing methods in future work.

In this chapter we have also highlighted that ensemble member dependence may lead to misleading conclusions when diagnosing the performance of ensemble forecasts with the popular rank histogram. We encourage the development of verification measures that accomodate ensemble member dependencies, although suspect that such developments may be difficult to implement in practice, as the 'true' dependence structure is not known and must therefore be estimated. Furthermore, we strongly encourage continuations in the development of methods for forecast verification in general. Weighted scoring rules are a promising development for assessing the skill of forecasts for extreme observations. These scores enable the user to weight those observations that are of particular interest while maintaining the desirable property of propriety, and so the scoring rule becomes a weighted average of the forecasts and observations, rather than the mean average as has been reported in this thesis. Diks et al. [2011] and Gneiting and Ranjan [2012] have investigated weighted versions of the likelihood and CRPS, respectively. The development of complimentary graphical tools, such as conditional rank and PIT histograms, seems highly desirable.

Acknowledgements

I am indebted to my lead supervisor, Chris Ferro, who has provided outstanding support throughout the duration of my doctoral studies. Chris has been generous with his time in both discussing ideas and reading my work, despite a busy schedule. I must also thank Chris for his constant support of my endeavours as a member of the England and Great Britain blind football teams, which has at times caused difficulties in maintaining momentum in my work. Chris's support in this regard has enabled me to be successful in both pursuits.

My second supervisor, Frank Kwasniok, is also thanked for his contributions, particularly during the early stages of my work. Frank has provided me with much useful feedback, and I have always found our discussions to be interesting and fruitful.

My thanks also go to several other members of the department. In particular, I have enjoyed many interesting and fruitful discussions with Stefan Siegert, some of which led to a recent publication. Theo Economou, Ben Youngman and Phil Sansom are also thanked for generously sharing their statistical expertise and knowledge of computational methods. During the course of my studies I have also enjoyed working alongside, among others, Lester Kwiatkowski, David Long, Alasdair Hunter, Joe Osborne, Maria Marklove and Robin Beaumont, who have helped to create an enjoyable and productive working environment.

Several authors are thanked for their assistance in sharing their articles in accessible formats. As a researcher who is blind, this has enabled me to study almost completely independently, which I have found very rewarding. In particular, Tilmann Gneiting, Dan Wilks, Tom Hamill and Jochen Bröcker have always been prompt in providing their articles. In addition, Eugenia Kalnay and John Buonaccorsi are thanked for providing the LaTeX source files of their excellent textbooks. My sincere thanks are also due to Alastair Irving for his development and maintainance of the excellent LaTeX-Access Python scripts. This package, which converts LaTeX source code to speech and Braille output, has significantly eased the process of reading and writing materials in LaTeX. My parents are of course thanked for their support, both mathematical (during my childhood) and otherwise. Finally, but by no means of least importance, my partner Danielle is thanked for her support and patience during the difficult process of pursuing my footballing aspirations while

also finishing my studies.

Bibliography

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

Bao, L., Gneiting, T., Grimit, E. P., Guttorp, P., and Raftery, A. E. (2010). Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Monthly Weather Review*, 138(5):1811–1821.

Berrocal, V. J., Raftery, A. E., and Gneiting, T. (2007). Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, 135(4):1386–1402.

Berrocal, V. J., Raftery, A. E., and Gneiting, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *The Annals of Applied Statistics*, 2:1170–1193.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519.

Bröcker, J. (2012). Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, 138(667):1611–1617.

Bröcker, J. and Kantz, H. (2011). The concept of exchangeability in ensemble forecasting. *Nonlinear Processes in Geophysics*, 18(1):1–5.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. CRC Press.

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

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

Diks, C., Panchenko, V., and Van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230.

Draper, N. R., Smith, H., and Pownell, E. (1998). *Applied regression analysis*. Wiley New York, 3rd edition.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48.

Epstein, E. (1969). Stochastic dynamic prediction. *Tellus*, 21(6):739–759.

Erickson, M. C. (1996). Medium-range prediction of PoP and max/min in the era of ensemble model output. In *Conference on weather analysis and forecasting*, volume 15, pages J35–J38. American Meteorological Society.

Ferro, C. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140:1917–1923.

Flowerdew, J. (2012). Calibration and combination of medium-range ensemble precipitation forecasts. *Met Office Forecasting Research Technical Report*, 567.

Fraley, C., Raftery, A. E., Gneiting, T., and Sloughter, J. M. (2007). ensembleBMA: An R package for probabilistic forecasting using ensembles and Bayesian model averaging.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23(7):579–594.

Gillard, J. (2014). Method of moments estimation in linear regression with errors in both variables. *Communications in Statistics-Theory and Methods*, 43(15):3208–3222.

Glahn, B., Peroutka, M., Wiedenfeld, J., Wagner, J., Zylstra, G., Schuknecht, B., and Jackson, B. (2009). MOS uncertainty estimates in an ensemble framework. *Monthly Weather Review*, 137(1):246–268.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133:1098–1118.

Gneiting, T. and Ranjan, R. (2012). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29:411–422.

Gneiting, T., Stanberry, L., Grimit, E., Held, L., and Johnson, N. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17(2):211–235.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14:107–114.

Grimit, E., Gneiting, T., Berrocal, V., and Johnson, N. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132(621C):2925–2942.

Hamill, T. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129:550–560.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau Jr, T. J., Zhu, Y., and Lapenta, W. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, 94(10):1553–1565.

Hamill, T. M. and Colucci, S. J. (1997). Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, 125:1312–1327.

Hamill, T. M. and Colucci, S. J. (1998). Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Monthly Weather Review*, 126:711–724.

Hamill, T. M., Whitaker, J. S., and Wei, X. (2004). Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, 132:1434–1447.

Harris, I. R. (1989). Predictive fit for natural exponential families. *Biometrika*, 76(4):675–684.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570.

Hoffman, R. N. and Kalnay, E. (1983). Lagged average forecasting, an alternative to monte carlo forecasting. *Tellus A*, 35(2):100–118.

Hou, D., Kalnay, E., and Droegemeier, K. K. (2001). Objective verification of the SAMEX'98 ensemble forecasts. *Monthly Weather Review*, 129(1):73–91.

Houtekamer, P., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. (1996). A system simulation approach to ensemble prediction. *Monthly Weather Review*, 124(6):1225–1242.

Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability*. Citeseer.

Kharin, V. V. and Zwiers, F. W. (2003). Improved seasonal probability forecasts. *Journal of Climate*, 16(11):1684–1701.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2):137–163.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Leith, C. (1974). Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, 102(6):409–418.

Lerch, S. and Thorarinsdottir, T. L. (2013). Comparison of nonhomogeneous regression models for probabilistic wind speed forecasting. *Tellus A*, 65:21206.

Lorenz, E. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141.

Lorenz, E. (1965). A study of the predictability of a 28-variable atmospheric model. *Tellus*, 17(3):321–333.

Lorenz, E. N. (1996). Predictability: A problem partly solved. In *Proc. Seminar on Predictability*, volume 1, pages 1–18. ECMWF.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22:1087–1096.

Messner, J. W., Mayr, G. J., Zeileis, A., and Wilks, D. S. (2014). Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Monthly Weather Review*, 142(1):448–456.

Möller, A., Lenkoski, A., and Thorarinsdottir, T. L. (2013). Multivariate probabilistic forecasting using ensemble bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, 139(673):982–991.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12:595–600.

Nelder, J. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Pinson, P. (2012). Adaptive calibration of (u, v)-wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138(666):1273–1284.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174.

Richardson, L. (2007). *Weather prediction by numerical process*. Cambridge University Press.

Roulston, M. S. and Smith, L. A. (2003). Combining dynamical and statistical ensembles. *Tellus A*, 55:16–30.

Schefzik, R., Thorarinsdottir, T., and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28(4):616–640.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1086–1096.

Scheuerer, M. and Büermann, L. (2014). Spatially adaptive post-processing of ensemble forecasts for temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(3):405–422.

Schuhen, N., Thorarinsdottir, T. L., and Gneiting, T. (2012). Ensemble model output statistics for wind vectors. *Monthly weather review*, 140(10):3204–3219.

Siegert, S., Sansom, P. G., and Williams, R. M. (2015a). Parameter uncertainty in forecast recalibration. *Quarterly Journal of the Royal Meteorological Society, in press*.

Siegert, S., Stephenson, D. B., Sansom, P. G., Scaife, A. A., Eade, R., and Arribas, A. (2015b). A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *arXiv preprint arXiv:1504.01933*.

Sloughter, J., Raftery, A., Gneiting, T., and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, 135(9):3209–3220.

Sloughter, M. J., Gneiting, T., and Raftery, A. E. (2013). Probabilistic wind vector forecasting using ensembles and bayesian model averaging. *Monthly Weather Review*, 141(6):2107–2119.

Stephenson, D. B., Coelho, C. A. S., Doblas-Reyes, F. J., and Balmaseda, M. (2005). Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus A*, 57:253–264.

Székely, G. (2003). E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, (03-05):2000–2003.

Thorarinsdottir, T. L. and Gneiting, T. (2010). Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):371–388.

Tippett, M. K., Goddard, L., and Barnston, A. G. (2005). Statistical-dynamical seasonal forecasts of central-southwest Asian winter precipitation. *Journal of Climate*, 18(11):1831–1843.

Wang, X. and Bishop, C. H. (2005). Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society*, 131:965–986.

Wilks, D. S. (2005). Effects of stochastic parametrizations in the Lorenz'96 system. *Quarterly Journal of the Royal Meteorological Society*, 131(606):389–407.

Wilks, D. S. (2006a). Comparison of ensemble-MOS methods in the Lorenz'96 setting. *Meteorological Applications*, 13:243–256.

Wilks, D. S. (2006b). *Statistical Methods in the Atmospheric Sciences*. Academic Press, 2nd edition.

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, 16:361–368.

Williams, R. M., Ferro, C. A. T., and Kwasniok, F. (2014). A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1112–1120.